Universitat Autònoma de Barcelona

Development of efficient eukaryotic and bacterial expression systems for functional studies of recombinant proteins of biomedical interest

Laura Mariño Puertas

UNIVERSITAT AUTÒNOMA DE BARCELONA

Departamento de Bioquímica y Biología Molecular
Programa de Doctorado en Bioquímica, Biología Molecular y Biomedicina

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

INSTITUTO DE BIOLOGÍA MOLECULAR DE BARCELONA

Structural Biology Unit
Proteolysis Lab

Development of efficient eukaryotic and bacterial expression systems for functional studies of recombinant proteins of biomedical interest

Memoria presentada por Laura Mariño Puertas para optar al título de doctor por la Universitat Autònoma de Barcelona

Doctorando:

Laura Mariño Puertas

Supervisores:

Prof.F.Xavier Gomis-Rüth **Dr.Theodoros Goulas** **Dr.Iñaki de Diego Martínez**

Tutor:

Dr.Sandra Villegas

Barcelona, Diciembre 2019

A todos los que siempre confiaron en mí,

« Wanderer above the Sea of Fog » (1817-18), C. D. Friedrich

"Life is not easy for any of us. But what of that? We must have perseverance and above all confidence in ourselves. We must believe that we are gifted for something and that this thing must be attained."

"La vida no es fácil, para ninguno de nosotros. Pero... ¡qué importa! Hay que perseverar y, sobre todo, tener confianza en uno mismo. Hay que sentirse dotado para realizar alguna cosa y que esa cosa hay que alcanzarla, cueste lo que cueste."

Marie Curie

ACKNOWLEDGEMENTS

No es fácil hacer una carta de agradecimientos, nada fácil, porque si soy sincera, ¡tengo mucho que agradecer!

En primer lugar, me gustaría dar las gracias de corazón a Xavi, por traerme de Inglaterra y hacerme un hueco en su laboratorio, y por no haberme dicho nunca que no a ninguna de mis ideas, ¡¡por muy cara que fuera!! (incubador de Expi, estancia en Suiza, curso EMBO en Oxford, curso de cristalografía en Granada…). ¡¡Me ha permitido crecer enormemente en el terreno profesional y aún más importante, a creer en mis ideas y en mí misma!! GRACIAS.

Tampoco puedo dejar en el tintero a mis supervisores de tesis, a Iñaki y Theo. Iñaki me ha enseñado a tomarme menos en serio y con él viví mi primer sincrotrón (en Grenoble, Francia) y eso nunca se olvida (aunque sólo sea por el miedo que dan las instalaciones.). De Theo he aprendido perseverancia, a ser un poco más paciente y, ¡mucho más práctica! Ambos me han hecho mejor científica, ¡¡así que estaré siempre en duda con ellos!!

A todos los compis actuales y pasados del lab, que sois mucho más que compañeros, sois amigos, compañeros de juegos y ha sido un placer compartir estos cinco años con vosotros… ¡y que estoy segura de que serán muchos más!

Tampoco me quiero olvidar de todos los compañeros y personal del Parc Cientific, que habéis conseguido que nunca me sintiera sola allí, ¡¡ni siquiera a horas intempestivas!!

Debo también agradecer a mis colaboradores, que me han prestado desinteresadamente, vectores, protocolos y resuelto dudas. Gracias, vosotros le dais sentido a la ciencia, porque si no fuese compartido el conocimiento, ¡no sería ciencia!

También le quiero agradecer a Alex su comprensión, ¡¡porque no es fácil tener una novia chiflada, que tiene ideas científicas cepillándose los dientes y que su idea de domingo sea ir a pasar las células!! Además, él ha sido parte fundamental en la corrección de la tesis, leyéndosela entera y dándome su feedback en inglés. ¡¡Sin él, esta tesis no sería ni parecida!!

Y, por último, quiero agradecer a mi familia por su apoyo incondicional. Porque no tiene que ser nada fácil tener una hija a la que le encanta estar lejos de casa, pasando aventuras y viviendo en la distancia sus logros y penas…¡¡pero sin ese apoyo en la distancia, jamás habría llegado hasta dónde lo he hecho!!

Y por favor, si me olvido de alguien, no me lo tengáis en cuenta, todos los que habéis formado parte de estos cinco años, para bien y para mal, ¡habéis sido importantes!

DECLARATION OF AUTHORSHIP

I, **Laura Mariño Puertas**, declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research:

"Development of efficient eukaryotic and bacterial expression systems for functional studies of recombinant proteins of biomedical interest"

I confirm that:

- This work was done wholly while in candidature for a research degree at this University;
- Where I have consulted the published work of others, this is always clearly attributed;
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

There are parts of this work that have been published as scientific publications:

- Goulas T, Garcia-Ferrer I, Marrero A, Marino-Puertas L, Duquerroy S, Gomis-Rüth FX. (2017). **Structural and functional insight into pan-endopeptidase inhibition by α2-macroglobulins**. *Biological Chemistry, 398*(9):975-994. doi: 10.1515/hsz-2016-0329.
- Marino-Puertas L., Goulas T and Gomis-Rüth FX. (2017). **Matrix metalloproteinases outside vertebrates.** *Biochimica Biophysica Acta Molecular Cell Research, 1864*(11 Pt A):2026-2035. doi: 10.1016/j.bbamcr.2017.04.003. Epub 2017 Apr 7. Review.
- Del Amo-Maestro, L., Marino-Puertas, L., Goulas, T. & Gomis-Rüth, F. X. (2019). **Recombinant production, purification, crystallization, and structure analysis of human transforming growth factor β2 in a new conformation**. *Scientific reports, 9*(1), 8660.
- Marino-Puertas, L., Del Amo-Maestro, L., Taulés, M., Gomis-Rüth, F. X., & Goulas, T. (2019). **Recombinant production of human α2-macroglobulin variants and interaction studies with recombinant G-related α2-macroglobulin binding protein and latent transforming growth factor-β2**. *Scientific reports, 9*(1), 9186.

Signed: Date:

ABSTRACT

In this thesis, efforts have been made to develop a system for heterologous expression of human glycoproteins with high yield, purity and homogeneity, as well as the necessary flexibility to engineer the protein at will. We established two eukaryotic expression systems, based in *Drosophila* Schneider 2 cells (inducible expression of stable cell lines by addition of Cu_2SO_4 and transient expression with polyethylenimine (PEI) transfection) and in mammalian cells (transient expression on Expi293F cells with PEI transfection).

Proteolysis is a fundamental part of the protein turnover in living organisms and its imbalance may lead to diseases, causing inflammation, neurodegenerative diseases, cancer or cell death among other disorders. Thus, mechanisms that regulate peptidase activity must be exquisitely regulated to ensure proper proteolytic activity at the right time and place.

The human subtilisin kexin-isozyme 1 (hSKI1) is the serine protease which catalyzes the first step in the proteolytic activation of the sterol regulatory element binding proteins (SREBPs), key regulator of the intracellular lipid metabolism. This protease is a transmembrane protein localized in endoplasmic reticulum and Golgi apparatus and its maturation is a very complex process. We achieved to express it soluble in the insect and mammalian expression systems and laid the foundations for future biochemical and structural studies.

In the second project, the main protease inhibitor of the plasma blood has been studied (the human α2-macroglobulin, hα2M), which has great impact on human physiology by participating in inhibition of a broad spectrum of endopeptidases, but also by modulating the activity of cytokines, hormones, growth factors and other proteins. Biochemical, biophysical, structural and binding assays with other interesting targets (G-related α2-macroglobulin binding protein from *Streptococcus pyogenes,* GRAB; human transforming growth factor-β2, TGFβ2; and human lipoprotein receptor-related protein 1, LRP1) in complex with wild-type and recombinant hα2Ms were performed to characterize these interactions. Therefore, we expressed the hα2M variants in eukaryotic cells and purified the endogenous protein from blood.

Furthermore, we achieved to express soluble pro-TGFβ2 in insect and mammalian cells and we managed to crystallize the mature part of this protein.

And finally, we reviewed the available literature on the matrix metalloproteinase (MMP) family outside vertebrates and performed database searches for potential MMP catalytic domains in invertebrates, plants, fungi, viruses, protists, archaea and bacteria.

Overall, the present thesis has established new and promising expression systems at the laboratory and has contributed substantially to the field at the expression, purification and crystallization levels on the serine protease hSKI1 and protease inhibitor hα2M, alone or interacting with GRAB and TGFβ2.

RESUMEN

En esta tesis, se han realizado esfuerzos para desarrollar un sistema para la expresión heteróloga de glicoproteínas humanas con alto rendimiento, pureza y homogeneidad, así como la flexibilidad necesaria para diseñar la proteína a voluntad. Desarrollamos dos sistemas de expresión eucariotas, uno en células de *Drosophila* Schneider 2, (expresión inducible mediante la adición de Cu_2SO_4 y expresión transitoria con transfección con polietilenimina (PEI)); y otro en células de mamífero, (un sistema de expresión transitoria basado en células Expi293F y transfección con PEI).

La proteólisis es una parte fundamental del recambio proteico en los organismos vivos y su desequilibrio puede conducir enfermedades, causando inflamación, enfermedades neurodegenerativas, cáncer o muerte celular, entre otros trastornos. Por lo tanto, los mecanismos que regulan la actividad de peptidasas deben estar exquisitamente regulados para garantizar una actividad proteolítica adecuada en el momento y lugar correctos.

La subtilisina kexina-isozima 1 humana (hSKI1) es la serina proteasa que cataliza el primer paso en la activación proteolítica de las proteínas de unión al elemento regulador de esteroles (SREBP), regulador clave del metabolismo de los lípidos intracelulares. Esta proteasa es una proteína transmembrana localizada en el retículo endoplásmico y el aparato de Golgi y su maduración es un proceso muy complejo. Logramos una expresión soluble en los sistemas de expresión de insecto y mamífero y sentamos las bases para futuros estudios bioquímicos y estructurales.

En el segundo proyecto, se ha estudiado el principal inhibidor de proteasas del plasma (la α2-macroglobulina humana, hα2M), que tiene un gran impacto en la fisiología humana ya que neutraliza un amplio espectro de endopeptidasas, pero también modula la actividad de las citocinas, hormonas, factores de crecimiento y otras proteínas. Se realizaron ensayos bioquímicos, biofísicos, estructurales y de unión con otras proteínas interesantes (G-related α2-macroglobulin binding protein de Streptococcus *pyogenes*; human transforming growth factor-β2, TGFβ2; and human lipoprotein receptor-related protein 1, LRP1) en complejo con endógena y recombinante hα2Ms para caracterizar estas interacciones. Por lo tanto, también expresamos las variantes hα2M en células eucariotas y purificamos la proteína endógena de sangre.

Además, expresamos pro-TGFβ2 en células de insecto y mamífero y cristalizamos la forma madura de TGFβ2.

Y finalmente, revisamos la literatura disponible sobre la familia de metaloproteinasas de matriz (MMP) fuera de los vertebrados y realizamos búsquedas en la base de datos para potenciales dominios catalíticos de MMP en invertebrados, plantas, hongos, virus, protistas, arqueas y bacterias.

En general, la presente tesis ha establecido nuevos y prometedores sistemas de expresión en el laboratorio y ha contribuido sustancialmente al campo en los niveles de expresión, purificación y cristalización en la serina proteasa hSKI1 y el inhibidor de proteasas de la sangre hα2M, solo o interactuando con GRAB y TGFβ2.

RESUM

Al llarg aquesta tesi s'han fet molts esforços per desenvolupar un sistema d'expressió heteròloga de glicoproteïnes humanes d'alt rendiment, puresa i homogeneïtat, així amb la flexibilitat necessària per modificar la proteïna a voluntat. Hem desenvolupat dos sistemes d'expressió a eucariotes, uno a cèllules de *Drosophila* Schneider 2, per a l'expressió inductible mitjançant l'addició de Cu2SO4 i per a l'expressió transitòria amb transfecció de polietilenimina (PEI); i un per a mamífer, un sistema d'expressió transitòria basat en cèl·lules Expi293F i transfecció de PEI.

La proteòlisi és una part fonamental de la rotació de proteïnes en organismes vius i el seu desequilibri pot comportar malalties, provocant inflamacions, malalties neurodegeneratives, càncer o mort cel·lular entre altres trastorns. Per tant, els mecanismes que regulen l'activitat de la peptidasa s'han de regular de manera estricta per garantir l'activitat proteolítica adequada en el moment i lloc adequats.

La queilis-isozima 1 de la subtilisina humana (hSKI1) és la serina proteasa que catalitza el primer pas a l'activació proteolítica de les proteïnes d'unió de l'element regulador d'esterol (SREBPs), regulador clau del metabolisme dels lípids intracel·lulars. Aquesta proteasa és una proteïna transmembrana localitzada al reticle endoplasmàtic i l'aparell de Golgi i la seva maduració és un procés molt complex. Vam aconseguir una expressió soluble als sistemes d'expressió d'insectes i mamífers i establim les bases per a futurs estudis bioquímics i estructurals amb garanties d'èxit.

Al segon projecte, el principal inhibidor de la proteasa del plasma (la α2-macroglobulina humana, hα2M), que neutralitza un ampli espectre d'endopeptidases, però també modula l'activitat de les citocines, hormones, factors de creixement i altres proteïnes i, per tant, per tant té un gran impacte en la fisiologia humana. Es van realitzar assajos bioquímics, biofísics, estructurals i d'unió amb altres dianes interessants (Streptococcus GRAB, TGFβ2 humà i LRP1 humà) en complexos amb hα2M autèntics i recombinants per a caracteritzar aquestes interaccions. Per tant, també vam expressar les variants hα2M en cèl·lules eucariotes i vam purificar l'autèntica proteïna de la sang.

A més, expressem pro-TGFβ2 a cèllules d'insecte i mamífer i cristalizamos la forma madura de TGFβ2.

I finalment, vam revisar la literatura disponible sobre la família de les metaloproteinases de matriu (MMP) fora de vertebrats i vam realitzar cerques a la base de dades de dominis catalítics de MMP potencials en invertebrats, plantes, fongs, virus, protistes, arqueus i bacteris.

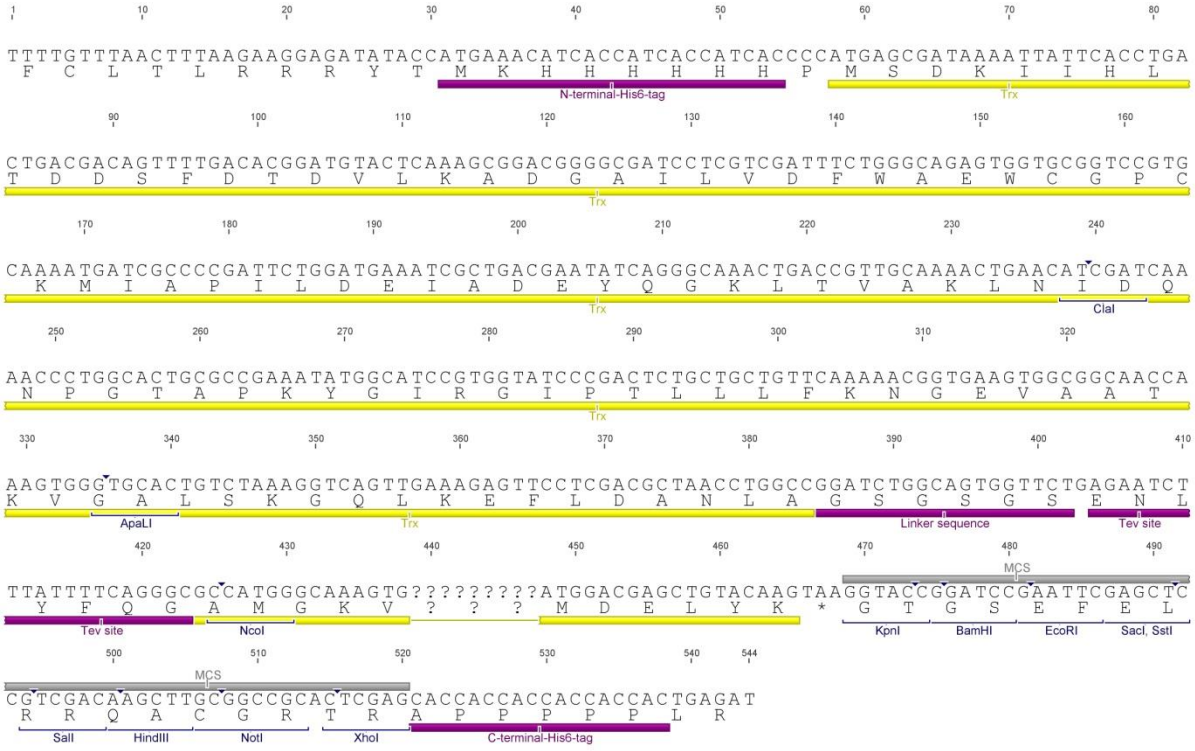
En general, la present tesi ha establert nous i prometedors sistemes d'expressió al laboratori, que han beneficiat i ha contribuït substancialment al camp als nivells d'expressió, purificació i cristal·lització de la proteïna serina hSKI1 i l'inhibidor de la proteasa hα2M, sols o interactuant amb GRAB i TGFβ2.

LIST OF CONTENTS

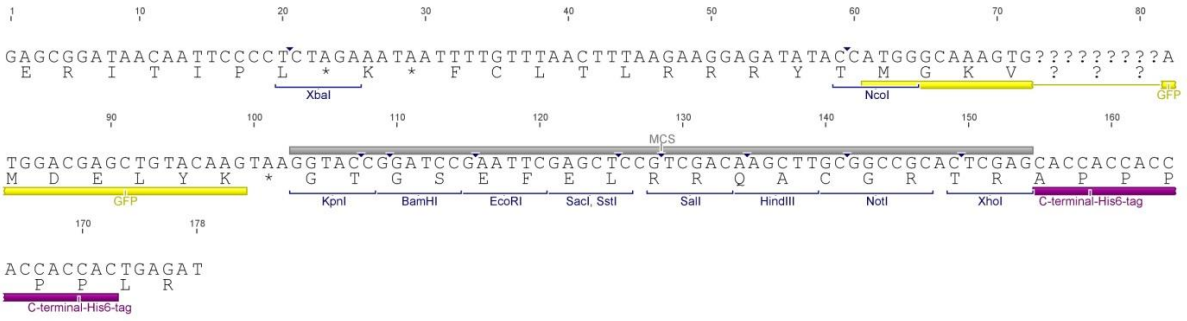
LIST OF FIGURES

LIST OF TABLES

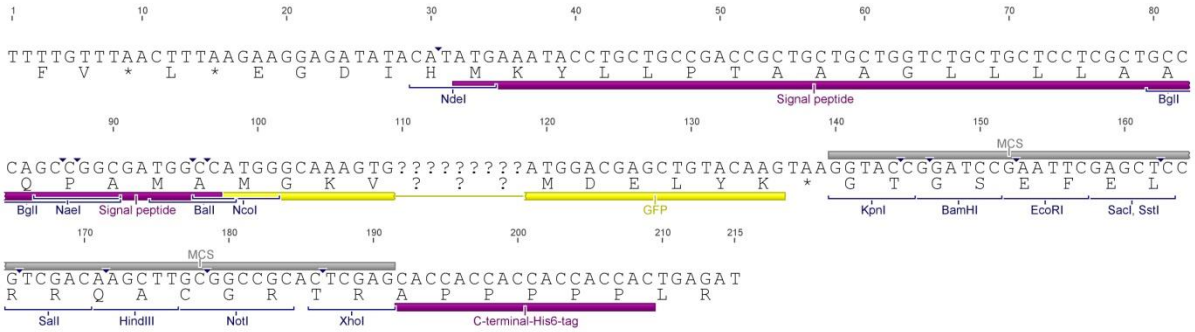
LIST OF EXPRESSION VECTORS



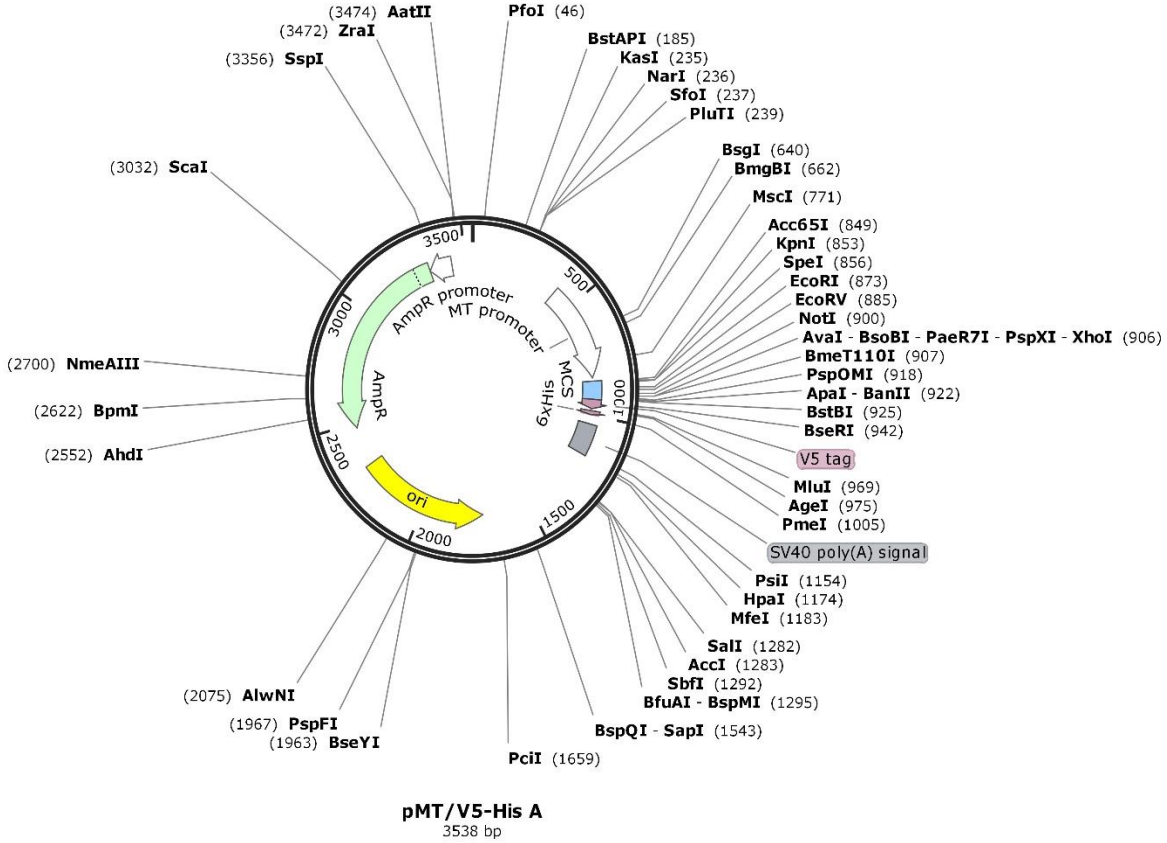
pCri (4a) map. Nucleotide sequences and their translations underlined with different colors are symbolized: in *yellow*, Trx (thioredoxin)-fusion and GFP sequence; in *violet*, linker sequence, Tev (*tobacco-etch virus* peptidase) cleavage site, N- and C-terminal His-tag; in *gray*, multiple cloning site (MCS). The question marks inside the GFP sequence indicate omitted sequence from the plasmid map. Further information in Goulas et al., 2014a and Table 8.



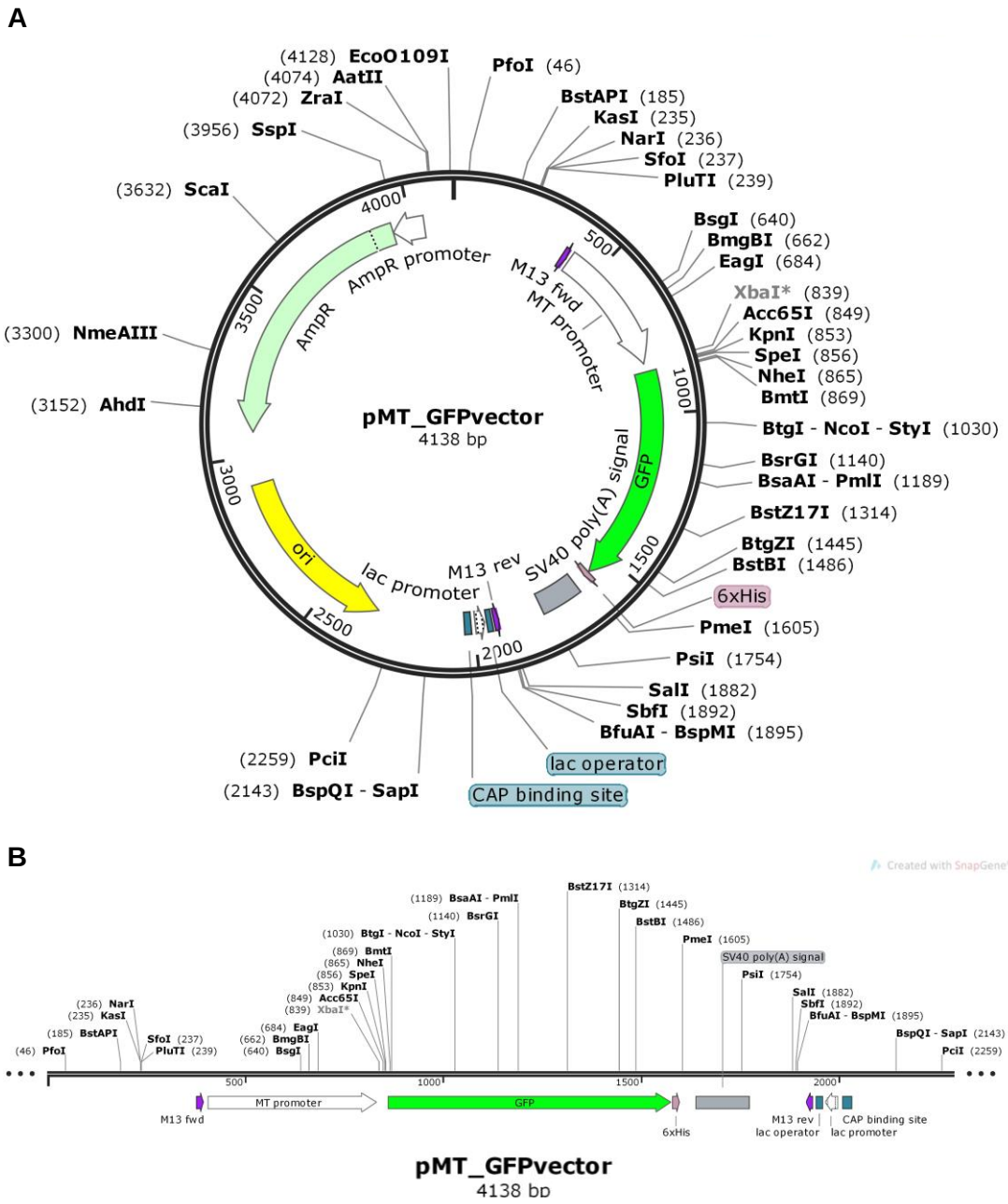
pCri (7a) map. Nucleotide sequences and their translations underlined with different colors are symbolized: in *yellow*, GFP sequence; in *violet*, C-terminal His-tag; in *gray*, multiple cloning site (MCS). The question marks inside the GFP sequence indicate omitted sequence from the plasmid map. Further information in Goulas et al., 2014a and Table 8.



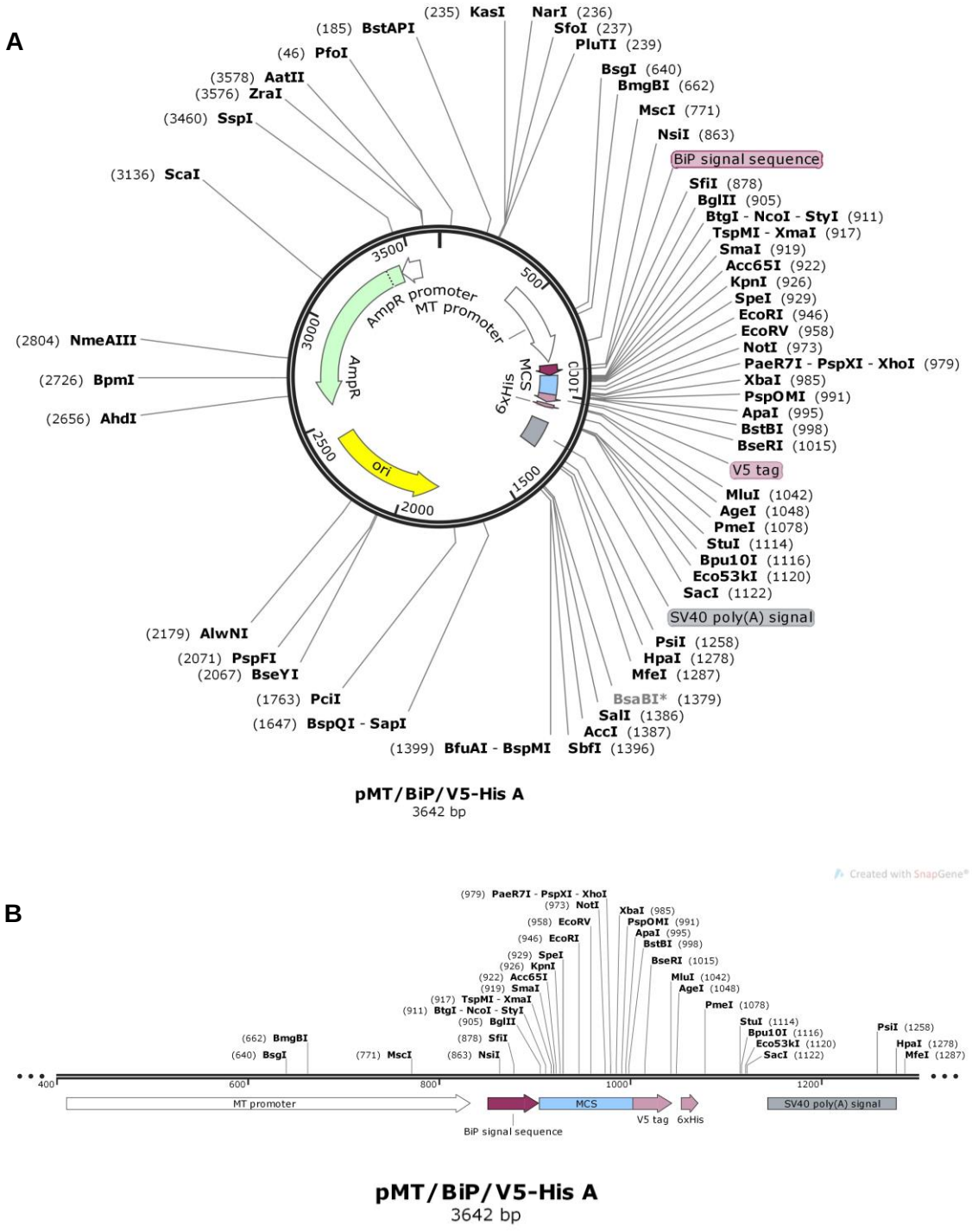
pCri (9a) map. Nucleotide sequences and their translations underlined with different colors are symbolized: in *yellow*, GFP sequence; in *violet*, signal peptide, linker sequence, Tev cleavage site and C-terminal His-tag; in *gray*, multiple cloning site (MCS). The question marks inside the GFP sequence indicate omitted sequence from the plasmid map. Further information in Goulas et al., 2014a and Table 8.



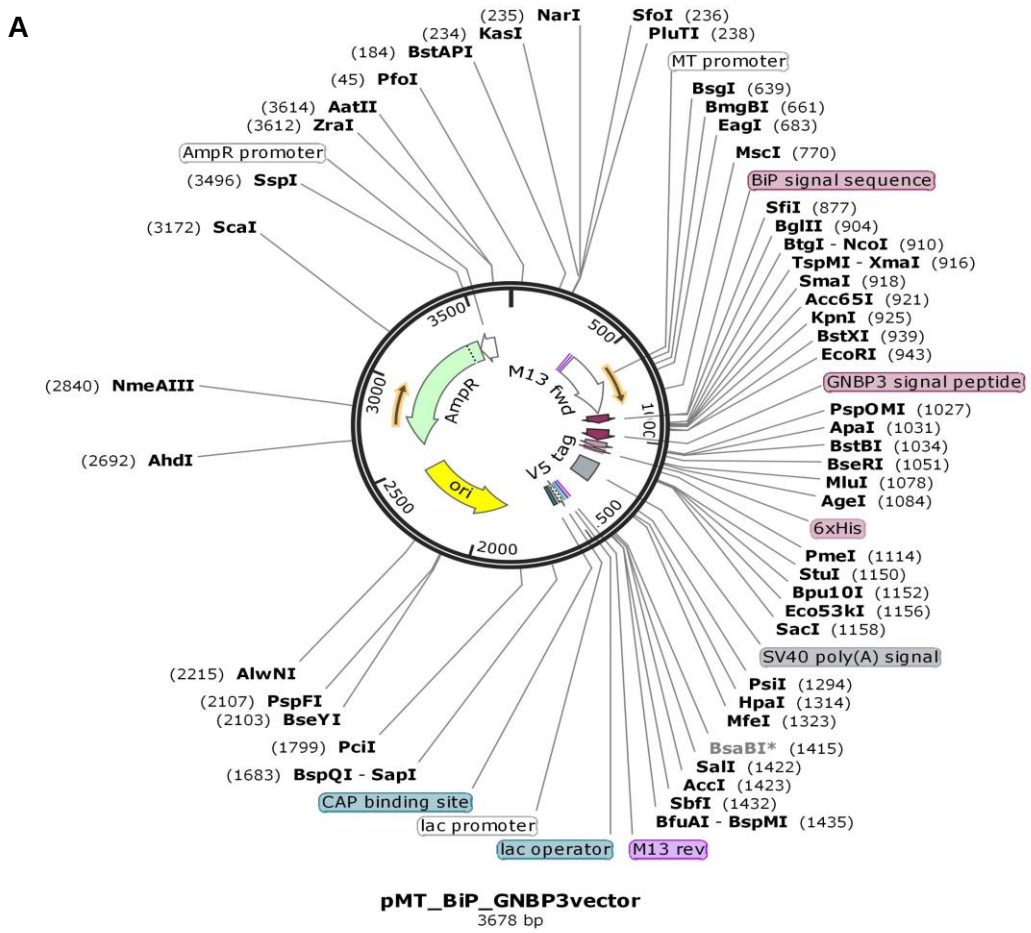
pMT/V5-His map. Feature plasmid map with different colors representing: in *green*, ampicillin resistance; in *yellow*, origin of replication; in *white*, *Drosophila* metallothionein (MT) and AmpR promoters; in *blue*, multiple cloning site (MCS); in *purple* C-terminal His-tag; and in *gray*, SV40 polyadenylation signal (poly(A) signal). Further information in Invitrogen and Table 8. Plasmid generously provided by Philippe Leone from AFMB, Marseille (France).

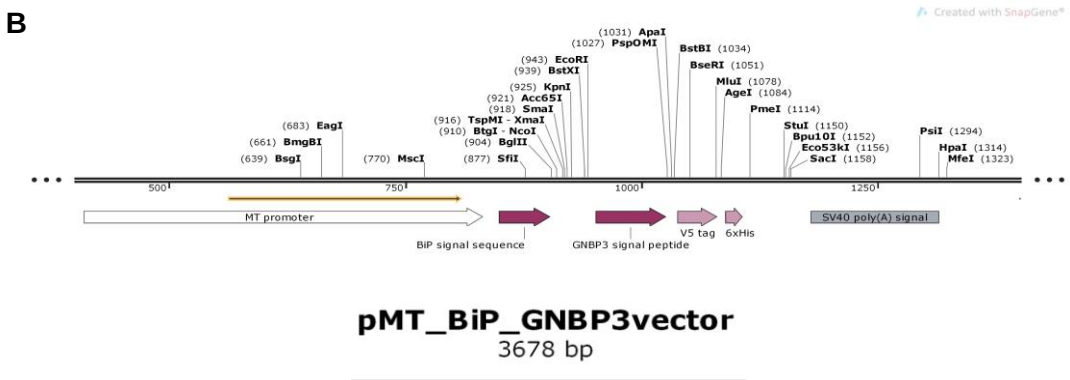


pMT/GFP-His map (version of pMT/V5-His vector with an introduced GFP sequence between the MT promoter and the His tag sequences). **(A)** Feature plasmid map with different colors representing: in *pale green*, ampicillin resistance; in *dark green*, GFP gen; in *yellow*, origin of replication; in *white*, *Drosophil*a metallothionein (MT) and AmpR promoters; in *blue*, multiple cloning site (MCS); in *purple* C-terminal His-tag; and in *gray*, SV40 polyadenylation signal (poly(A) signal). **(B)** Magnification of the area of interest for cloning. Plasmid generously provided by Philippe Leone from AFMB, Marseille (France).

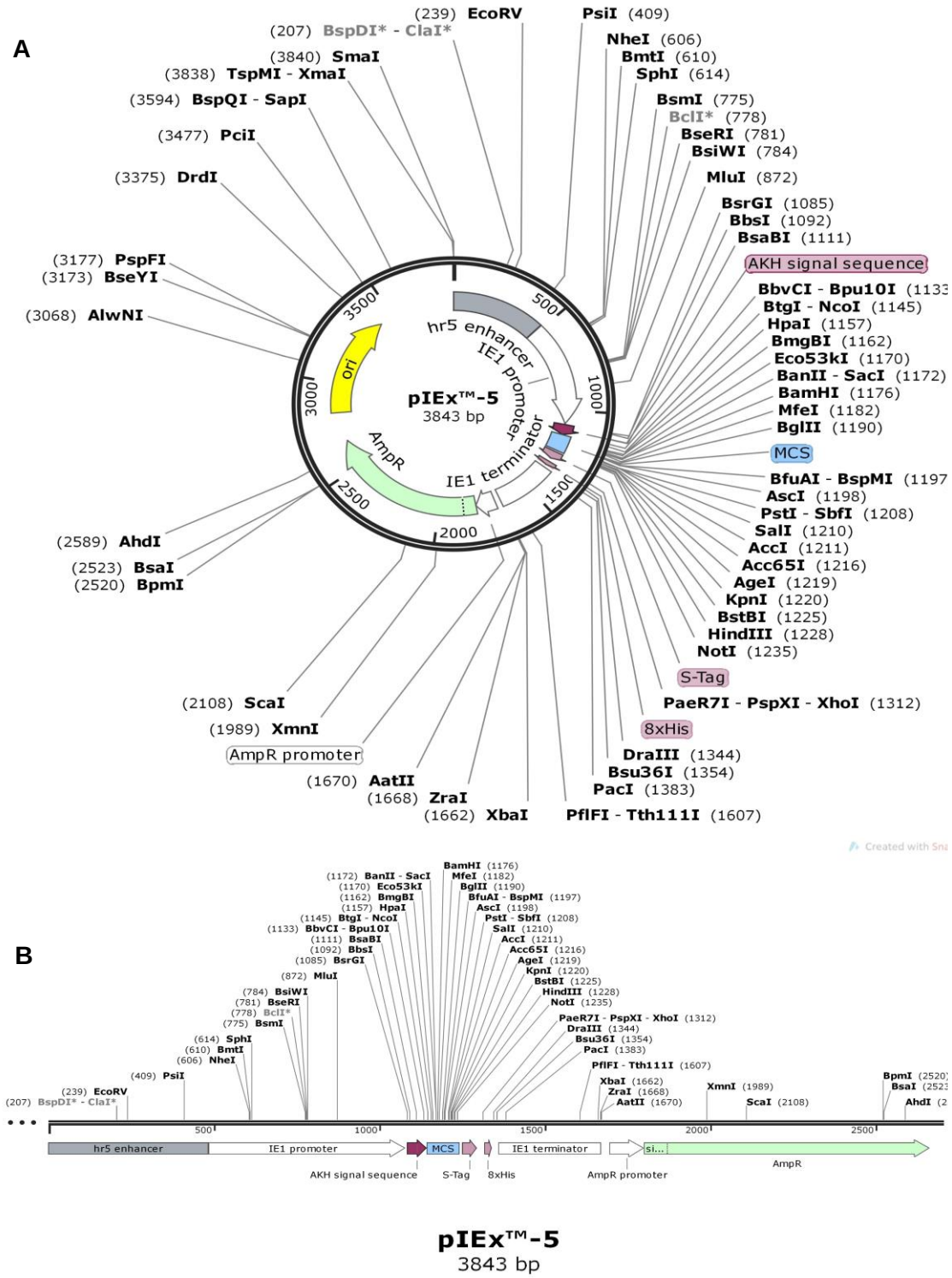


pMT/BiP/V5-His map. **(A)** Feature plasmid map with different colors representing: in *green*, ampicillin resistance; in *yellow*, origin of replication; in *white*, *Drosophila* metallothionein (MT) and AmpR promoters; in *dark purple*, *Drosophila* BiP signal sequence; in *blue*, multiple cloning site (MCS); in *pale purple* C-terminal His-tag; in *gray*, SV40 polyadenylation signal (poly(A) signal). Further information in Invitrogen and Table 8. **(B)** Magnification of the area of interest for cloning. Plasmid generously provided by Philippe Leone from AFMB, Marseille (France).



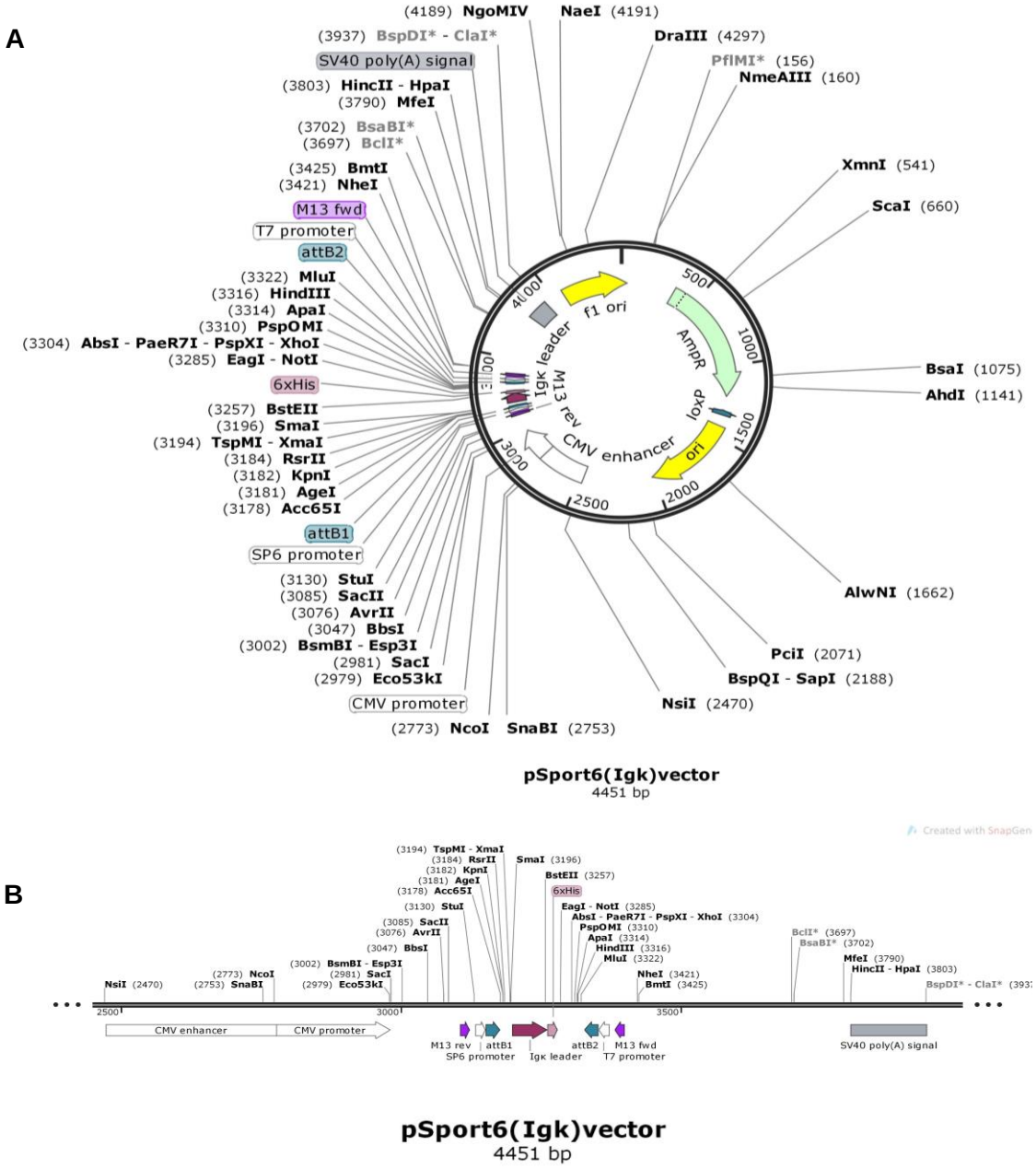


pMT/GNBP3-His map (version of a MT/BiP/V5-His vector with an introduced sequence for GNBP3 signal peptide before the C-terminal His tag sequence)**. (A)** Feature plasmid map with different colors representing: in *green*, ampicillin resistance; in *yellow*, origin of replication; in *white*, *Drosophila* metallothionein (MT) and AmpR promoters; in *dark purple*, *Drosophila* BiP and then GNBP3 (*Drosophila* melanogaster signal peptide from Gram-negative bacteria-binding protein 3) signal sequences; in *blue*, multiple cloning site (MCS); in *pale purple* C-terminal His-tag; in *gray*, SV40 polyadenylation signal (poly(A) signal). **(B)** Magnification of the area of interest for cloning. Plasmid generously provided by Philippe Leone from AFMB, Marseille (France).



pIEx map. **(A)** Feature plasmid map with different colors representing: in *green*, ampicillin resistance; in *yellow*, origin of replication; in *gray*, hr5 enhancer; in *white*, IE1 promoter and IE1terminator; in *dark purple,* adipokinetic hormone (AKH) signal sequence; in *blue*, multiple cloning site (MCS); in *pale purple* C-terminal His-tag. Further information in Novagen and Table 8. **(B)** Magnification of the area of interest for cloning.

XXX



pCMV-SPORt6 map. **(A)** Feature plasmid map with different colors representing: in *green*, ampicillin resistance; in *yellow*, origin of replication; in *white*, CMV (human cytomegalovirus) enhancer and CMV promoter; in *dark purple*, Igκ ladder (ladder sequence from mouse immunoglobulin κ light chain); in *pale purple* C-terminal His-tag; ; in *gray*, SV40 polyadenylation signal (poly(A) signal). Further information in Invitrogen and Table 8. **(B)** Magnification of the area of interest for cloning. Plasmid generously provided by Nadia Sukusu Nielsen (University of Aarhus).

LIST OF ABBREVIATIONS

Å	Ångstrom
aa	Amino acid
AcMNPV	Baculovirus *Autographa californica* multicapsid nucleopolyhedrovirus
ADAMT	A desintegrin-like and metallopeptidase with thrombospondin type 1 motif
AKH	Adipokinetic hormone
AMPs	Antimicrobial peptides
ATF6	Activating transcription factor 6
α2M	α2-macroglobulin
BDNF	Brain-derived neurotrophic factor
BEVS	Baculovirus expression vector system
BRD	Bait region domain
BSA	Bovine serum albumin
BSC40	African green monkey cells derived from BSC-1 cells
BS^3	Bis(sulfosuccinimidy] suberate
BSS	Before shedding site
BTMD	Before transmembrane domain
bα2M	Bacterial α2-macroglobulin
βOG	Octyl-β-glucoside
C-	Carboxy-
C3, C4 and C5	Complement components 3, 4 and 5
$CaCl_2$	Calcium chloride
CD	Catalytic domain
CHO	Chinese hamster ovary cells
C- hα2M	C-terminal recombinant protein of human α2-macroglobulin
CMV	Cytomegalovirus

CREBs	Cyclic AMP-responsive element binding proteins
Cryo-EM	Cryo-electron microscopy
CT	Cytosolic
CTS	C-terminal sub-domain
CUB domain	Domain first described in complement C1R/C1s, Uegf, Bmp1 proteins
Da	Dalton
DES	*Drosophila* Expression System®
DHFR	Dihydrofolate reductase activity
DNA	Deoxyribonucleic acid
DNTB	5,5'-dithiobis-(2-nitrobenzoic acid) (Ellman's reagent)
dNTPs	Deoxynucleotide triphosphates,
DTT	1,4-Dithiothreitol
EC	Enzyme commission
ECAM	*Escherichia coli* α2-macroglobulin
ECM	Extracellular matrix
E.coli	*Escherichia coli*
EDTA	Ethylenediaminetetraacetic acid
EM	Electron microscopy
EMA	European Medicines Agency
ER	Endoplasmic reticulum
FBS	Fetal bovine serum
FDA	Food and drug administration
FRET substrate	Fluorescence resonance energy transfer substrate
FT	Flow through
GA	Golgi apparatus
GAGs	Glycosaminoglycans
GdmCl	Guanidinium chloride

GF	Growth factor moiety
GFP	Green fluorescent protein
GNBP3	Gram-negative bacteria-binding protein 3 *(Drosophila melanogaster).*
GNPTAB	N-acetylglucosamine-1-phosphate transferase subunits alpha and beta
GPC	Envelope glycoprotein precursor
GRAB	G-related α2M binding protein from *Streptococcus pyogenes*
HCV	Hepatitis C virus
HEK 293	Human embryonic kidney 293 cells
H5	HighFive™
hr5	Homologous region 5
hSKI1	Human Subtilisin Kexin Isozyme-1
hα2M	Human α2-macroglobulin
Igκ	Immunoglobulin κ
IEC	Ion exchange chromatography
IMAC	Immobilized metal affinity chromatography
IPTG	Isopropyl-β-D-thiogalactopyranoside
I-α2M	Induced α2-macroglobulin
k_a	Association rate constant
k_d	Dissociation rate constant
K_D	Equilibrium dissociation constant
LAP	Latency-associated domain
LB	Lysogeny broth medium
LIC	Lepidopteran insect cells
LD	Lipid droplets
LDLR	Low-density lipoprotein receptor
LoVo	Human colon adenocarcinoma cells
LRP	Lipoprotein receptor-related protein

LTBP	Latent TGF-β binding protein
MA	Methylamine
MA- α2M	Methylamine induced α2-macroglobulin
mAb	Monoclonal antibody
MALDI-TOF-MS	Matrix-Assisted Laser Desorption/Ionization Time of Flight Mass Spectrometry
MALLS	Multiangle laser light scattering
MBTPS1	Membrane-bound transcription factor site-1 protease
MCS	Multicloning site
MG domain	Macroglobulin-like domain
M&M	Material and Methods
MMP	Matrix metalloproteinase
MP	Metallopeptidase
MT promoter	Metallothionein promoter
MW	Molecular weight
N-	Amino-
Ni-NTA	Nickel-nitrilotriacetic acid
N-hα2M	N-terminal recombinant protein of human α2-macroglobulin
NTS	N-terminal sub-domain
OmpA	Outer membrane protein A
O/N	Overnight
PACE4	Paired basic amino acid cleaving enzyme 4
PAMAM	Polyamidoamine
PBS	Phosphate buffer saline buffer
PC	Proprotein convertase
PCR	Polymer chain reaction
PCSK9	Proprotein convertase subtilisin kexin 9
PD	Prodomain

PDB	Protein data bank
pDNA	Plasmid deoxyribonucleic acid
PEI	Polyethylenimine
pI	Isoelectric point
PMF	Peptide mass fingerprint
pro-TGF-β2	Latent transforming growth factor-β2
PTM	Post-translational modification
PZP	Pregnancy zone protein
RAP	Receptor-associated protein
RBD	Receptor binding domain
r-BV	Recombinant Baculovirus
r.p.m.	Revolutions per minute
RER	Rough endoplasmic reticulum
R_{eq}	Response at equilibrium (steady state) $\mathbf{R_{max}}$
RCL	Reactive centre loop
$\mathbf{R_{max}}$	Maximum response
rmsd	Root-mean-square distance
RP-HPLC	Reversed phase high performance liquid chromatography
RT	Room temperature
RU	Response units
S1P	Site 1-protease
S2P	Site 2 protease
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
SEC	Size exclusion chromatography
Sf9 cells	*Spodoptera frugiperda* cells
Sf900 II	Sf-900™ II SFM medium
SGE	Stable gene expression

SKI1	Subtilisin/kexin-isozyme 1
SREBPs	Sterol regulatory element-binding proteins
SRD-12B cells	Chinese hamster cell cultures
SP	Signal peptide
SPR	Surface plasmon resonance
Sulfo-NHS-AMCA	Sulfosuccinimidyl-7-amino-4-methylcoumarin-3-acetate
S2 cells	*Drosophila melanogaster* Schneider 2 embryonic cells
TCA	Trichloroacetic acid
TCEP	Tris[2-carboxyethyl]phosphine or 1,4-dithiothreitol
TED	Thioester domain
TEP	Thioester containing protein
TEV	*Tobacco etch virus*
TIMP	Tissue inhibitor of metalloproteinase
TGE	Transient gene expression
TM	Transmembrane
TS50	TubeSpin bioreactor 50 tube
TS600	TubeSpin bioreactor 600 tube
UP	UniProt
WB	Western blot
WT	Wild type
λem	Emission wavelength
λex	Excitation wavelength
χ^2	The average deviation of the experimental data from the fitted
XIAP	X-linked inhibitor of apoptosis protein

The twenty physiological and proteinogenic amino acids with their respective three-letter and one-letter codes:

Alanine	**Ala**	**A**
Arginine	**Arg**	**R**
Asparagine	**Asn**	**N**
Aspartic acid	**Asp**	**D**
Cysteine	**Cys**	**C**
Glutamic acid	**Glu**	**E**
Glutamine	**Gln**	**Q**
Glycine	**Gly**	**G**
Histidine	**His**	**H**
Isoleucine	**Ile**	**I**
Leucine	**Leu**	**L**
Lysine	**Lys**	**K**
Methionine	**Met**	**M**
Phenylalanine	**Phe**	**F**
Proline	**Pro**	**P**
Serine	**Ser**	**S**
Threonine	**Thr**	**T**
Tryptophan	**Trp**	**W**
Tyrosine	**Tyr**	**Y**
Valine	**Val**	**V**

LIST OF PUBLICATIONS

- Goulas T, Garcia-Ferrer I, Marrero A, Marino-Puertas L, Duquerroy S, Gomis-Rüth FX. (2017). **Structural and functional insight into pan-endopeptidase inhibition by α2-macroglobulins**. *Biological Chemistry, 398*(9):975-994. doi: 10.1515/hsz-2016-0329.

- Marino-Puertas L., Goulas T and Gomis-Rüth FX. (2017). **Matrix metalloproteinases outside vertebrates.** *Biochimica Biophysica Acta Molecular Cell Research, 1864*(11 Pt A):2026-2035. doi: 10.1016/j.bbamcr.2017.04.003. Epub 2017 Apr 7. Review. First author.

- Del Amo-Maestro, L., Marino-Puertas, L., Goulas, T. & Gomis-Rüth, F. X. (2019). **Recombinant production, purification, crystallization, and structure analysis of human transforming growth factor β2 in a new conformation**. *Scientific reports, 9*(1), 8660. doi: 10.1038/s41598-019-44943-4. First co-author.

- Marino-Puertas, L., Del Amo-Maestro, L., Taulés, M., Gomis-Rüth, F. X., & Goulas, T. (2019). **Recombinant production of human α2-macroglobulin variants and interaction studies with recombinant G-related α2-macroglobulin binding protein and latent transforming growth factor-β2**. *Scientific reports, 9*(1), 9186. doi:10.1038/s41598-019-45712-z. First author.

Figure reproduced from “flirck.com”

INTRODUCTION

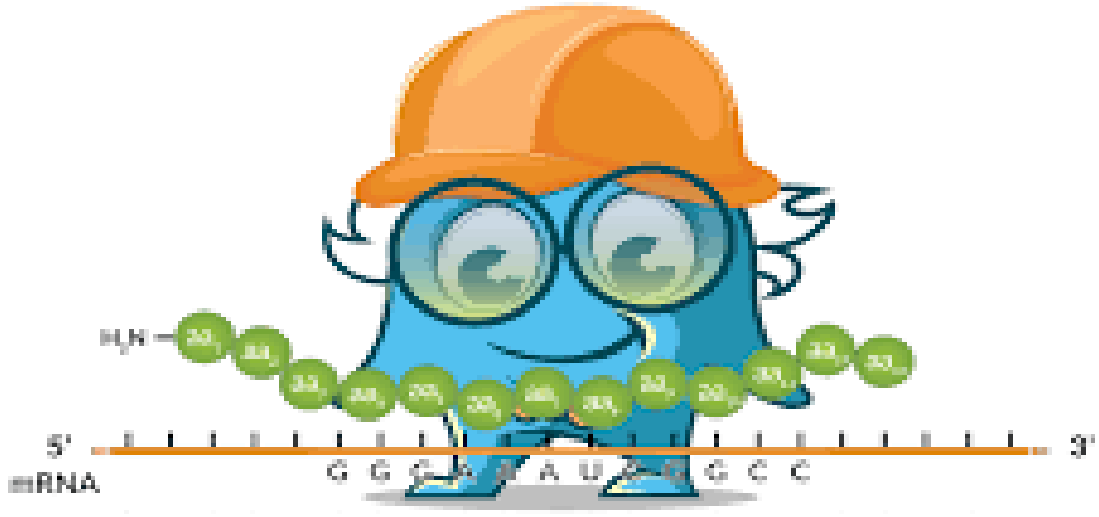


Figure reproduced from Bosterbio

INTRODUCTION

1.1 Technological advances on therapeutic recombinant protein production

1.1.1 Recombinant therapeutic protein industry

Since the approval of the first human recombinant protein (r-protein) in 1982 (Humulin, a recombinant human insulin), the r-protein industry has grown at a rapid pace (Pavlou and Reichert, 2004). Therapeutic r-proteins represent the core and fastest growing segment of the pharmaceutical biotechnological industry, with 18.2% in overall growth (Aggarwal, 2014). This industry reached a total of $99 billion in global sales in 2009 (Walsh, 2010), $120 billion in 2012, $150 billion in 2015 (Butler and Meneses-Acosta, 2012) and is predicted to reach $1.18 trillion in 2024, where immuno-oncology drugs are going to significantly contribute to this growth (www.evaluate.com/PharmaWorldPreview2019). 57 biopharmaceuticals gained approval between 2006 and 2010 (Walsh, 2010) and over 200 were approved in 2013 (Kling, 2014). With the emergence of biosimilars, the profitability and therapeutic potential of biotechnological drugs has increased (more than 20 biosimilars have been approved by the European Medicines Agency (EMA) since 2006 (Kling, 2014)) and protein-based drugs represented 8 of the top 10 best-selling drugs globally by the year 2012 (Lindsley, 2013).

Expression systems based on mammalian cells and *Escherichia coli* (*E. coli*) remain the major workhorses of biopharmaceutical production (Wash, 2010; Wurm, 2004). Prokaryotic cell hosts have been mainly used to produce low-complexity proteins, such as insulin or growth hormones (Wilson et al., 1990). However, for producing therapeutic proteins, animal cells are currently preferred over other cell types since "human-like" post-translational modifications (PTMs), such as complex glycosylation, are required in order to reduce the risk of immunogenicity (Frenzel et al., 2013).

They also offer specific chaperone capabilities and all the cellular machinery to achieve proteins with native-like activities. Since the first FDA-approved recombinant therapeutic proteins were produced in stable transfected Chinese hamster ovary (CHO) cells in 1987 (Activase-tPA), CHO cells have been the manufacturing host system of choice for more than 70% of therapeutic r-proteins on the market (De Jesus and Wurm, 2011). By 2010, more than 50% of the biopharmaceuticals on the market relied on mammalian cell-based processes. Of the top 10 sold drugs globally in 2018, eight were monoclonal antibodies (mAb) for cancer treatments or autoimmune disorders (Fig. 1A) (Urquhart, 2019). Humira (adalimumab) by AbbVie's is in the first place with US $20 billion in sales, the inhibitor Keytruda (pembrolizumab) by Merck is in the third with $7.2 billion sales, and Herceptin (trastuzumab), Avastin (bezacizumab) and Rituxan (rituximab) mAbs by Roche with a total of $21 billion in sales are in the fourth, fifth and sixth places, respectively. Thus, recombinant protein production, and especially mAb mainly generated in mammalian cells), are huge areas of interest within biopharmaceutical companies (Fig. 1B).

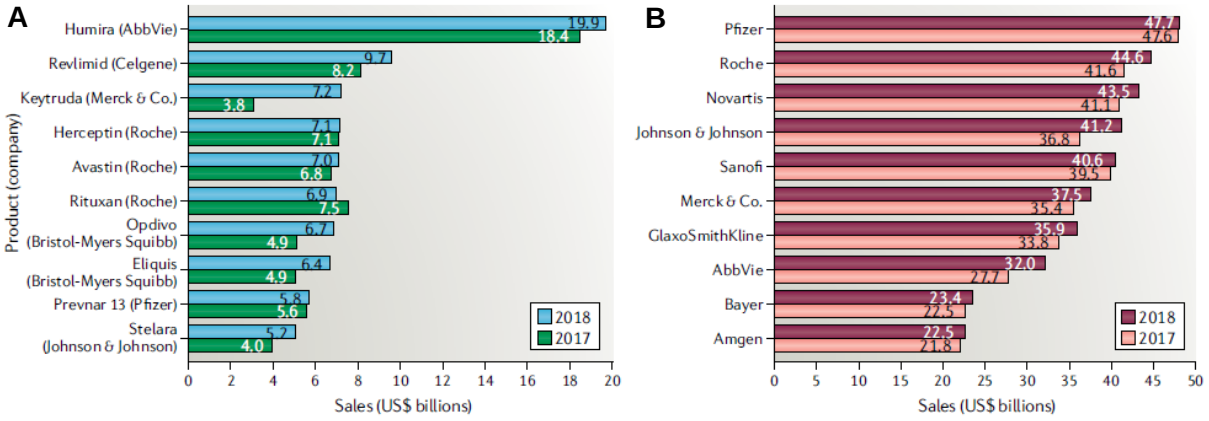


Figure 1. Top drugs and companies by sales in 2018 and 2017. (Figure duplicated from Urquhart, 2019). **(A)** Top ten drugs by sales globally. **(B)** | Top ten companies by sales of biopharmaceutical drugs.

1.1.2 Choice of cell hosts for recombinant protein production

There are important differences among host cells and the choice of expression system depends fundamentally on the function of the desired r-protein. A comparison of protein expressions using different cell hosts is summarized in Table 1 (Fernandez and Hoeffler, 1999).

The predominantly cell hosts used for human r-protein are insect and mammalian cells. Insect cells are the major animal cell hosts for small-scale protein production for exploratory research, drug discovery and structural biology (Hunt, 2005); whereas mammalian cells are the main host for expression and sale of human therapeutic proteins (Dawn M Ecker, 2015).

Table 1. Comparison of protein expression systems.

Cell host	***E. coli***	**Yeast**	**Insect cells**	**Mammalian cells**
Cell growth	Rapid (30 min)	Rapid (90 min)	Slow (18-24h)	Slow (24h)
Complexity of growth medium	Low	Low	High	High
Cost of growth medium	Low	Low	High	High
Protein folding	Refolding required	Refolding required	Proper folding	Proper folding
N-linked glycosylation	None	High mannose	Simple, no sialylation	Complex
O-linked glycosylation	No	Yes	Yes	Yes
Phosphorylation	No	Yes	Yes	Yes
Acetylation	No	Yes	Yes	Yes
Acylation	No	Yes	Yes	Yes
γ- carboxylation	No	Yes	No	Yes
Yield (mg/L)	50-500	10-200	10-200	0.1-100
Project cost	Low	Low	High	High

1.1.3 Protein expression in bacterial cells

The advantages of *E. coli* for protein production are many: it has extremely fast growth kinetics, high cell densities can be achieved, culture media and reagents are inexpensive and transformation with expression constructs is straightforward. For the whole biochemistry and biology fields, the easy

incorporation of labeled and non-natural amino acids (such as selenomethionine and selenocysteine for X-ray crystallography) into the proteins through expression cultures is a definitive advantage (Wang, 2009). It is therefore no surprise that most of the crystal structures deposited at the Protein Data Bank (PBD) come from material purified from overexpressing *E. coli* cells.

E. coli must always be used in a first attempt, as it provides a fast and cheap route to the expressed target if it works. Many strains and plasmids are available to introduce several PTMs, to enable the secretion of expressed proteins (especially when they are small), to assist the folding of disulphide-bridge containing proteins and to assemble coexpression constructs.

Some protein types are however known to be a challenge to *E. coli*, like protein kinases (which tend to kill growing *E. coli* cells), large eukaryotic multisubunit complexes (which typically lead to no expression or very poor yields) and membrane proteins, many of which require specific PTMs and lipid compositions.

Optimization of the expression construct is recommended and the strategies are very numerous, from the choice of a fusion tag and its placement (N- or C-terminal small tag or larger fusion proteins), to truncation (from either one or both ends), mutagenesis (to remove glycosylation sites), insertion of *Tobacco etch virus* (TEV) recognition site and for cleavage (to facilitate the purification), to use different signal peptides to guide the expression and special *E. coli* strains for expression of toxic proteins. Only when a reasonably large set of expression experiments has been performed without promising results is recommended to move on another expression host.

1.1.4 Protein expression in yeast cells

Yeasts share many useful features with *E. coli*, including fast growth rates, inexpensive culture media, many useful molecular biology tools (including, e.g. promoters, selection markers) and very accessible genomes to target genetic manipulation. These properties, combined with their eukaryotic nature, which allow them to perform many PTMs and to fold properly the proteins, make them a usual alternative after the *E. coli* expression system.

Many eukaryotic proteins (including human) and peptide hormones that are typically secreted can usually be produced using yeasts, including many protein factors from the immune complement system. There are two types of yeast: methylotrophic (able to use in methanol and other reduced one-carbon compounds as their main carbon source) and non-methylotrophic (Fernandez and Vega, 2013). Methylotrophic yeasts were discovered in the late 60s and early 70s (Kato et al., 1974; Ogata et al., 1969). Non-methylotrophic yeasts include *Saccharomyces cerevisiae*, *Kluyveromyces lactis* and

Yarrowia lipolytica. They have very well-known genetics and metabolism, so they are usually used to engineer rapid protein production screenings and high-throughput settings. In contrast, methylotrophic yeasts (mainly *Komagataella*, also termed *Pichia pastoris,* and *Ogataea polymorpha*) grow to higher densities and typically produce greater yields for most proteins (Cregg, 2007). The two types of yeasts can fortunately be made into a powerful combination. Non- methylotrophs can be used for fast screening and optimization purposes and methylotrophs can be employed mostly for the generation of overproducing strains and to increase protein product yields.

K. pastoris, in particular, is the preferred yeast for recombinant overexpression of proteins and has been extensively used for the production of heterologous proteins in general (Grishaev, 2008). It has also been extensively used for applications in food and feed industries (Grishaev, 2008) in addition to be used for the production of labeled proteins to solve atomic structures (Basak, 2003; Larsson et al., 2002). The fact that it grows in mineral defined (minimal) medium, makes it easy to introduce labeled amino acids into the medium.

However, many eukaryotic proteins and protein complexes have special requirements in terms of PTMs, chaperone assistance and folding properties, which require insect or mammalian cells for their correct folding, processing and assembly.

1.1.5 Protein expression in insect cells

Insect expression systems represent an adequate compromise between bacterial and mammalian systems (Becker-Pauly and Stöcker, 2011). In insect cells, signal peptides are cleaved as in mammalian cells, disulphide bonds are formed in the endoplasmic reticulum and proprotein-converting enzymes are available for proteolytic processing. Insect cell cultures are less demanding than mammalian cells under standard laboratory conditions, they grow at lower pH (6.2-6.7), and there is generally no need for a CO_2 atmosphere. Also, baculovirus infection does not increase the biosafety level *per se*.

The first report on the production of a human r-protein in insect cells was published in 1983 (Smith et al., 1983). They expressed human interferon in butterfly cells with the help of insect pathogenic viruses and established the baculovirus expression system.

Many other human r-proteins have been expressed in insect cells, like tissue factor and antibodies (Kost et al., 1997; Masroori et al., 2010), but one important focus has been the development of subunit vaccines, including the human papillomavirus VLP vaccine against cervical cancer (Cervarix^{TM}), the vaccine against seasonal influenza (Flublok^{R}) and the recombinant virus compensation for lipoprotein lipase deficiency (Glybera^{R}).

The most common established insect cell lines are derived from pupal ovarian tissue of the fall armyworm *Spodoptera frugiperda* (Sf9, Sf21) (Vaughn et al., 1977), the embryos of *Trichoplusia ni* (Tn5 or HighFive™) and the late embryonic stages of *Drosophila melanogaster* (S2) cells (Schneider, 1972). These cells can be used for virus production (titration) or, after infection with baculovirus or transient/stable transfection with suitable plasmids, for r-protein production.

1.1.5.1 Baculovirus expression vector system technology platform

The baculovirus expression vector system (BEVS) has emerged as a reliable technology for r-protein production in insect cells over the last two decades (Smith et al., 1983) . There are two main systems to generate recombinant baculoviral particles: homologous recombination or specific site transposition. In the first system, the target protein is cloned in a transfer vector (under control of a viral promoter) flanked by homologous regions to the viral genome. Co-transfection of insect cells with transfer vector and linearized baculoviral DNA (usually from baculovirus *Autographa californica* multicapsid nucleopolyhedrovirus (AcMNPV)) induces the generation of recombinant viruses expressing the target protein, that are selected afterwards. In the second system, much faster and widely used, transposition at Tn7 sites is the method to obtain recombinant virus. The target gene is cloned into a transfer vector (pFASTBAC™) with mini-Tn7 flanking sites. A specific bacterial strain (DH10Bac™), containing a helper vector (coding for transposases) and a subgenomic DNA from baculovirus that contains mini-attTn7 regions (called bacmid or BaculoGold™ DNA), is transformed with the transfer vector and transposition of the target gene onto the bacmid takes place. Colonies containing recombinant bacmids are selected with antibiotics and identified with white/blue screening (due to disruption of LacZ-alpha gene by the transposition event). From these colonies, bacmid DNA is prepped and used to transfect insect cells to generate the baculovirus. In both systems, the recombinant viral DNA within the insect cells is able to produce viral particles that are competent for re-infection. After virus titration, r-protein expression is achieved by infecting fresh cultures of lepidopteran insect cells (LIC) at the appropriate cell densities with the recombinant Baculovirus (r-BVs).

Compared to other systems, the LIC-BEVS offers several advantages. R-BV vectors allow the expression of single r-proteins over 20 kDa and multimeric protein assemblies over 1 million Daltons. R-BVs are relatively stable, and each virus stock can be amplified several times. A volumetric yield of hundreds of milligrams of r-protein per liter, in suspension cell cultures, can often be achieved, but yields can vary from protein to protein (Harrison and Jarvis, 2007a; Harrison and Jarvis, 2007b). Gene

intron splicing, or post-translational modifications like peptide cleavage, phosphorylations, glycosylations and acylations are achieved via r-BV mediated expression in insect cells.

Nevertheless, the BEVS has several limitations. One of the main disadvantages is that the yields of most membrane-bound and secreted proteins are significantly lower than the yields obtained with intracellular proteins (Harrison and Jarvis, 2007a). Another disadvantage is that Baculovirus is destroying the cells and cell lysis releases hydrolytic enzymes, like proteases, into the medium. This might hydrolyse the product and compromise the final yield. r-BVs are also known to be unable to replicate in mammalian cells, but they can however enter mammalian cells in culture and deliver DNA sequences leading to a potential biosafety concern. The last important disadvantage is they are highly time and space consuming. This is because for a single r-protein production is necessary the amplification and selection of recombinant virus, their storage, and then finally their use for infecting the LIC system.

1.1.5.2 Stable gene expression in insect cells

Stable gene expression (SGE) involves a stable transfection of insect cell lines with a plasmid encoding a protein of interest under the control of a constitutively or inducible active promoter. In the SGE system, r-proteins are expressed by recombinant cell lines in which the transgenes have been stably integrated into the genome of the host cells. Cells that have undergone gene integrations are genetically selected from transfected cell pools using a selection marker encoded on the transfected plasmid. The clonal cell lines with the desired phenotype for protein expression and cell growth are then used for protein production (Moraes at al., 2012).

Cell lines that are used for this system include: Sf9, Sf21, H5, S2 and S3 cells. Different promoters can be used in these cells, like the strong constitutive actin 5C promoter (Krasnow et al., 1989; Winslow et al., 1989) or the inducible (e.g. by addition of $CuSO_4$) metallothionein promoter (MT) (Bunch and Goldstein, 1989; Bunch et al., 1988). With this system, cells are not lysed being better for secreted r-proteins and sustained r-protein expression can be also achieved. In addition, SGE cell cultures are reproducible with homogeneous product and protein expression, permitting an easier production scale-up. However, the use of stable insect cell lines in a non-lytic system is not preferred by the industry, because these insect cell lines normally express less protein compared to BEVS. They also require a long development time for the establishment of stably transfected lines (around 3 weeks) and each recombinant cell line must be stored separately.

1.1.5.3 Non-viral transient gene expression

Transient gene expression (TGE) allows for quick production of a desired r-protein within days or weeks. Because the entry of naked plasmid DNA into cells is not very efficient, various physical means (e.g. microinjection and electroporation) and chemical agents (e.g. calcium phosphate, cationic lipids and cationic polymers) have been developed to facilitate effective gene delivery. Among these, 25 kDa linear polyethylenimine (PEI) (the main cationic polymer) is one of the most widely used agents for DNA delivery (Baldi et al., 2007). It has also been shown to transfect adherent Sf9 and S2 cells (Ogay et al., 2006; Patino et al., 2014). In addition to PEI, other widely used polymers for gene delivery are polyamidoamine (PAMAM) dendrimers and polylysine, but have rarely been tested on insect cell lines (Pack et al., 2005).

1.1.5.4 Glycosylation differences between insect and mammalian cells

Protein glycosylation is influenced by host-cell type and the presence of signal sequences on the protein. Different host systems can express different types of glycosylation enzymes and transporters, contributing to the specificity and heterogeneity in glycosylation profiles and subsequent clinical effectiveness of the therapeutic product. Most human proteins are glycosylated, and they have one or more oligosaccharide chains (glycans) attached to their polypeptide backbone. They are composed of combinations of seven monosaccharides: glucose (Glc), galactose (Gal), mannose (Man), fucose (Fuc), N-acetylglucosamine (GlcNAc), N-acetylgalactosamine (GalNAc) and sialic or neuraminic acids (SA). Glycans attached to the amide nitrogen atom of asparagine (Asn) residues are termed N-linked glycans, while glycans attached to the oxygen atom of serine (Ser) or threonine (Thr) residues are O-linked glycans. The possible variations in monosaccharide composition, glycosidic linkages and glycan branching gives rise to many possible glycan structures (Brooks, 2004).

The initial steps of N-linked glycosylation are the same for all eukaryotes, starting with the co-translational transport of glycoproteins to the rough endoplasmic reticulum (RER). In the RER, the same glycosylation sites on the nascent peptide-chain are used throughout eukaryotes and a branched oligosaccharide is added to the asparagine residue in this motif (Aebi and Burda, 1999). The oligosaccharide in the ER and Golgi apparatus (GA) is then trimmed in a similar way in eukaryotes (Kubelka et al., 1994). Next, new N-linked carbohydrate side chains are added, resulting in glycoproteins with simple oligo-mannose sugar chains in insects (they are also called pauci-mannose),

while mammals have complex sugar groups with terminal galactose and/or sialic acids (Betenbaugh et al., 2004; Kost et al., 2005; Van Oers, 2011) (Fig. 2).

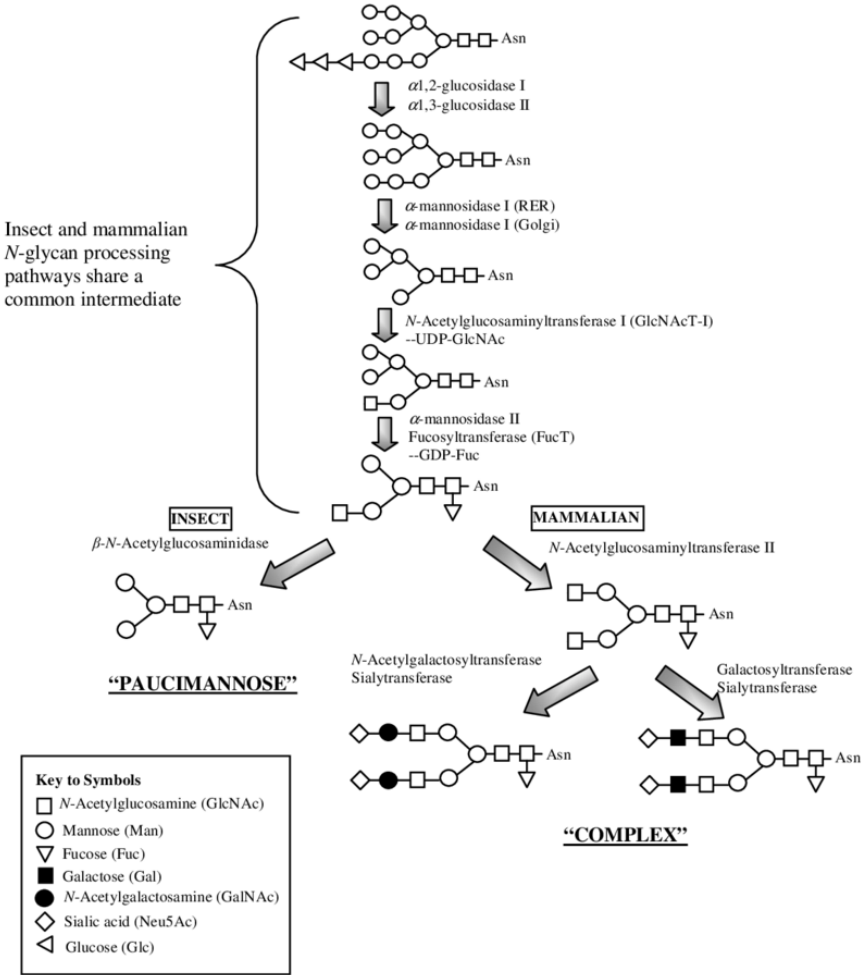


Figure 2. N-glycosylation differences in insect and mammalian cells (Figure reproduced from Gillich, 2005).

1.1.6 Protein expression in mammalian cells

Expression of proteins in mammalian cells is an important resource for many studies on human genes. These studies include the mapping of protein interactions, solving protein structures (for example, by X-ray crystallography, solution phase NMR or cryogenic electron microscopy) and generation of human proteins, including antibodies. It is always most appropriate to use a mammalian expression system that offers the machinery, such as chaperones, binding partners, secretion apparatus and post-translational modifications for correct protein folding and activity of human r-proteins. Although, successes have been achieved by the use of *E. coli* (Dyson et al., 2004), yeast or insect cells expression systems, however, especially for secreted or membrane containing proteins,

the non-mammalian expression systems lack the appropriate machinery for complex glycosylation (Byrne et al., 2007; Marchal at al., 2001; Nallet et al., 2012).

For a long time, expression of proteins in mammalian cells was too expensive, requiring special equipment and staff. It also gave a poor yield for transient transfection or took too long to isolate a stable, high-producing cell line. During the last years a series of publications has luckily emerged and gave the correct combination of cell line, transfection method, expression vector and cultivation conditions, making it possible for the scientific community to achieve excellent yields of protein in a relatively quick and moderately affordable procedure.

The main expression cell lines are human embryonic kidney (HEK) 293 cells and Chinese hamster ovary (CHO) cells. Originally the HEK293 cell line was established by the transformation of human embryonic kidney cells with sheared human adenovirus DNA, resulting in expression of the adenovirus E1A and E1B genes (Graham et al., 1977). Subsequently the HEK293E (Yates at al., 1985) and HEK293T (Rio et al., 1985) cell lines were isolated by the integration of genes encoding the Epstein-Barr Virus nuclear antigen 1 (EBNA1) or simian virus 40 (SV40) large T antigen (LT) respectively. The function of EBNA1 and SV40 LT is to bind to their specific origins of replication (ori) and promote DNA replication by the recruitment of the DNA replication machinery and potentially increase protein expression yield. Suspension adapted cell lines able to grow in serum free conditions were generated (FreeStyle293F™ and Expi293F™ cells) (Liu et al., 2008) to enable high density cell growth and this way, increase protein expression yield per unit volume of cell culture media with no viral elements.

CHO cells were originally isolated from spontaneously immortalized cells of primary Chinese hamster ovarian cultures (Puck et al., 1958). CHO-K1 (Kao and Puck, 1968) and CHO-S (Gottesman, 1987) were derived from the original CHO cell line. A CHO-K1 derivative cell line has also been constructed with integrated EBNA1 and glutamine synthase genes for enhanced transgene expression (Daramola et al., 2014). A CHO cell line adapted for suspension growth in serum free conditions, designated CHO-S cells, has been also generated. In addition, the CHO-DG44 cell line (Urlaub et al., 1983) is used for the generation of stably transfected cell lines (they lost dihydrofolate reductase (DHFR activity), which is recovered with the transfection of a plasmid with that cassette and the gene of interest).

1.1.6.1 Transient and stable gene expression

Protein expression in mammalian cells is divided into transient and stable gene expression. Both modes start with the introduction of plasmid DNA into the cells. For transient expression, improving yield is governed by improving the transfection step (increasing the percentage of cells that are transfected with high biomass, decreasing the cost of the transfection reagent and obtaining high quality and high transfectable DNA). For stable expression, a highly efficient transfection is not as important, because the successfully transfected cells with the target of interest and the gene resistance are going to be selected for weeks until achieving a full cell line stably transfected.

Efficient transfection can be achieved with cationic lipid formulations for mammalian cells using reagents such as FuGene HD (Roche) or Lipofectamine™ 2000 (Invitrogen) (Geisse and Voedisch, 2012). These methods are however not scalable due to their cost. Calcium phosphate has been used successfully to transfect HEK293 and CHO cells (Jordan et al.,1996), but requires the presence of serum which is not compatible with many serum free media compositions normally used. PEI is another common transfection reagent used (Hacker et al., 2013; Rajendra et al., 2011), due to its low cost and high transfection efficiency. Among all the different types of transfection reagents, the linear 25 KDa PEI is the most reported for large-scale protein production (Backliwal et al., 2008).

The majority of small to medium scale expression experiments can be performed by transient gene expression protocols and are usually performed in HEK293 cells. This is because they adapt well to suspension culture, are highly transfectable, express well, and have relatively slow growth rates (that is important in order not to lose the protein expression by dilution of the transfected genes). However, when bigger quantities of protein is required, it can be tedious to produce the large amounts of plasmid DNA required for large scale transfection, and thus it can be more convenient to isolate a stable cell line (of monoclonal or polyclonal origin) expressing the gene of interest (Büssow, 2015). Stable cell line development involves transfection with linear DNA encoding a gene of interest and a selectable marker (that confers a drug resistance, used in the cell selection: NeoR for G418, DHFR for metrotrexate). CHO cells have been traditionally used for stable gene expression, as opposed to HEK293, because they also adapt well to suspension culture, and despite they yield lower transfection efficiencies, they grow much more rapidly, with doubling times of 16–19 h, compared to the 23–29 h exhibited by HEK293. As a consequence, CHO cells are the usual model for therapeutic large-scale bioprotein production. Biopharmaceutical companies have used and standardized the protein production for CHO cells , one of the best hosts for human r-proteins, avoiding DNA cross contamination. However, there are glycosylation differences between HEK293 and CHO cells, which give different features to the r-proteins, which have to be considered.

1.1.6.2 Glycosylation differences among mammalian cell lines

Nowadays, most of the therapeutic proteins are produced in mammalian expression systems. These r-proteins are similar, but not identical. This is because differences in glycan structures have been detected depending to the used mammalian cell line (Gaudry et al., 2008; Van den Nieuwenhof et al., 2000). There are also reports about culture conditions, that can affect the glycosylation pattern (temperature, media composition…). Apparently, changes in culture conditions could modify the distribution of the isoforms (Schiestl et al., 2011).

As mentioned, the usual mammalian cell lines used for r-protein expression are HEK and CHO cells, one from human and the other from hamster and from kidney and ovary cells, respectively. These two cell lines represent two different species and tissues and it is known that glycosylation is depending on the cell type and tissue where the protein is expressed (Brooks, 2004; Brooks, 2006). Most r-proteins are not expressed in their native cell types, and thus changes in the glycosylation pattern compared to the native protein could affect the protein behaviour. For instance, CHO cells do not express β-galactoside α2,6-sialyltransferase (Gal α2,6 ST), α1,3/4 fucosyltransferase or β-1,4-N-acetylglucosaminyltransferase III (GnT-III) which are enzymes expressed in HEK cells (Campbell and Stanley, 1984; Sallustio and Stanley, 1989; Stanley et al., 2005). Therefore, recombinant proteins from CHO cells may lack or have different glycosyl structures.

The sialic acid content has been shown to influence protein activity *in vitro* and it is known to influence the biological half-life of proteins *in vivo* (Lee et al., 2002; Wide et al., 2010). Proteins that lack terminal sialic acids are bound to the receptor in the liver and are eliminated, reducing their half-life (Ashwell and Harford, 1982). In general, the proteins expressed in CHO cells had more acidic isoforms compared to the HEK derived protein, which could favour the *in vivo* half-life of proteins produced in CHO compared to HEK cells. However, this is not an absolute rule. This also must be considered for the purification, when ionic exchange chromatography step is included, because this technique is based on the charge of the protein and the total charge of the protein is modified with the different glycosylation distribution.

To summarize, the major differences in the glycosylation profile between the CHO and HEK cell lines are the size and number of glycosylations, together with the amount of sialic acids, which make the protein population more or less acidic. This could change their protein properties, half-life and influence protein interactions (Rutishauser, 1996). Giving differences in molecular weight, mass peak profile and isoelectric point (pI). Glycoproteins expressed in HEK cells generally show more complex glycosylation profiles than in CHO cells (Croset et al., 2012; Uchida et al., 1979) and on the other hand,

the CHO derived proteins are usually more sialyladed than HEK derived proteins (Croset et al., 2012; Uchida et al., 1979).

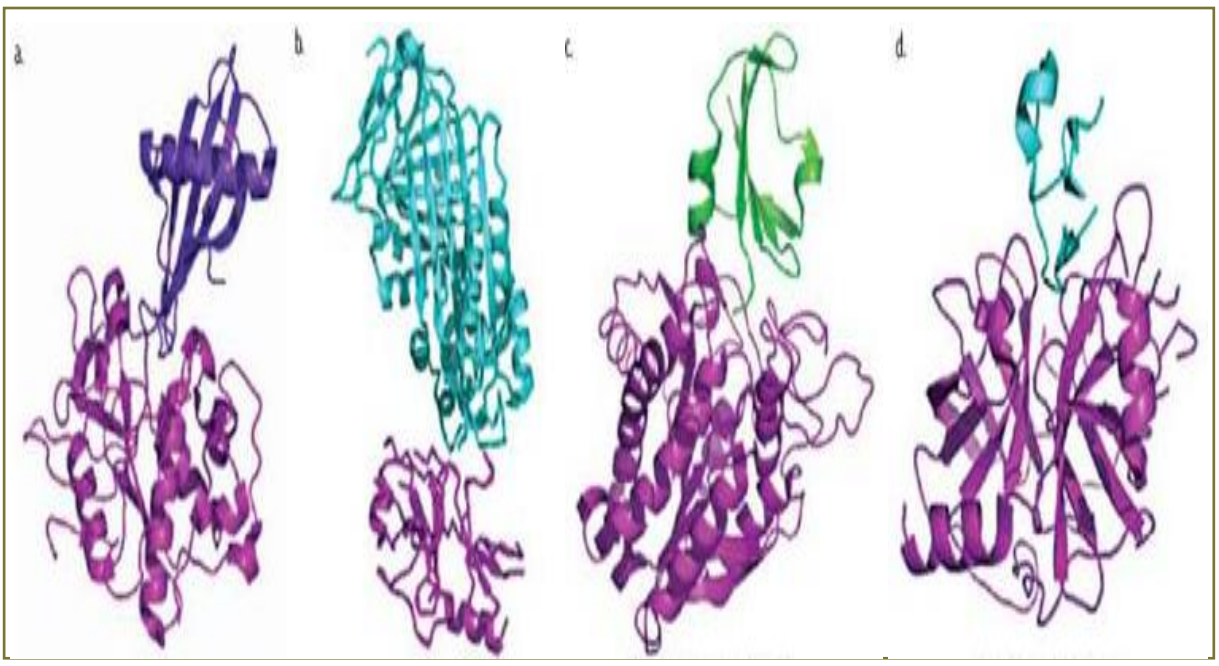


Figure reproduced from “Revista cubana de Ciencias Biológicas”

1.2 Proteolytic enzymes and their control

1.2.1 Proteolysis as essential process

Proteolysis, or protein degradation is a series of processes that result in the hydrolysis of one or more peptide bonds in a protein. This can be through catalysis by proteolytic enzymes called peptidases or nonenzymatically, for instance at very low or very high pH. Proteolysis is a part of protein turnover in living organisms (Varshavsky, 2001) (Fig. 3).

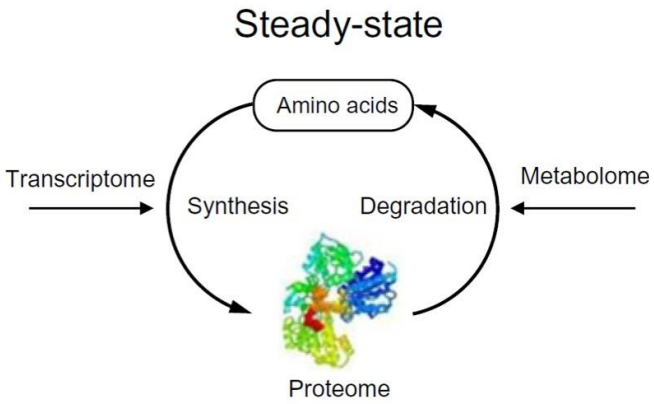


Figure 3. Steady state in protein turnover (Figure adapted from Chou et al., 2012). Protein synthesis equalling degradation with neither the related transcriptome nor the metabolome affected.

Complete degradation of a protein to its constituent amino acids, by the action of peptidases, allows these amino acids to be reutilized. For example, proteins from the diet are hydrolysed to amino acids or short peptides in the gastrointestinal tract by peptidases like trypsin and pepsin (Peters, 1970).

Intracellular proteins can be selectively destroyed through proteolysis with the ubiquitin system (also called the ubiquitin-proteasome system) in the cytosol and nucleus. Intracellular proteins can also be degraded by lysosomal proteases and other hydrolytic enzymes of the lysosome (Schmidt and Finley, 2014; Turk and Turk, 2009).

One major role of the cellular proteolytic systems is the detection and elimination of damaged (misfolded or aggregated) or abnormal proteins. The damaged proteins are potentially toxic to the cell because they might interact with inappropriate ligands or, if aggregated, become mechanical impediments to the normal cellular processes. These proteins are recognized and destroyed by the ubiquitin system (Varshavsky, 2001).

Proteolysis can involve hydrolysis of most of the protein's peptide bonds or only some of them. Site-specific proteolysis underlies a great variety of biological processes, like the modification of newly formed proteins in preparation for their function inside or outside the cell. For example, many

hormones are produced from a precursor, and at a appropriate time and intracellular compartment, are cleaved and activated by proteases.

Site-specific proteolysis is also a frequent feature of extracellular regulatory systems, as the clotting of blood (Green, 2006), which is mediated by a complex cascade of proteins (proteases and their protein inhibitors).

Proteolytic imbalance may lead to the malfunction of individual cells or whole organisms, causing inflammation, neurodegenerative diseases, cancer or cell death among other disorders (López and Bond, 2008; Yadav et al., 2014). Moreover, some peptidases are important in host-pathogen interaction, being either virulence factors or involved in defence mechanisms (Amstrong, 2006; Urban, 2009). In this context, mechanisms that regulate peptidase activity are crucial to ensure that the proper substrates are processed at the right time and place.

Furthermore, proteolytic enzymes are used as important reagents in laboratory, clinical, and industrial processes (detergent, food, dairy, leather, baking, brewing, meat tenderization, photography and management of waste industries) (Kirti et al., 2012). Proteases represent one of the three largest groups of industrial enzymes and account for about 60% of the total worldwide sale of enzymes (Rao et al., 1998).

1.2.2 Proteases, proteinases and peptidases

Proteolytic enzymes are ubiquitous in all organisms and constitute 2–4% of the encoded gene products (Farady and Craik, 2010).

The term “protease” was established by Grassman and Dyckerhoff for enzymes that degrade proteins by hydrolysis of peptide bonds, and they recognized that there are two very different types of these enzymes (Grassman and Dyckerhoff, 1928). They noted that some proteolytic enzymes act best on intact proteins, whereas others show a preference for small peptides as substrates. The result of that was the creation of the term “proteinase” for proteases that show specificity for intact proteins, and the term “peptidase” for proteases that show preference for small peptides (Grassmann and Dyckerhoff, 1928). However, the nomenclature peptidase is used in a more general context to refer to proteases in general.

The nomenclature of cleavage site positions of the protein substrate by proteolytic enzymes was formulated by Schechter and Berger (Schechter and Berger, 1967). The binding region of the active site is divided into subsites (S), or specificity sites, per amino acid residue or blocking group of the substrate. The subsites on each side of the catalytic site are designated S1 and S1´. The rest of the

subsites are numbered from this point in both directions as S1, S2..., etc., and S1´, S2´..., etc. The positions of the amino acids in the substrate that interact with the protease are numbered to correspond with the subsites they occupy on the active site and are designated P1 and P1´. Cleavage is catalyzed between P1 and P1´ (Farady and Craik, 2010) (Fig. 4).

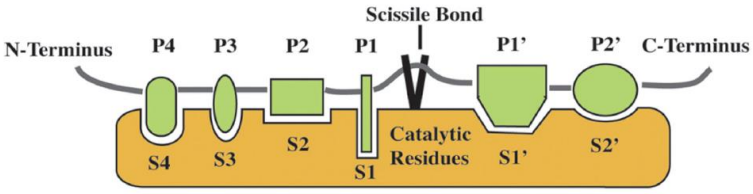


Figure 4. Schematic representation of a protease active site and nomenclature (Figure reproduced from Farady and Craik, 2010).

1.2.2.1 Classification of proteases

Peptidases can be classified according to four different criteria (Fig.5):

1.2.2.1.1 Specificity

Proteases were initially divided on the basis of the position of the peptide bond hydrolysis in two groups: “endopeptidases” (they cleave internal peptide bonds) and “exopeptidases” (they specifically cleave at the N- or C-terminus of a polypeptide) (Barrett and McDonald, 1985; Bergmann and Ross, 1936). Exopeptidases were further separated into “aminopeptidases” and “carboxypeptidases” depending on whether they cleave off N- or C-terminal residues (Barrett et al., 2013).

1.2.2.1.2 Catalytic mechanism

Peptidases can also be classified according to their active site configuration and catalysis mechanism. These include: serine, threonine, cysteine, aspartic, glutamic and metalloproteases. In the case of the three first mentioned proteases, the nucleophile group that initiates the attack on the peptide bond is the hydroxyl on the side chain of a serine or threonine, or the thiol on the side chain of a cysteine located in the active side of the protease (=”protein nucleophiles”) (Rawlings and Barrett, 2013; Seemuller et al., 1995). The last three proteases use a water molecule as a nucleophile for catalysis, which is activated either by the side chain of an aspartate residue, a glutamic acid residue or a coordinated metal ion, respectively (Fujinaga et al., 2004). In 2011, a new type of proteolytic enzymes was described, the asparagine proteases. They employ an asparagine as nucleophile, but they act with a mechanism that cannot be classified as hydrolysis, thus being lyases (Rawlings et al., 2011).

The most common classes of proteases are (in this order): serine, metallo and cysteine proteases (Rawlings et al., 2014).

1.2.2.1.3 Catalysed reaction and substrate

Another classification of peptidases was proposed by the Enzyme Commission (EC). They classified all kinds of enzymes based on the type of reaction they catalyze and assigned a unique EC number composed of three digits separated by dots. According to this, peptidases are a subclass within hydrolases (EC 3 hydrolases; Subclass 3.4) and they are sorted into 19 sub-subclasses to describe substrate specificity. This database can be accessed at the following website: https://www.qmul.ac.uk/sbcs/iubmb/enzyme/EC3/4/.

1.2.2.1.4 Evolutionary relationship

Rawlings and Barrett proposed an evolutionary scheme for proteases based on amino acid sequence data (Rawlings and Barrett, 1993). In this system, proteases/peptidases (they do not make any difference) are classified into families and clans. The term “family” is used to describe a group of enzymes which shows sequence homology on the part of the molecule bearing the active-site residues (the “peptidase unit”). Whereas the term “clan” comprises a group of families which shows evolutionary relationship, despite the lack of statistically significant similarities in sequence.

Families are named with a letter denoting the catalytic type (S, C, T, A, G, M, P, N, U or X, for serine, cysteine, threonine, aspartic, glutamic, metallo, mixed catalytic type, asparagine lyases, unknown catalytic type or compound peptidase, respectively), The second letter is assigned sequentially as each clan is identified. An example of a clan identifier is CA, which includes cysteine peptidases with a papain-like fold (Rawlings et al., 2018). The peptidase number assigned by the MEROPS database can be found in the following website: http://www.ebi.ac.uk/merops/. In addition, the MEROPS database include a collection of known cleavage sites in substrates, including proteins, peptides and synthetic substrates (Rawlings, 2009). They are currently (on April 26, 2019), over 270 families and 50 clans of peptidases that have been described, comprising more than 900,000 sequences (Rawlings et al., 2018).

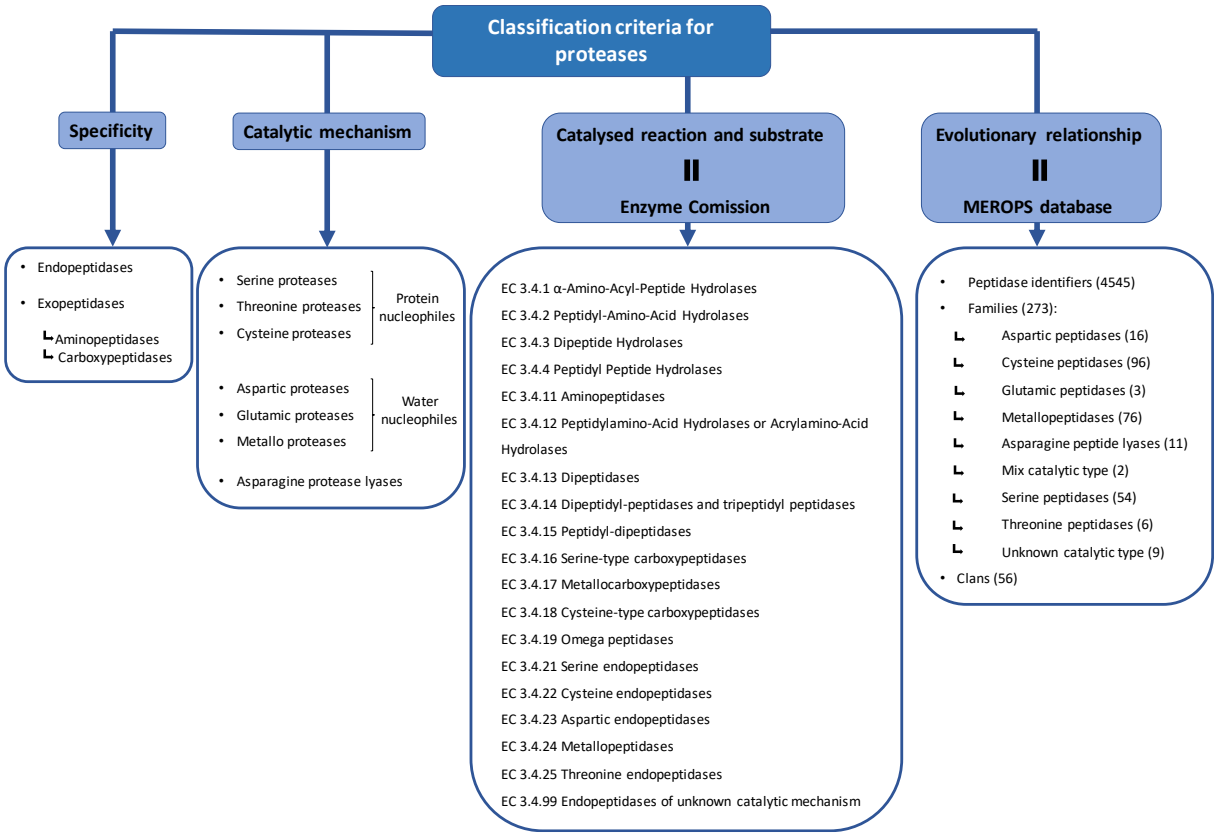


Figure 5. Classification systems for proteases (Figure adapted from Garcia-Ferrer, 2015). Scheme of the four classification criteria applied to proteases. MEROPS database was updated on April 26, 2019.

1.2.3 Mechanisms of proteolytic regulation

The action of proteases is tightly controlled to prevent indiscriminate peptidase activity. A wide variety of regulatory mechanisms are used, from the transcriptional level by differential expression, to the protein level *via* activation of inactive zymogens, competitive binding of inhibitors, allosteric or competitive binding to regulators and cofactors, or protein compartmentalization (Fig. 6).

1.2.3.1 Regulation at the transcriptional level

The amount of transcripts coding for a peptidase depends on the gene-copy number. The transcription can be modulated by molecular sensors and signalling proteins that are able to detect specific cellular or environmental conditions and activate its transcription. An example of this regulation is the mammalian MMP (matrix metalloproteinase) expression, which is induced in tumour and stromal cells by cytokines, chemokines, growth factors and oncogene products (Overall and López-Otín, 2002).

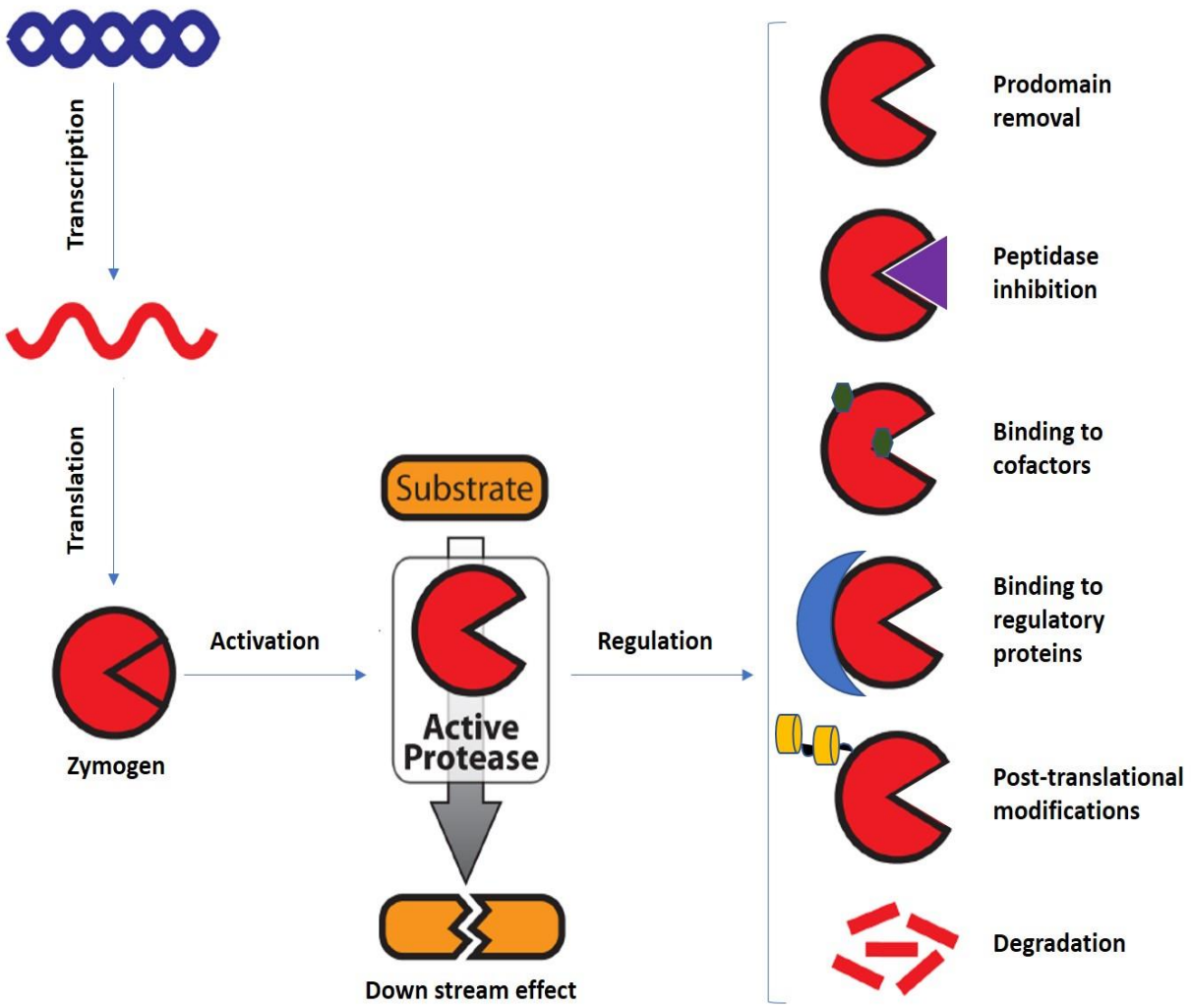


Figure 6 Mechanisms of proteolytic regulation (Figure adapted from Deu et al., 2013).

1.2.3.2 Zymogenicity and other post-translational modifications

The most widely spread mechanism of proteolytic regulation is the synthesis of proteolytic enzymes as inactive precursors, or "zymogens," in order to prevent unwanted protein degradation and as well, to enable spatial and temporal regulation of proteolytic activity.

The activation of zymogens usually occurs by proteolytic cleavage of a peptide bond, which is normally placed into the N-terminal prosegment (also called "prodomain", PD) of the protein. The PD sterically blocks the active site and, thereby prevents the binding of substrates. The PD removal by proteolytic cleavage is a process called "maturation" and can be spontaneous (by a simple change in pH), autolytic or heterolytic (for instance, catalyzed by another protease). These PDs often guide the folding process of the rest of peptidase domains, acting as intramolecular chaperones (Bryan, 2002; Eder and Fersht, 1995).

Post-translational modifications are also able to regulate protease activity, like phosphorylation, acetylation, myristoylation or ubiquitination (Schmidt and Finley, 2014). Changes in red-ox state, like oxidation of catalytic cysteines can also influence proteolytic activity.

Another important post-translational modification is oligomerization, which can promote peptidase activation, inhibition or substrate binding. A good example are the caspases, which need homodimerization to trigger activation (Pop and Salvesen, 2009).

1.2.3.3 Peptidase inhibitors

Peptidase inhibitors are the major regulators of protease activity. They inhibit peptidases by multiple mechanisms, ranging from irreversible, covalent binding (shown in suicide inhibitors like serpins), to competitive and/or reversible binding (Otlewski et al., 2005). Also, some inhibitors can act as high affinity transition state analogues.

Peptidase inhibitors can be proteins or small molecules, synthesized by bacteria, fungi and plants (Sabotic and Kos, 2012). They can be classified into endogenous or exogenous inhibitors. Until now, only 105 endogenous human inhibitors have been identified, which is substantially less than the number of proteases identified. This might be explained by one inhibitor being able to inhibit several proteases because of their low specificity for their target proteases. For example, there are more than 180 metallo-endopeptidases identified, inhibited for only four proteins (tissue inhibitor of metalloproteinase 1, 2, 3 and 4; TIMP1, TIMP2, TIMP3 and TIMP4).

1.2.3.4 Cofactors

Protease activity is also regulated by cofactors, molecules that bind to proteases or inhibitors and affect their final activity, often in an allosteric and reversible manner. The most well-known examples are in the blood coagulation cascade, with the tissue factor protein regulation of Factor VIIa activity (Eigenbrot and Kirchhofer, 2002).

1.2.3.5 Regulatory proteins

Regulatory subunits are proteins that usually bind the catalytic domain, affecting their activity (López-Otín and Bond, 2008). This regulatory mechanism is mainly found in multisubunit proteolytic assemblies like the 26S proteasome.

Some proteins termed “activators” allow with their association, the cleavage of the substrates (Schmidt and Finley, 2014). Other proteins called “adaptors”, bind the catalytic subunit and the substrate, facilitating the encounter and recognition (Battesti and Gottesman, 2013). In other cases, the regulation is not executed by independent proteins, but by regulatory domains, present within the peptidase polypeptide chain. A good example of this regulation is the ancillary domains of ADAMT peptidases (a desintegrin-like and metallopeptidase with thrombospondin type 1 motif) (Apte, 2009).

1.2.3.6 Allosteric regulators

Allosteric regulators outside the enzyme’s active site. They are therefore not involved in substrate recognition, but they induce conformational changes on the former, thus altering its function.

A good example are the glycosaminoglycans (GAGs), which are extremely important allosteric regulators of proteases and their physiological inhibitors. Their best characterized interaction is with antithrombin, the major inhibitor of several blood coagulation proteases (Olson et al., 2002). One of the best characterized GAGs is heparin and its derivatives and they are often referred to as indirect and allosteric inhibitors of blood coagulation proteases, primarily Factor X.

1.2.3.7 Protein compartmentalization and peptidase trafficking

As already mentioned, proteolysis must be highly regulated and accurately separated in space and time. Compartmentalization and peptidase trafficking between subcellular compartments with specific enzymes and biochemical conditions promote optimal peptidase activity. These biochemical conditions are: calcium concentrations, redox balance (from oxidizing, in the compartments of the secretory pathway and extracellular space, to reducing, within endo-lysosomal compartments) and pH (values from basic to highly acidic). For example, cathepsins are only activated into the acidic lysosomal lumen by PD cleavage (Pungercat et al., 2009), which also gives them the optimal pH for their peptidase activity (Jordans et al., 2009).

1.2.4 Peptidase inhibitors

Peptidase inhibitors are ubiquitous and present in multiple forms in numerous tissues of animals, plants and in microorganisms. They constitute up to 1% of the genes in metazoan. They appeared many

times during evolution and most of these events are linked to the eukarya (Rawlings, 2010; Rawlings et al., 2004). They accomplish a valuable function, the prevention of unwanted proteolysis.

Peptidase inhibitors of proteolytic enzymes are the largest and structurally most diverse group of enzyme inhibitors that appear naturally. A list of 99 inhibitor families is available at the MEROPS database with their inhibitor structures, modes of inhibition, kinetic and thermodynamic parameters. The nature of the enzyme–inhibitor complexes is very diversified, but there is a limited number of inhibition modes.

1.2.4.1 Classification of peptidase inhibitors

Peptidase inhibitors can be classified according to several criteria: by mechanism of inhibition, physiological role or evolutionary relationship (Otlewski et al., 2005; Rawlings et al., 2004; Turk, 2006).

1.2.4.1.1 By mechanism of inhibition

Regarding the mechanism of action, peptidase inhibitors must be firstly classified into reversible and irreversible inhibitors (Fig. 7).

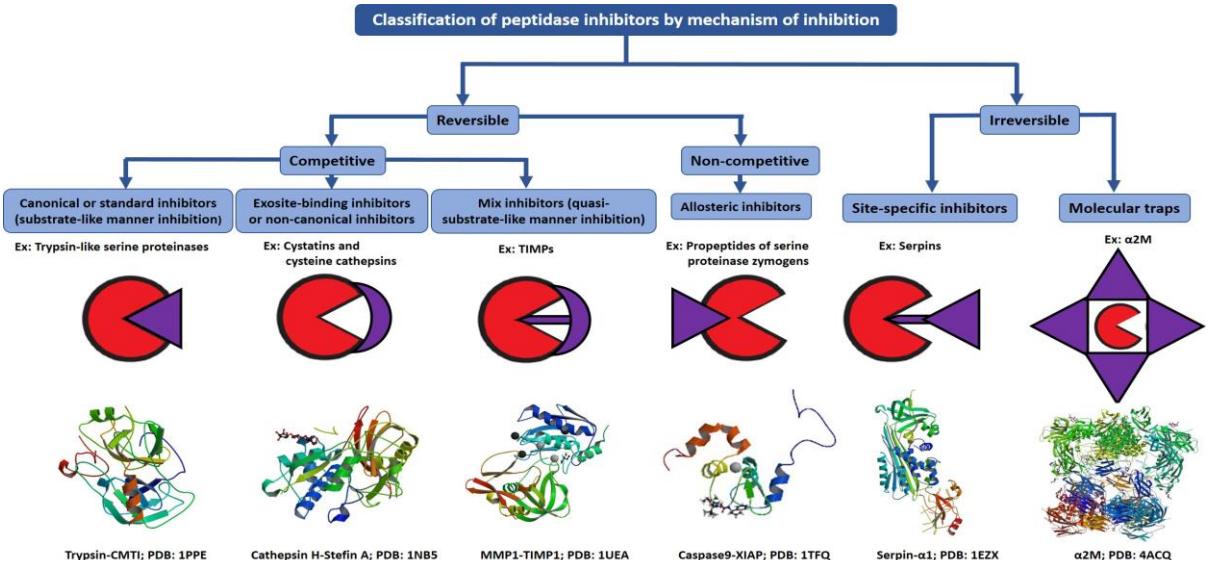


Figure 7. Classification of peptidase inhibitors attending to their mechanism of inhibition (Figure adapted from Garcia-Ferrer, 2015).

Reversible inhibitors

The vast majority of peptidase inhibitors are reversible and competitive. They establish tight non-covalent contacts with peptidase residues, that are part of the catalytic site or close enough to block

the access to the catalytic centre, this is the reason why they are called "competitive inhibitors". Reversible inhibitors bind to the protease to modulate its activity, but under some circumstances, dissociate (Bode and Huber, 1992). Up until now, three different mechanisms for reversible competitive protease inhibition have been described (Bode and Huber, 2000) and one for reversible non-competitive inhibition (Shiozaki et al., 2003):

- **Canonical or standard inhibitors with a substrate-like manner inhibition (lock-and-key inhibition mechanism):** The most studied is the standard inhibition mechanism of serine proteases (18 families that include Kazal, Kunitz, and Bowman-Birk families) (Qi et al., 2005; Ranasinghe and McManus, 2013), that it is also used to inhibit cysteine and metalloproteases (Otlewski et al., 2005). This mechanism of inhibition is based on the insertion of a reactive loop into the active site of the protease in an extended β-sheet conformation, which is complementary to the substrate specificity of the target protease. When this reactive loop is bound to the protease, the "scissile bond" is hydrolysed very slowly, the hydrolysed products are not released, and the amide bond can be re-ligated (Zakharova et al., 2009).

- **Exosite-binding inhibitors (also termed as non-canonical competitive inhibitors):** Most reversible inhibitors block access to the peptidase active site through their binding to subsites and catalytic residues adjacent to the catalytic centre, covering it partially (Turk, 2006). Exosite binding increases the surface area of the protein-protein interaction, leading to a greater affinity, and it can improve the specificity of the inhibitor. This inhibition mechanism was observed initially in cystatins and cysteine cathepsins (Stubbs et al., 1990). The cystatins, an inhibitor family of papain-like cysteine proteases, insert a wedge-like face (consisting of its N-terminal end and two hairpins) into the active site of the peptidase. The N-terminal residues bind in the S3-S1 pockets in a substrate-like manner, but the two hairpin loops bind to the prime-side of the active site. Thus, both the prime and non-prime sides of the active site are occupied, but no interactions are made with the catalytic machinery of the enzyme (Bode and Huber, 2000).

- **Inhibitors that combine the canonical and exosite binding mechanisms (or quasi-substrate-like manner):** A good example of this mechanism is the TIMPs (Gomis et al., 1997). They bind to their target enzymes (MMPs) in a two-step mechanism. The N-terminal end of TIMPs bind in the P1-P3' pockets of the protease, but also interfere with the catalytic Zn^{2+} ion and exclude a catalytic water molecule from the active site. A second loop of the TIMPs bind P3 and P2

pockets. Despite of the similarities with the mechanism of inhibition of the cystatins, it is different because the TIMPs interfere with the catalytic machinery of MMPs by chelating the catalytic Zn^{2+} (Brew et al., 2000).

- **Allosteric inhibitors:** Allosteric inhibition is a non-competitive modulation of the peptidase activity. This binding is at a different site from the protease active centre and induces a conformational change in the protease, which reduces the affinity of the active site for its substrate. E.g. X-linked inhibitor of apoptosis protein (XIAP) that inhibits caspase 9. XIAP binds selectively to the caspase-9 monomer, thereby preventing the formation of the catalytically active caspase-9 homodimer (Shiozaki et al., 2003).

Irreversible inhibitors

This type of inhibition leads to covalent modification of the target protease. These inhibitors act as a substrate, using the catalytic machinery of the peptidase to trap and inhibit it (Farady and Craik, 2010).

Irreversible or trap inhibitors are also called suicide substrates, which are usually large molecular-weight proteins. They are two families known with very different mechanisms of inhibition:

- **Site-specific inhibitors:** The best example of this group are the serine peptidase inhibitors called serpins. In humans, there have been identified at least 36 serpins, that bind and inhibit irreversibly primarily serine proteases (Gettins, 2002). Serpins have a large reactive centre loop (RCL) that is presented to their target protease for proteolytic processing. When the RCL is cleaved and its N-terminal half, still attached to the protease as an acyl-enzyme intermediate, very fast conformational change on the serpin, several orders of magnitude faster than the acyl-intermediate hydrolysis, is able to completely deform the catalytic triad of the protease, that becomes irreversibly inactive.

- **Molecular traps:** The universal inhibitor called α-2-macroglobulin (α2M) is the prototypic example (Laskowski and Kato, 1980; Travis and Salvesen, 1983). α2M and its relatives are responsible for clearing excess proteases from plasma, working as a “protease sponge”. When a protease cleaves one of these reactive loops (called the “bait region”) (Sottrup-Jensen, 1989), it triggers a conformational change, and the protease becomes covalently linked and trapped into the cage of the inhibitor. The proteases are still active and can hydrolyze small molecule

substrates that are already in the cage, but their action is limited. After binding, α2M and its cargo are quickly cleared from the blood stream and degraded in the lysosomes (Kolodziej et al., 2002).

1.2.4.1.2 By physiological role

Depending on temporal and spatial co-localization of the protease and its inhibitor, their relative concentrations and binding kinetics, protease inhibitors are classified in two major categories of physiological inhibitors (Turk et al., 2002):

Emergency type inhibitors

They are characterized by no co-localization with proteases, rapid binding and large excess concentration of inhibitor. Their major function is to block any protease activity in an inappropriate compartment. For example: cystatins (Turk et al., 2002).

Regulatory type inhibitors

They are characterized by often being co-localized with proteases and are responsible for the fine regulation of protease activity. Regulatory type inhibitors are subdivided into four sub-categories (Turk, 2006):

- **Threshold inhibitors:** Fast binders and at low concentration. Their main function is to neutralize accidental protease activation. For example: X-linked inhibitors of apoptosis (X-IAP), through the inhibition of caspases (Deveraux et al., 1997).
- **Buffer-type inhibitors:** Weak binders. Their function is to temporarily block proteases to prevent inappropriate activity. For example: Propeptides of cathepsins (Turk et al., 2000).
- **Delay-type inhibitors:** Slow binders and often irreversible. Their function is to enable protease activity for a limited amount of time. For example: Antithrombin (Olson et al., 2002).
- **Pro-inhibitors:** They are synthesized as inactive and require proteolytic processing to become active. For example: Invariant chain p41 fragment resulting from the proteolytic processing of major histocompatibility class II molecules (Bevec et al., 1996).

1.2.4.1.3 By evolutionary relationship

This is the most systematic manner to classify the peptidase inhibitors (Rawlings, 2010; Rawlings et al., 2004). The same approach is used for peptidases in the MEROPS database. Protease inhibitors are separated into units, families and clans.

According to the number of inhibitory units, they are classified into simple or compound inhibitors (multiple inhibitory units). Protein sequences longer than fourteen amino acids with one single inhibitory reactive site are identified as individual inhibitory units and given a unique identifier. Compound inhibitors, on the other hand, can be sub-classified into:

- Homotypic inhibitors: All inhibitory units are from the same family.
- Heterotypic inhibitors: Inhibitory families are from at least two different families.

Inhibitory units are then grouped by sequence similarity into families and structurally related families, into “clans” (Rawlings et al., 2018).

At the time of writing, more than 721 inhibitory units were assigned. From these identifiers, 670 corresponded to simple inhibitors and 51 for compound inhibitors.

Figure reproduced from “ksiopeaslight.com”.

1.3 The human subtilisin kexin-isozyme 1 (hSKI1)

1.3.1 Name, proteome and distribution

The human subtilisin kexin-isozyme 1 (SKI1), called membrane-bound transcription factor site-1 protease (MBTPS1) (EC:3.4.21.112), human site 1-protease (S1P) or endopeptidase S1P is a serine protease belonging to the subtilisin/kexin type of proprotein convertases (PC) (Seidah and Prat., 2012).

The human SKI1 gene is localized in the chromosome 16. It is more than 60 kbp long and contains 23 exons and 22 introns. Analysis of the promotor region revealed a highly G/C rich region containing a binding site for sterol regulatory element-binding protein 1 (SREBP-1) (Nakajima et al., 2000). Therefore, the expression of the SKI1 gene may be under the control of SREBP-1, which is a key regulator of the expression of genes essential for intracellular lipid metabolism. Moreover, knockout of SKI1 is as lethal for mice in the blastocyst stage as it is for zebrafish, revealing crucial functions (Mitchell et al., 2001; Schlombs et al., 2003).

SKI1 is widely expressed and particularly abundant in liver, anterior pituitary, thyroid and adrenal glands (Seidah et al., 1999). Regarding intracellular location, SKI1 localizes in ER, Golgi and other organelles, including lysosomal and endosomal compartments (Pullikotil et al., 2007).

1.3.2 Protein family and group databases

Serine proteases are characterized by the presence of an active serine, which is used for the hydrolysis of peptide bonds in substrates. They are one of the most diverse class of enzymes (1% of all proteins and 200 of these enzymes in humans) (Long and Cravatt, 2011). Serine proteases are divided into two major families: one related to the trypsin or chymotrypsin fold and another family that is closer to bacterial subtilisin, which are usually called subtilases (Siezen and Leunissen, 1997; Wright et al., 1969). MEROPS peptidase database classifies subtilases in a S8 family (Rawlings et al., 2010), which is in turn divided into two subfamilies: the prokaryotic subfamily S8A (including bacterial subtilisins) and the eukaryotic subfamily S8B (further split in the archetype yeast kexin and the mammalian proprotein convertases) (Fuller et al., 1989; Seidah, 2011).

SKI1 belongs to the mammalian proprotein convertase family (PC), which is constituted to date by nine mammalian secretory calcium-dependant serine proteases harbouring a distinctive Ser/His/Asp catalytic triad that mediates peptide bond scission (Seidah and Prat, 2002). The first seven members (proprotein convertase 1 (PC1), PC2, furin, PC4, PC5, paired basic amino acid cleaving enzyme 4 (PACE4) and PC7) cleave their substrates at single or paired basic residues. However, the other two members of the PC family (SKI1 and proprotein convertase subtilisin kexin 9 (PCSK9)) cleave at non-basic residues.

The PC family is involved in the activation and sometimes inactivation of multiple polypeptide hormones, growth factors, adhesion molecules and proteins from viral, parasitic and bacterial pathogens. As a consequence, they are also important for the progression of various diseases like pathogenic infections, inflammation and cancer (Artenstein and Opal, 2011).

1.3.3 Physiological role of SKI1

SKI1 catalyzes the first step in the proteolytic activation of the sterol regulatory element-binding proteins (SREBPs). Other known substrates are brain-derived neurotrophic factor (BDNF), N-acetylglucosamine-1-phosphate transferase subunits alpha and beta (GNPTAB), activating transcription factor 6 (ATF6) and the cyclic AMP-responsive element binding proteins (CREB3, CREB3L1, CREB3L3 and CREB4) (Ye et al., 2000; Seidah et al., 2007).

In contrast to basic PCs, SKI1 cleaves after hydrophobic or small residues with an Arg or Lys in the position P4 and a Val/Leu at P2 of the consensus motif: (R/K) -X- (V/L)-X ↓, where X is variable (da Palma et al., 2014; Pasquato et al., 2006). Known cleavage substrate motifs are: **Arg**-Ser-**Val**-Leu↓ (SERBP-2), **Arg**-His-**Leu**-Leu↓ (ATF6), **Arg**-Gly-**Leu**-Thr↓ (BDNF) and its own propeptide after **Arg**-Arg-**Leu**-Leu↓ (in bold positions P4 and P2 of the consensus motif)..

SKI1 has a high preference for cleavage at luminal sites of membrane bound transcription factors. This cleavage is usually followed by a second cleavage via site 2 protease (S2P), which is a metalloproteinase in the medial-Golgi. This releases a cytosolic fragment that is translocated into the nucleus, activating the transcription and production of proteins that regulate multiple processes in the cell, such as, cellular ER/Golgi stress response (regulated by ATF6 and CREBs) (Seidah and Prat, 2012) and cholesterol and fatty acid biosynthesis (regulated by SREBPs) (Sakai et al., 1998). Other funtions of SKI1 are the regulation of the targeting of proteins from Golgi to lysosomes (via maturation of the GNPTAB, which is crucial for mannose-6-phosphate labeling of lysosomal proteins required for correct trafficking) (Marschner et al., 2011), bone mireralization (regulated by SREBP and CREB/ATF family transcription factors) (Gorski at al., 2010) and finally, neurance guidance and survival (via cleavage of repulsive guidance molecule A) (Tassew et al., 2012).

SKI1 is also involved in viral infections, mainly human Arenavirus (Lassa virus, lymphocytic choriomeningitis virus, Machupo virus, Guaranito virus), Bunyavirus, dengue virus, hepatitis C virus, human immunodeficiency virus and rotavirus (Lenz et al., 2001; Vincent et al., 2003). Arenavirus cause several hundred thousand infections each year in Africa and South America and the only available treatment is the ribavirin (antiviral drug to treat the hepatitis C), that is exclusively effective in early

disease and with multiple side effects. SKI1 is crucial for producing viral infection, via processing of the viral envelope glycoprotein precursor (GPC) and overstimulating of the host lipid metabolism to assembly new viral capsids (Beyer et al., 2003; Kunz et al., 2003; Rojek et al., 2008). Thus inhibitors of SKI1 are seen as promissing drugs for early and late stages of viral infection, like for example, secreted Spn4A.RRLL which forms a covalent complex with mature SKI1 in Golgi and the small-molecule inhibitor PF-429242 (Blanchet et al., 2015; Olmstead et al., 2012; Rojek et al., 2010) (Fig. 8).

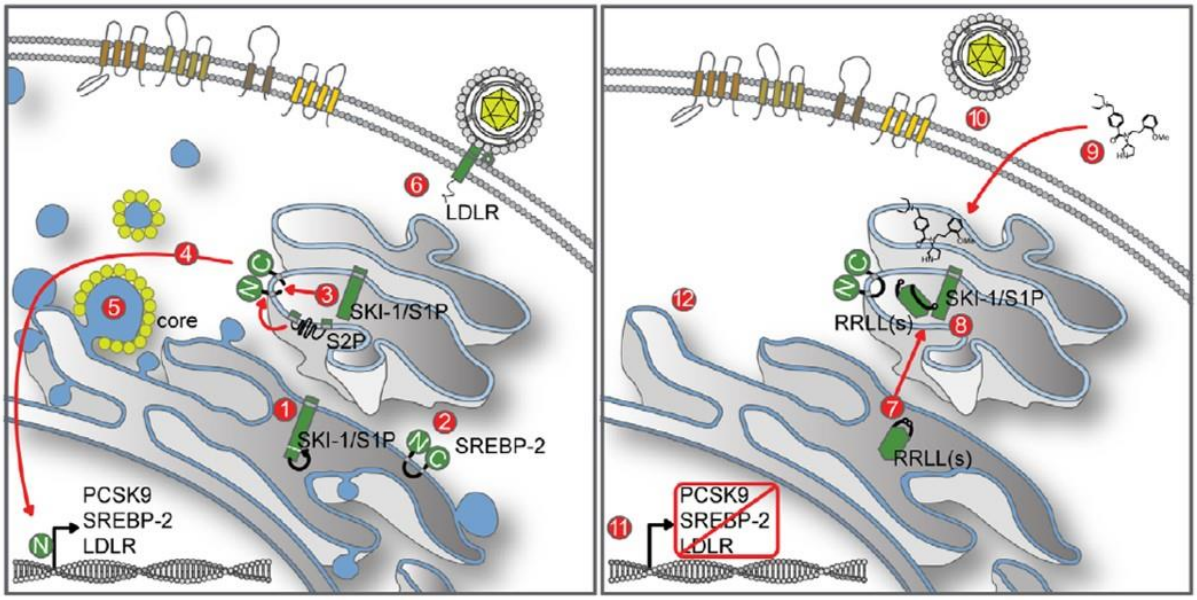


Figure 8. Mechanisms of SKI1 inhibition for hepatitis C virus (HCV) infection treatment (Figure reproduced from Olmstead et al., 2012). **(1)** Inactive SKI1 zymogen is biosynthesized in the RE and traffics to the Golgi apparatus where it finishes its autocatalytic maturation (cleavage of the prodomain) and becomes fully active. **(2**) During HCV infection, the SREBP pathway is activated for the overstimulation of host lipid metabolism (HCV requires host lipid droplets (LD) for assembly of nascent viral particles). **(3)** To activate genes involved in lipid biosynthesis, the N-terminal domain of SREBP-2 must be released through sequential cleavages of, first SKI1 and then, S2P. **(4)** The released N-terminal domain of SREBP-2 is translocated to the nucleus to activate lipid metabolism. **(5)** Activation of the lipid synthesis increases LD formation where the HCV core protein localizes, and it orchestrates HCV assembly and secretion. **(6)** Biosynthesis of low-density lipoprotein receptor (LDLR), which is proposed to be a receptor for HCV entry, is also activated by SREBP signalling. **(7) (8)** Spn4A.RRLL is a secreted engineered serpin to inhibit irreversibly SKI1 in Golgi apparatus, preventing the cleavage of SREBP-2. **(9)** PF-429242 can be added extracellularly and travels to Golgi where it blocks SKI1 endoproteolytic activity. **(10)** SKI1 inhibition also blocks the expression of LDLR, reducing HCV entry. **(11)** Expression of SREBP-regulated genes are also blocked. **(12)** Downstream lipid synthesis is stopped, reducing intracellular cholesterol-ester and triglyceride abundance, resulting in a decrease of LD abundance, which impedes assembly and secretion of infectious HCV particles.

1.3.4 Zymogen activation of SKI1

SKI1 is initially synthesized as an inactive precursor of 1052 amino acids, comprising a signal peptide (aa 1-17), followed by a prodomain (aa 18-186), a luminal domain (aa 187-998) (which includes a serine protease domain between Asp[218]-Ser[414] residues and a subtilase domain between Pro[190]-Tyr[472]), a transmembrane domain (aa 999-1021) and finally, a basic cytosolic tail (aa 1022-1052) (Fig. 9A)

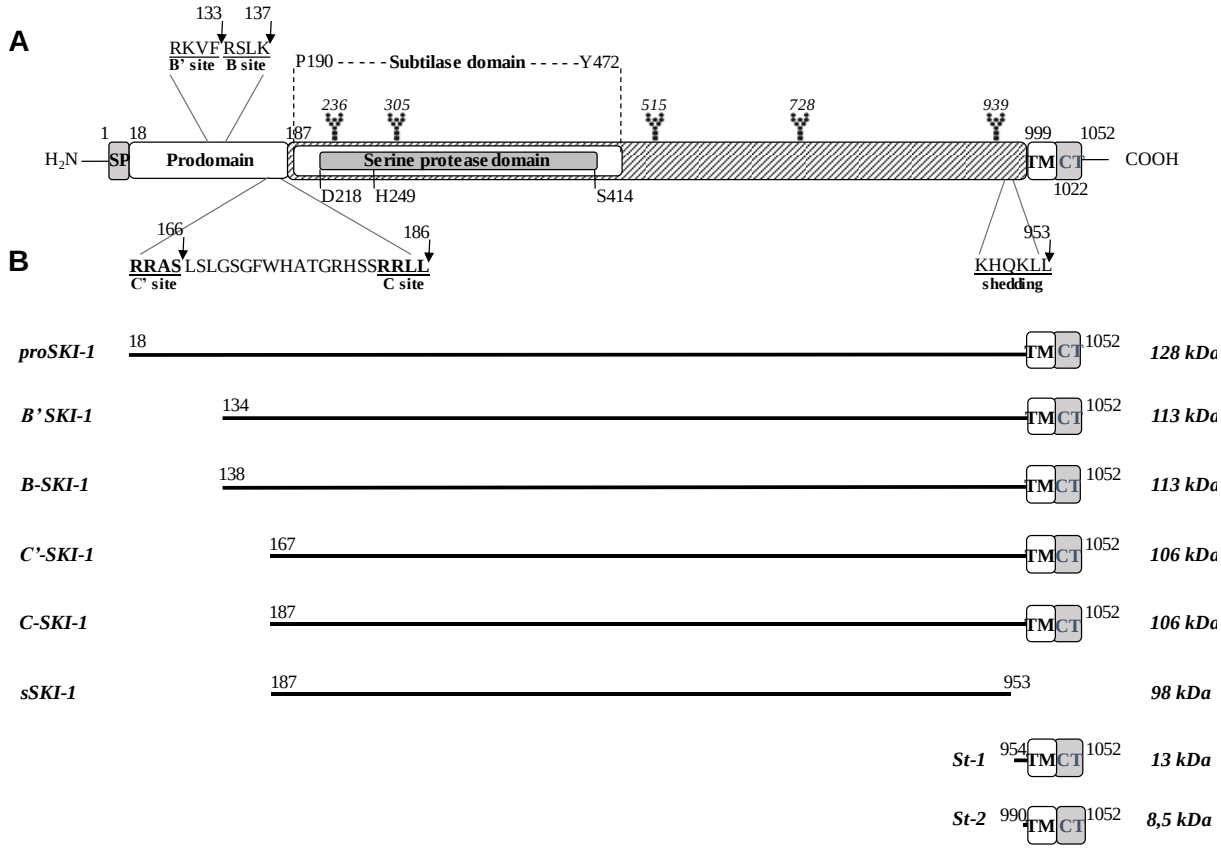


Figure 9. Schematic representation of hSKI1 protein and the processed forms of SKI1 after the multistep maturation process. The residue numbers correspond to UniProt (UP), entry Q14703. **(A)** Functional domains of zymogen SKI1 are the signal peptide (SP), prodomain (PD), transmembrane (TM) and cytosolic tail (CT). N-linked glycosylation sites are highlighted with a sugar chain. Cleavages and shedding motifs are shown in magnification and underlined. The exact cleavage is also pinpointed with a black arrow. **(B)** Scheme depicting the basic domain structure of the different processed SKI1 forms and stubs that result after the B'/B and C'/C cleavages and the subsequent shedding. Amino acid length and molecular weight are given for each processed form of SKI1.

Normal zymogen activation of PC require excision of their prodomain for activity. However, SKI1 has an unusual mechanism for zymogen activation, that includes a multistep process with different subcellular locations and several protein intermediates (da Palma et al., 2014; da Palma et al., 2016). Upon translocation of the SKI1 zymogen (aa 18-1052) into the ER, SKI1 undergoes autocatalytic maturation by sequential cleavages in *cis* of the N-terminal prodomain at sites B'/B (**R**K**V**F^{133}↓ **R**S**L**K^{137}↓). Thereupon, a secondary cleavage in *trans* of the prodomain at sites C' (**RRA**S^{166} ↓) (da Palma et al., 2014) and C (**R**R**L**L^{186}↓) occur in the *cis*- and *medial*- Golgi (in bold positions P4 and P2 of the consensus motif: (R/K) -X- (V/L)-X↓, where X is variable). The resulting product of all these cleavages is the fully mature enzyme or membrane-bound SKI1 (aa 187-1052) (Elagoz et al., 2001; Touré et al., 2000), in addition to all the intermediates with prodomain fragments of different lengths (aa 134- 1052, aa 138-1052 and aa 167-1052). Nevertheless, in contrast to other zymogen PCs, all incompletely matured intermediates of SKI1 retaining fragments of the prodomain have showed catalytic activity toward cellular and viral substrates in distinct subcellular compartments (da Palma et al., 2014).

Membrane-bound SKI1 is predominantly found in Golgi and partially in endosomal compartments. A part of this enzyme undergoes autocatalytic shedding (KHQKLL^{953}↓) close to the transmembrane domain, leaving a membrane-associated stub (St-1: aa 954- 1052) behind (Elagoz et al., 2001). The resulting protein is the secreted full-active SKI1 (aa 187-953), which can sort to endosomes, lysosomes and to the extracellular surface. There is a second stub (St-2: aa 990-1052) resulted by an alternative translation of the primary transcript, which starts at Met^{990}. This mechanism has been seen in viruses to express variant proteins, but it is very unusual for secretory proteins in mammals (Pullikotil et al., 2007) (Fig.9B).

As a consequence of this unusual mechanism of maturation, SKI1 can act in at least three different compartments of the secretory pathway, depending on the substrate (processing of the Lassa virus GPC occurs in the ER/*cis*-Golgi, whereas SREBP-2 and GNPTAB activations occur in the median or late Golgi, respectively) (Burri et al., 2012). It has also been published by da Palma (2014) that SKI1 forms are in association with prodomain fragments of different lengths. Mature SKI1 is a heterogenous collection of mature and immature proteins with prodomain pieces rather than a monomer of defined length and composition. Moreover, they proposed a model for the autocatalytic activation of SKI1 with cleavages in *cis* (for B'/B) and in *trans* (for C'/C) (Fig. 10).

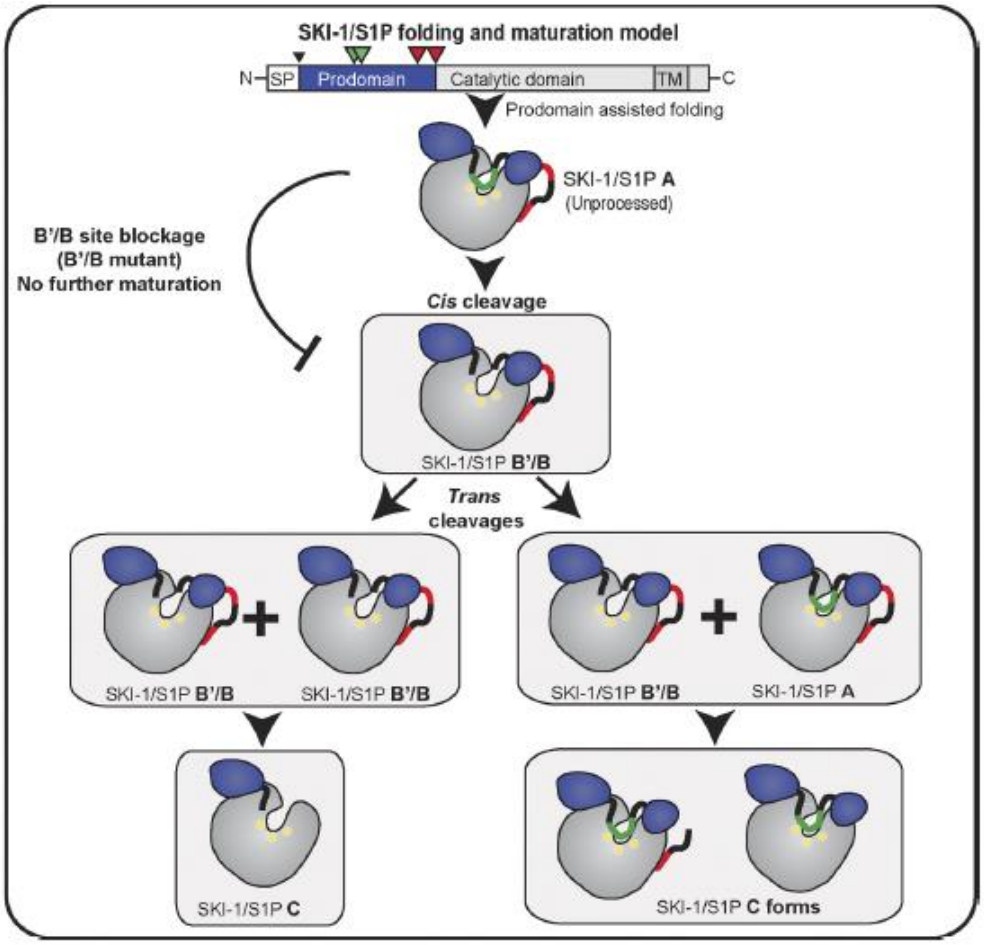


Figure 10. Model for SKI1 zymogen activation (Figure reproduced from da Palma et al., 2014). Schematic representation of SKI1 forms (prodomain B'/B and C'/C site cleavages in green and red, respectively). SKI1 suffers a multistep maturation process. According to this model, when the *cis* cleavages at B'/B take place, the SKI1 B forms are generated. The B forms can process other SKI1 forms at the C'/C site in *trans,* generating processed and unprocessed SKI1 forms that might coexist both spatially and temporally.

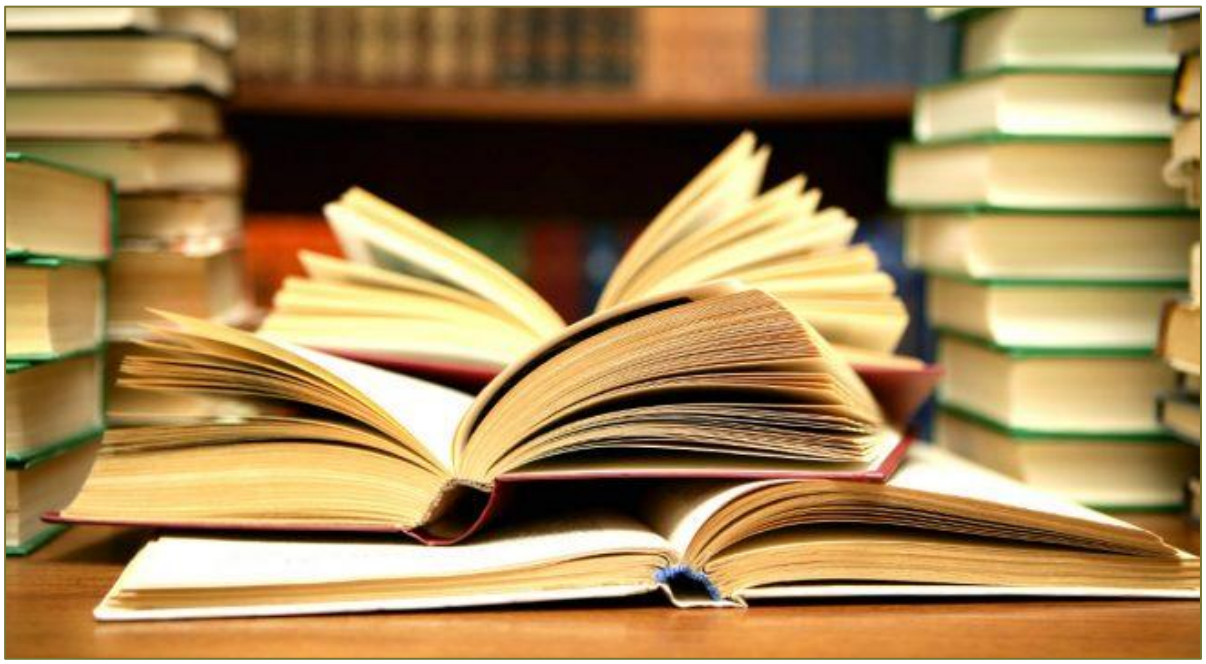


Figure reproduced from “pxhere.com”.

INTRODUCTION

1.4 Matrix metalloproteinases

Marino-Puertas L., Goulas T and Gomis-Rüth FX. (2017). Matrix metalloproteinases outside vertebrates. *Biochimica Biophysica Acta Molecular Cell Research, 1864*(11 Pt A):2026-2035. doi: 10.1016/j.bbamcr.2017.04.003. Epub 2017 Apr 7. Review.

1.4.1 Molecular characteristics of matrix metalloproteinases

The matrix metalloproteinases (MMPs) are a widespread family of zinc-dependent metallopeptidases (MPs), which either broadly degrade extracellular matrix components or selectively activate or inactivate other proteins through limited proteolysis (Morrison et al., 2009; Nagase et al., 2006; Tallant et al., 2010b). MMPs were discovered in 1962 as an active principle in frog metamorphosis (Gross and Lapière, 1962), and they contain a central zinc-dependent catalytic domain (CD) of ~165 residues, which is mostly furnished at its N-terminus with a ~20-residue signal peptide for secretion and an ~80-residue zymogenic prodomain (PD). Some MMPs possess a "furin recognition motif" (R-X-R/KR; (Krysan et al., 1999) for proteolytic activation between the PD and the CD. Into this minimal configuration, distinct MMPs have inserted extra segments and domains, such as fibronectin-type-II inserts within the CD and C-terminal unstructured linker regions, ~200-residue hemopexin domains (HDs) and other domains, as well as glycosyl phosphatidylinositol (GPI) anchors and transmembrane segments (Fanjul-Fernández et al., 2010; Nagase et al., 2009; Tallant et al., 2010b; 6). Overall, this variability divides human MMPs into subfamilies: archetypal MMPs, gelatinases, matrilysins and furin-activatable MMPs, which include membrane-type MMPs (Fanjul-Fernández et al., 2010; Nagase and Woessner, 1919).

MMPs belong to the metzincin clan of MPs, which currently comprises twelve structurally characterized families with common features (Bode et al., 1993; Cerdà-Costa and Gomis-Rüth, 2014; Gomis-Rüth, 2003; Gomis-Rüth, 2009; Stöcker et al., 1995). However, this structural similarity is not reflected at sequence level, because pairwise identities are typically< 20%, which is below the twilight zone of relevant sequence similarity (25–35%; (Rost, 1999)). Overall, metzincins share a globular CD of ~130–270 residues, which is split into a structurally conserved N-terminal sub-domain (NTS) and a diverging C-terminal sub-domain (CTS) by a horizontal active-site cleft. This cleft contains the catalytic zinc ion and accommodates protein or peptide substrates for catalysis (Fig. 11A). NTSs contain a mostly five-stranded β-sheet, whose strands (βI-βV) parallel the active-site cleft. The lowermost strand βIV shapes the upper rim of the cleft and runs antiparallel to the other strands and to a bound substrate, thus establishing inter-main-chain interactions to fix it. NTSs also possess a "backing helix" (αA) with structural functions and an "active-site helix" (αB) with an extended "zinc-binding motif", H-E-X-X H-X-X-G/N-X-X-H/D-Φ. This motif includes the general base/acid glutamate required for catalysis and three metal-binding protein residues (in bold), as well as a "family-specific residue" (Φ; (Bode et al., 1993; Cerdà-Costa and Gomis-Rüth, 2014; Stöcker et al., 1995). Some families further comprise an additional "adamalysin helix", which was first identified in the adamalysin/ADAM family (Cerdà-Costa and Gomis-Rüth, 2014; 14]. In contrast to this structural conservation, CTSs largely deviate in length and structure but share the "Met-turn" (Tallant et al., 2010a), a loop with a methionine placed just below the

catalytic zinc, and a "C-terminal helix" (αC) that terminates the CD. Into this common minimal scaffold, specific molecular elements decorate each metzincin family in the form of loops, ion-binding sites, additional regular secondary structure elements, extra domains, etc. (Bode et al., 1993; Cerdà-Costa and Gomis-Rüth, 2014; Gomis-Rüth, 2003; Gomis-Rüth, 2009; Stöcker et al., 1995).

To date, structural studies on MMP CDs are restricted to mammals (human, mouse, rat and pig), with the notable exception of bacterial karilysin (Cerdà-Costa and Gomis-Rüth, 2014; see also Section 3.3.2.5 in Results). These structures revealed that MMPs are very similar (Fig. 11B), have much longer NTSs (~127 residues) than CTSs (~37 residues), and include an "S-shaped double loop", which connects the third and fourth strands of the NTS β-sheet (Fig. 11A). This loop binds essential structural zinc and calcium cations, after which the polypeptide forms a prominent bulge ("bulge-edge segment") that protrudes into the active-site groove and assists in substrate specificity. A further hallmark of MMPs is a second calcium site within the NTS and that the family-specific residue is a serine or threonine (Tallant et al., 2010b; Gomis-Rüth, 2003). This residue makes a strong hydrogen bond between side chains with the first of two consecutive aspartates at the beginning of helix αC (Fig. 11C). This aspartate is also engaged in fixing the N-terminus of the mature CD upon proteolytic activation. In turn, the second aspartate participates in a double electrostatic interaction with main-chain atoms of the Met-turn (Fig. 11A, C). This intricate electrostatic network is structured around the "connecting segment", which links the third zinc-binding histidine with the Met-turn methionine and spans only seven residues; thus, it is the shortest of the metzincins (Gomis-Rüth, 2003). Further typical of MMPs are two tyrosines, respectively four positions downstream of the Met-turn methionine and featuring the penultimate residue of the CD. Finally, the Met-turn and helix αC are connected by a loop subdivided into the "S1'-wall-forming segment" and the downstream "specificity loop" (Fig. 11A). Taken together, all these characteristic structural features yield an extended sequence pattern, H-E-2X-H-2X-G-2X-H-S/T-6X-M-3X-Y-9X-D-D-7X-Y-X (Fig. 11D), which distinguishes the C-terminal segment of MMP CDs from other metzincin families and MPs in general (Cerdà-Costa and Gomis-Rüth, 2014).

In addition to the CD, a further hallmark of MMPs is that latency is maintained through the PD (Tallant et al., 2010b), which together with transcriptional regulation and dedicated protein inhibitors regulates physiological MMP activity (Mittal et al., 2016). In most MMPs, zymogenicity is exerted by a "cysteine-switch" mechanism centered on a cysteine within a consensus motif at the end of the PD, P-R-C-G-X-P-D (Van Wart and Birkedal-Hansen, 1990), which is found in all human MMPs except MMP-23B. Structurally, the polypeptide chain spanning these residues shields the active-site cleft by adopting a U-shaped conformation mediated by a double salt bridge between the arginine-aspartate pair and by the glycine, whose missing side chain prevents collision with upper-rim strand βIV (Tallant

et al., 2010b),. In addition, the cysteine Sγ atom coordinates and thus blocks the catalytic zinc (see Fig. 3B in (Tallant et al., 2010b). MMP activation entails PD removal to free access to the cleft.

Figure 11. Matrix metalloproteinases (MMP) catalytic domain structure. (A) Cross-eye stereographic Richardson-plot of the catalytic domain of T. forsythia karilysin (Y35-P200; PDB 2XS3; (Cerdà-Costa et al., 2011) in standard orientation (Gomis-Rüth et al., 2012) as a representative of MMP catalytic domains (see Tallant et al., 2010b). Green arrows represent β-strands (labeled βI-βV), brown ribbons stand for α-helices (αA-αC). The two zinc cations found across MMPs are shown as magenta spheres, the calcium cations found in most MMPs but missing in karilysin are depicted in red. Relevant chain segments are shown in distinct colors and labeled (Met-turn in blue, specificity loop in red, S1'-wall-forming segment in yellow, S-loop in white, and bulge-edge segment in cyan). The side chains of the zinc-binding histidines, the general base/acid glutamate, the Met-turn methionine, the family-specific serine, and the two tyrosines and aspartates further included in the extended signature characteristic of MMPs are shown as stick models. **(B)** Superposition of the Cα-traces of the catalytic domains of bacterial karilysin (in yellow) and human MMP-1 (PDB 966C; (Lovejoy et al., 1999), MMP-3 (PDB 1CIZ; (Pavlovsky et al., 1999)), MMP-7 (PDB 1MMQ; (Browner et al., 1995)), MMP-8 (PDB 1JAN; (Reinemer et al., 1994)), MMP-9 (4H3X; (Antoni et al., 2013)), MMP-10 (PDB 1Q3A; (Bertini et al., 2004)), MMP-11 (PDB 1HV5; (Gall et al., 2001)), MMP-12 (PDB 1Y93; (Bertini et al., 2005)), MMP-13 (PDB 2D1N; (Kohno, et al., 2006)) and MMP-16 (PDB 1RM8; (Lang et al., 2004)), all in cyan. The two consensus zinc and calcium cations are shown as magenta and red spheres, respectively; the consensus N- and C-termini are indicated by red arrows. **(C)** Close-up view of (A) highlighting the structural features of the C-terminal sub-domain characteristic for MMPs. The hallmark electrostatic interactions within the CTS are shown as cyan lines, residues included in the extended sequence pattern are depicted for their side chains and labeled. **(D**) Color-coded mapping of the MMP extended sequence pattern (H-E-2X-H-2X-G-2X-H-S/T-6X-M-3X-Y-9X-D-D-7X-Y-X) onto the karilysin polypeptide chain. Each residue type is shown in one color and labeled, random residues (X) are alternatively in light green and yellow.

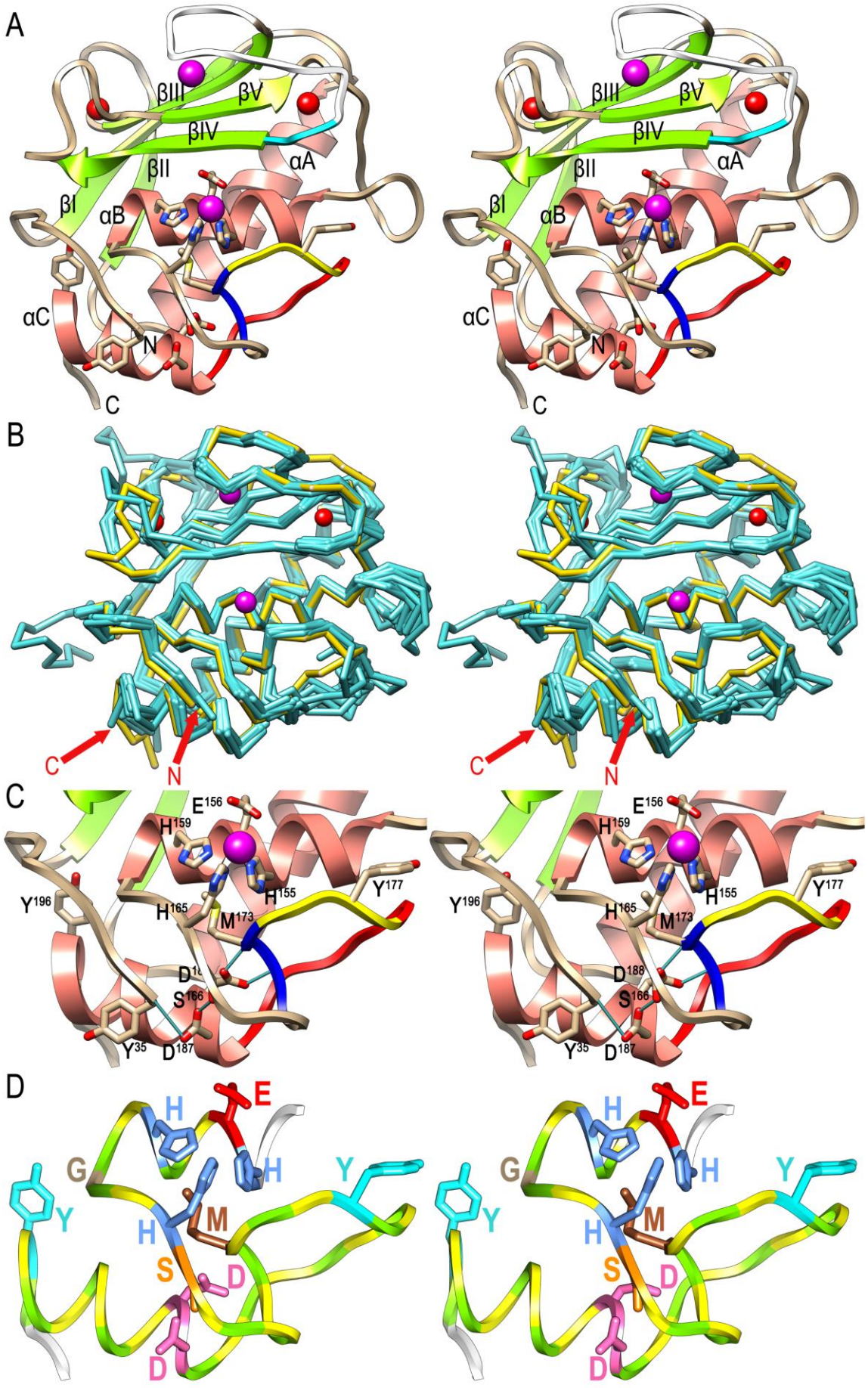

A
βIII
βV
βIV
αA
βII
βI
αB
αC
N
C
B
C
N
C
E156
H159
H155
Y177
Y196
H165
M173
D188
S166
Y35
D187
D
H
E
G
H
Y
Y
H
M
S
D
D

Figure reproduced from “lavanguardia.com”.

1.5 The human α2-macroglobulin (hα2M)

Goulas T, Garcia-Ferrer I, Marrero A, Marino-Puertas L, Duquerroy S, Gomis-Rüth FX. (2017). Structural and functional insight into pan-endopeptidase inhibition by α2-macroglobulins. *Biological Chemistry, 398*(9):975-994.

1.5.1 Name and proteome

The α2-macroglobulin protein (α2M) is alternatively called C3 and PZP-like alpha-2-macroglobulin domain-containing protein 5.

The human α2M gene is localized in the chromosome 12. Analysis of the sequence revealed regulatory sequences homologous to the IL-6 response element as well as to the HP-1 element, a metal responsive element, and a potential Sp1 binding site in the 5' flanking region of exon 1 (Borth W, 1992).

1.5.2 Protein family, origin and distribution

α2M protein belongs to the "α2-macroglobulin family" (MEROPS family I39), which is a large (>180 kDa) multi-domain protein family which act as broad-spectrum endopeptidase inhibitors (Barrett and Starkey, 1973; Rehman et al., 2013). α2Ms also belong to the thioester-containing proteins (TEPs) whose members share a common evolutionary origin and with conserved structural and functional features (they are characterised by the presence of an intrachain β-cysteinyl γ-glutamyl thioester bond in a conserved motif with sequence CXEQ (Fig. 12) (Sekiguchi et al., 2012)) The α-macroglobulin members of this family in humans are the α1-macroglobulin (α1M), α2-macroglobulin (α2M), complement components (C3, C4 and C5), pregnancy zone protein (PZP), CPAMD8 and CD109 (Neves et al., 2012).

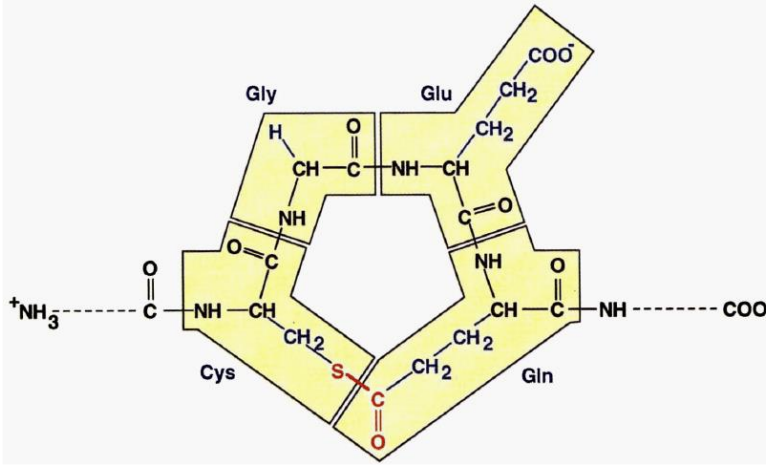


Figure 12. Representation of the thioester ring of the C3 and C4 complement proteins (Figure reproduced from Law and Dodds, 1997). Thioester ring is composed of the residues: Cys-Gly-Glu-Gln (Gly could be replaced by Ala, as in the case of mouse C4). In black, the polypeptide backbone; in blue, the side chains; and in red, the thioester bond.

α2Ms are ancient diverging molecules. To date, they have been characterized from metazoans (mammals, reptiles, amphibians, fishes and invertebrates) and in colonizing Gram-negative bacteria, but not in free-living microbes (Barret and Starkey, 1973; Garcia-Ferrer et al., 2017; Goulas et al., 2017;

Sottrup-Jensen et al., 1989). It is worth to note that phylogenetic studies revealed that the gene distribution in bacteria violates the vertical descent model, which suggests that bacterial α2Ms (bα2Ms) might have been acquired from metazoans by horizontal gene transfer through long contacts between colonizing bacteria and their-blooded animal hosts (Budd et al., 2004; Doan and Gettins, 2008).

Regarding the oligomerization state of α2Ms, some inhibitors have been reported to be tetrameric (like the human blood α2M), dimeric (such as human PZP) (Sand et al., 1985) and monomeric (as the *Escherichia coli* α2M, also termed as ECAM) (Garcia-Ferrer et al., 2015).

The best and first studied member of the α2Ms family is the human α2M (hα2M) (Barrett, 1981; Barrett and Starkey, 1973; Starkey and Barrett, 1973). hα2M was first isolated in 1946 by Cohn and colleagues (Cohn et al., 1946) and first named in 1955 by Schultze and co-workers (Schultze et al., 1955).

hα2M is an abundant 720 kDa homotetrameric secreted protein that is found at ~3mg/mL and 1.0–3.6 μg/mL in human blood plasma and cerebral spinal fluid, respectively (Ganrot and Schersten, 1967; Garton et al., 1991; Sottrup-Jensen, 1989). It is mainly synthesized in the liver (Barret and Starkey, 1973; Sottrup-Jensen, 1989; Travis and Salvesen, 1983) by hepatocytes but other cell lineages, such as lung fibroblasts, monocytes-macrophages and astrocytes are also able to produce significant amounts of α2M (Matthijs et al., 1992; Sottrup-Jensen, 1989; Sottrup-Jensen, 1987).

1.5.3 Biological functions of hα2M

hα2M is the main proteinase inhibitor of the blood plasma, but other hα2M can be found in epidermis and pregnancy serum (Barrett et al., 1979; Cray et al., 2009; Galliano et al., 2006; Sand et al., 1985; Tunstall et al., 1975). hα2M targets are proteinases of the four major classes (serine, cysteine, metallo and aspartic), which form 1:1 or 2:1 complexes with tetrameric hα2M (Sottrup-Jensen, 1989; Travis and Salvesen, 1983). Here, the specificity depends on the access to the protease to the bait region and cleavage, initiating the process of entrapment. Hence, α2Ms regulate proteolysis in complex biological processes such as digestion, blood homeostasis, signalling, tissue remodelling and defence against toxins and other virulence factors during infection and envenomation (Garcia-Ferrer et al., 2017). They are distantly related to the innate-immunity proteins of host defence, in particular the complement system, despite of not being peptidase inhibitors (Armstrong and Quigley, 1999; Sottrup-Jensen et al., 1985; Vogel et al., 2014).

In addition to peptidase binding and inhibition, hα2M modify and modulate the activity of cytokines (Gonias et al., 2000; LaMarre et al., 1991b), growth factors (Barcelona and Saragovi, 2015; Skornicka et al., 2002; Westwood et al., 2001; Wolfand and Gonias, 1994), lipid factors (Krimbou et al., 1998) and misfolded proteins (French et al., 2008; Whiten et al., 2018; Wyatt et al., 2013a; Wyatt et al., 2013b; Yerbury et al., 2009), and thus have a great impact on human physiology.

hα2M is one of the few secreted proteins that are known to possess holdase-type chaperone activity, which is the ability to stabilise misfolded proteins and prevent their aberrant aggregation (Wyatt et al., 2014). Playing an important role in human diseases like Alzheimer, Parkinson and motor neuron diseases where the pathology is caused by the accumulation of misfolded proteins.

Besides its physiological role in the human body, hα2M can be used by bacteria for their own protection by peptidases against host immune response. This is the case of the G-related α2M-binding protein (GRAB), which facilitates the *Streptomyces pyogenes* infection with this mechanism (Müller and Rantamäki, 1995; Rasmussen et al., 1999).

1.5.4 Structural aspects of hα2M

Primary structure

hα2M is a 720 kDa homotetramer protein, where each subunit (around 180 kDa) is a multidomain glycosylated molecule of 1474 residues. There are some important motifs in its sequence for protein functionality. One of the most relevant motifs is the one which encodes the hyperactive thioester bond (C-X-E-Q: where X is for G or L) (Sottrup-Jensen et al., 1985), which is present in almost all the TEPs, as we previously described.

Each 180 kDa monomer starts with a 23 residues-signal peptide for secretion, either to the blood serum and hemolymph in vertebrates and invertebrates, respectively. It is also stabilized by thirteen disulphide bonds per monomer (Doan and Gettins 2008), where eleven are intra-chain disulphide bonds and two are inter-chain bonds between adjacent monomers (Cys278 and Cys431) (Marrero et al. 2012). In addition to this, hα2M has eight glycosylations per monomer, that contribute to protein solubility and stability (Goulas et al. 2014; Marrero et al. 2012).

Secondary and tertiary structures

hα2M is formed by eleven domains (Fig. 13 and 14A, B, C) (Marrero et al. 2012):

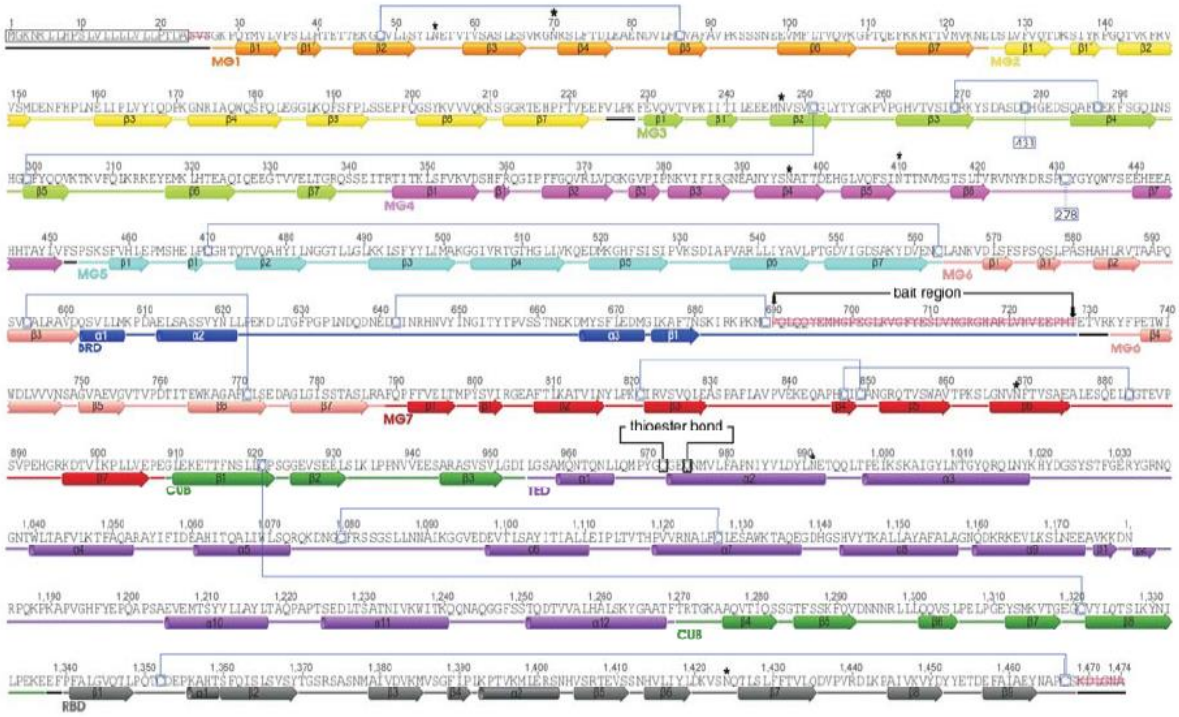


Figure 13. Secondary structure elements of hα2M (Figure reproduced from Marrero et al. 2012). Protein domains are represented in different colors. α-helix and β-strand elements are represented as barrels and arrows, respectively. Linker regions are shown as black lines. The bait region and the thioester bond are tagged with arrows and boxes in black, respectively. Intra- and inter-molecular disulphide bonds are highlighted with square boxes linked by a solid or dotted line in blue, respectively. N-glycosylation sites are marked by asterisks. Residues not defined by electron density in the crystallized protein (PDB 4ACQ) are shown with a double strikethrough in red

- The first seven domains are termed macroglobulin-like (MG) domains (hence MG1-MG7 domains). They are constituted by seven-stranded antiparallel β-sandwiches, containing a three- and four-stranded sheet, and they are arranged like a central scaffold, into which other elements are inserted (Fig. 14D).
- There is an insertion from Q^{602} to T^{728} residues in the domain MG6, which is an extended flexible domain called the bait-region domain (BRD). A very disordered segment is found at the end of this structure, which is the 39-residue bait-region (P^{690}-T^{728}) (Marrero et al., 2012). This domain is recognized by endopeptidases of the four classes and one of the responsibles for the conformational change of α2M when it is cleaved and the exposition of the buried thioester (further explained). It is formed by three helices and it is stabilized through interactions with adjacent MG domains and the C-terminal receptor-binding domain (RBD).
- Following the MG7 domain, the β CUB domain (C1r/C1s, Uegf and Bmp1 found domain) is placed. This domain is formed by two four-stranded antiparallel β-sheets (Fig. 14E).

- The thioester domain (TED) is found inserted in the CUB domain and bellow. It owns a 315-residue helical domain with a α/α-toroid topology (Fig. 14F). This domain consists of six concentric α-hairpins, arranged as a six-fold α-propeller around a central axis. This is the part for the protein that is responsible for the covalent binding to the protease to the inhibitor through the thioester bond (**C^{972}-G^{973}**-E^{974}-Q^{975} residues). The thioester bond is buried in the native state of hα2M and after its cleavage, the resulting free residues, C^{972} and modified Q^{975}, protrude out from the surface of TED.
- After TED, the polypeptide chain re-joins CUB, which ends in the C-terminal domain, the RBD. This domain is also known as MG8 domain because its β-sandwich architecture is very similar to the MG domains, but with a β-α-β insertion, resulting in a four-stranded and five-stranded twisted sheet (Fig. 14G). When α2M is induced, this domain is exhibited and accessible to be bound to the low-density lipoprotein receptor-related protein (LRP).

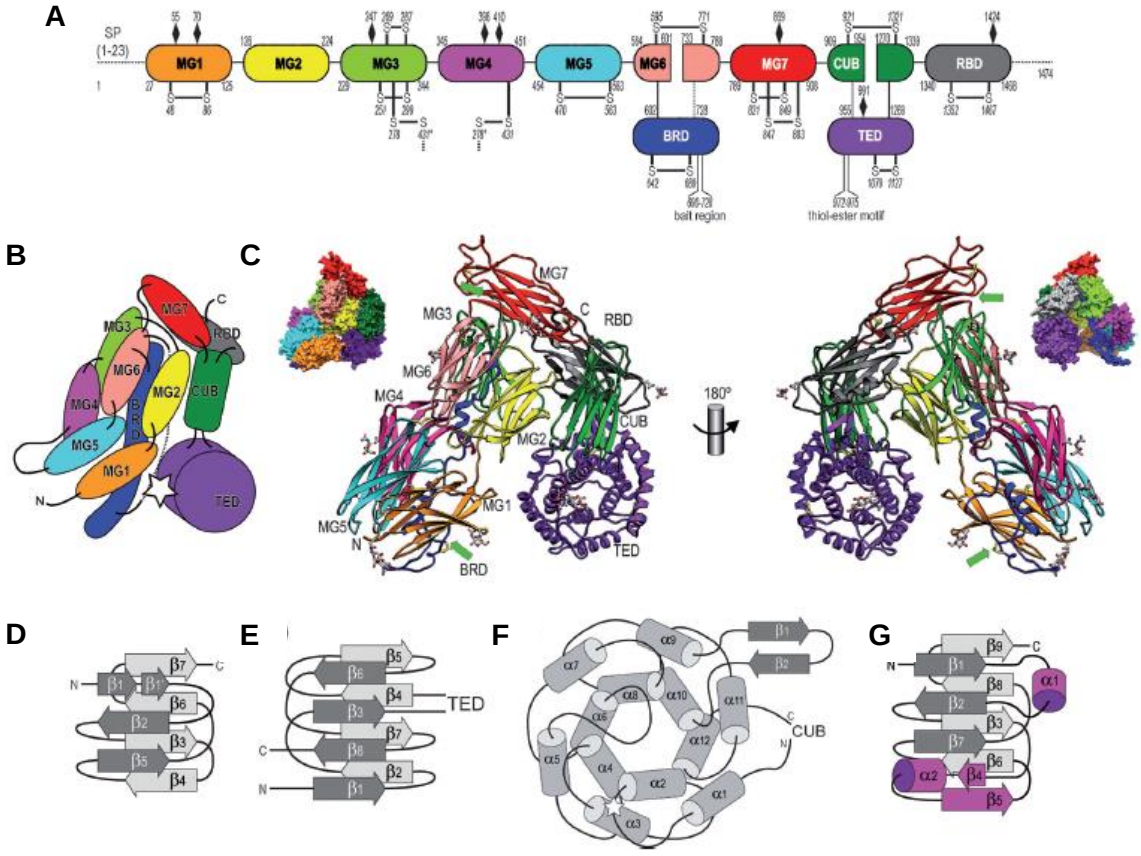


Figure 14. Domain organization of hα2M monomer, arrangement of its domains and its secondary-structure elements (Figure reproduced from Marrero et al. 2012). **(A)** Domain organization of human a2M protein, representing the residues of each domain, the disulphide bonds, the N-linked glycosylation sites (black rhombuses) and the insertion sites of the bait region and the thioester motif. **(B)** Arrangement of the domains forming the a2M-MA (methylamine) monomer in front view. The thioester site is shown as a white star and the flexible bait region, as a dashed line. **(C)** Molecular surface and Richardson-type plot of an a2M-MA monomer in front view (convex face; left) and in back view (concave face; right), where green arrows tag the anchor points of the flexible bait region. **(D-G)** Secondary-structure elements representation of the MG domains **(D),** the CUB domain **(E),** the TED domain **(F),** and the RBD domain **(G).** The differential elements between RBD and MGs are coloured in magenta.

Quaternary structure

The first low resolution models appeared in 1968 and they were obtained by cryo-electron microscopy (cryo-EM) and negative-staining (Andersen et al., 1995; Bloth et al., 1968). However, a crystal structure of a medium resolution (4.3 Å) of methylamine-induced hα2M (closed state or fast form) was not reported before 2012 (PDB 4ACQ; Marrero et al. 2012) (Fig. 15).

hα2M is a homotetramer of approximately 210 Å in length and 140 Å in width and depth (Fig. 15A,B) (Marrero et al. 2012). It is also considered a "dimer of dimers" with a top and a bottom dimer, in a head-to-tail orientation. Each dimer is composed of two monomers covalently bound by two symmetric interchain disulphide bonds (through C^{278} in MG3 domain from one monomer and C^{431} in MG4 domain from the other monomer). In turn, the dimers are joined by non-covalent interactions to form a tetrameric molecule (Marrero et al. 2012; Wyatt et al., 2012). Native tetramers of hα2M can dissociate into dimers at low protein concentration (10 μg/mL), by reduction of the intra-chain disulphide bonds, by salt-induced dissociation or by pH (Boisset et al., 1996; Kolodziej et al., 2002). The opposite will happen (dimers will form tetramers) in presence of peptidases and chemical inducers (like methylamine, MA). It has been reported that the induced hα2M tetramers are more robust and quite more difficult to dissociate (Barret et al., 1979).

The four thioester-bond sites are in the middle of the tetramer (called "prey chamber", ~60 Å) and are easily accessible for reactive lysines from peptidases (Fig. 15D, E, F, G). In addition to this space, there are other four "substrate ante-chambers" (Fig. 15G) (Marrero et al., 2012) (Goulas et al., 2017). Overall, the prey chamber and thus, the tetramer, have space for two peptidases of 20-30 kDa (one per disulphide-linked dimer) (Fig. 15H) or only one large molecule of 80-90 kDa (Garcia-Ferrer et al., 2017; Goulas et al., 2017; Marrero et al., 2012). Bound proteases are trapped, but still active so that, they can process small substrates of 6-9 kDa (Barrett and Starkey, 1973). These small substrates can arrive to the prey chamber through one of the twelve apertures of hα2M (Fig. 15A, B, C). The interaction of the 20 kDa-soybean trypsin inhibitor with trapped trypsin in the prey chamber has been reported (Fig. 15I) (Bieth et al., 1981; Song and Suh, 1998). After the activation/induction of hα2M, the aperture size is reduced. This prevents non-covalently bound peptidases or non-peptidases from escaping but also allow small protein substrates or inhibitors to enter (25-30 Å) and interact with the internalized proteins (Barret and Starkey, 1973; Bieth et al., 1981; Travis and Salvesen, 1983).

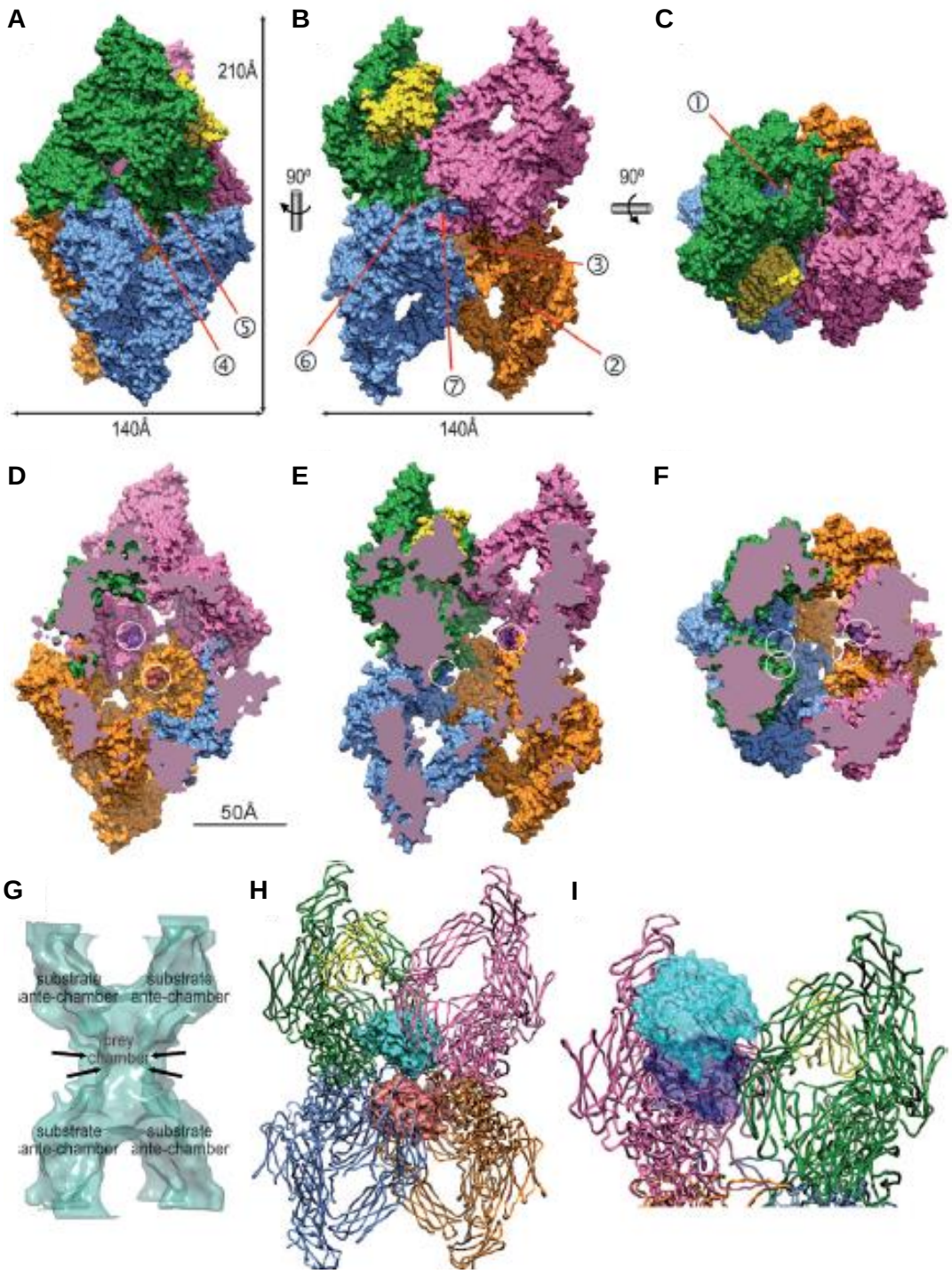


Figure 15. Crystal structure of methylamine-induced hα2M (4.3 Å) published by Marrero and her colleagues in 2012 (PDB 4ACQ); (Figure reproduced from Marrero et al. 2012). Surface plot of the tetramer in X-view (A), H-view (B) and end view (C). Each monomer is represented with a different color. The RBD domain is only visible in one monomer and it is shown in yellow. The entrances to the prey chamber are marked by circled numbers. The size of the tetramer is also indicated. Surface representation of the tetramer in X-view (D), H-view (E) and end-view (F) showing the encircled thioester sites into the central cavity. (G) Representation of the inner-cavity surface in H-view, localizing the central prey chamber and the four substrate ante-chambers. Arrows tag the narrowest part of the prey chamber. (H) Representation of the tetramer in H-view with two trapped HIV-1 proteinases. (I) Proposed interaction between the soybean trypsin inhibitor with porcine trypsin (in blue: inhibitor; in purple: proteinase) (PDB 1AVW; Song and Suh, 1998).

1.5.5 Mechanism of action, peptidase entrapment and conformational changes of hα2M

The hα2M tetramer operate through a unique irreversible "Venus flytrap" mechanism (Conrad, 1998; Marrero et al., 2012), through which it traps peptidases into its inner cavity, but at the same time, this mechanism does not block them, because the peptidases remain active against small substrates or inhibitors that are able to cross the hα2M entrances, and in this way, they are removed from the circulation. This mechanism is a three-step process (Armstrong, 2010) that involves, in turn, three reactive sites of the hα2M structure (Fig. 16):

1) Bait region: The peptidase attack of the bait region in the BRD domain, which contains many potential recognition sites, results in a fast-conformational change from the native to the nascent or induced states of hα2M. The cleavage of the bait region also exposes an internal and buried thioester bond within the TED domain, this is called the "nascent" state of hα2M (Goulas et al., 2017; Marrero et al., 2012; Sottrup-Jensen et al., 1981) (Fig. 16A2).
2) Thioester bond: This β-cysteinyl-γ-glutamyl thioester bond is very labile and ready for being cleaved by small nucleophiles from surface lysines of the trapped endopeptidase (Armstrong and Quigley, 1999). Cleavage of the thioester bond results in a free cysteine and a ε-lysyl-γ-glutamyl cross-link between the side chains of the thioester glutamine and the lysine from the endopeptidase (covalently bound) (Salvesen et al., 1981; Sottrup-Jensen, 1994). This cleavage causes the conformational state change of hα2M from nascent to induced (Goulas et al., 2017) (Fig. 16A3, B2).

   Independently of the bait region hydrolysis, the internal thioester bond can also be cleaved by small nucleophiles, like the methylamine (MA), ethylamine or ammonia (Barret et al., 1979;) and by heat (Doan and Gettins, 2008), resulting in an induced hα2M with an intact bait region, but this hα2M no longer inhibits peptidases (Armstrong and Quigley, 1999; Sottrup-Jensen et al., 1980; Travis and Salvesen, 1983). In this case, the conformational change is slower and with intermediate species (Van Leuven, 1982).

   Some peptidases can be efficiently trapped after the bait region cleavage, without covalent binding from the thioester bond. This is the case for example of human neutrophil elastase that is able to cut the bait and to induce the conformational change but due to the lack of a surface lysine, there is no covalent binding with the ha2M (Enghild et al., 1989). Nevertheless, inhibition can be exerted efficiently as long as the protease inf engulfed within hα2M.

   The conformational switches of the hα2M tetramer are revealed by low-resolution electron microscopy (Kolodziej et al., 2002), native electrophoretic mobility studies (Barret et al., 1979; Leuven et al., 1981) (Fig.16D) and sedimentation coefficients (Björk and Fish, 1982). Thus, the two states of hα2M can easily being distinguished between "native" (or "slow "mobility in the

native gel because of its "opened" conformation) (Fig. 16A1, B1, C1, D1) and "induced" (or "fast" mobility in the native gel because of its "closed" conformation) (Fig. 16A3, B2, C2, D2).

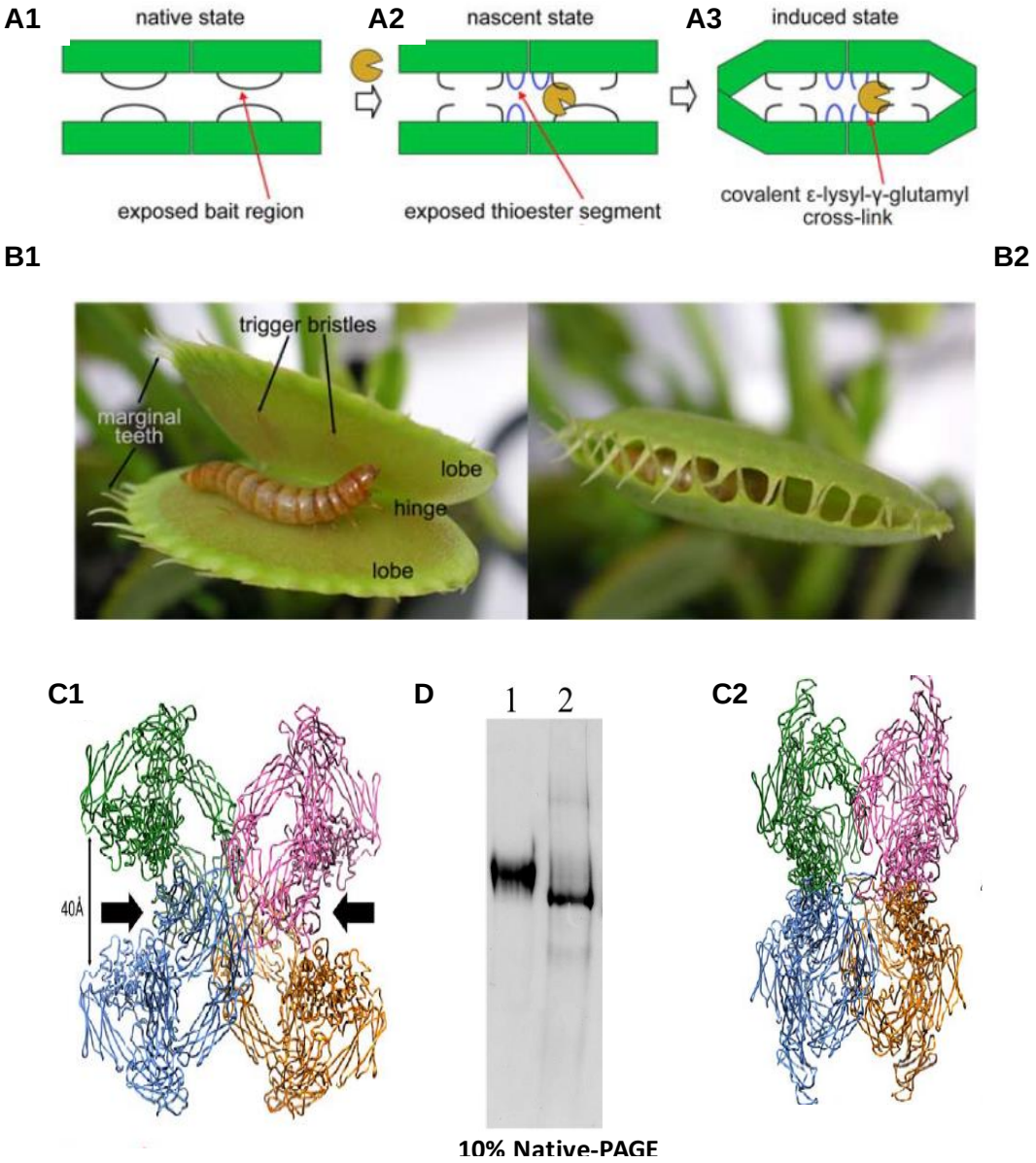


Figure 16. "Venus flytrap" mechanism of hα2M (Figure reproduced from Goulas et al., 2017)**. (A)** Representation of endopeptidase neutralization through tetrameric hα2M. **(A1)** The native state is opened, and the bait region is exposed and accessible for a prey peptidase (yellow Pac- Man). **(A2)** Once the bait region is cleavage, the inhibitor enters a nascent state, in which the thioester-segment—which was hidden in the native state—is exposed and becomes targetable by lysines on the surface of the peptidase. **(A3)** This results in a covalent cross-link. Overall, conformational rearrangement leads to the closed induced state of the tetramer. **(B)** Comparison of mechanisms (hα2M & Venus flytrap *(Dionaea muscipula*) (the photo is courtesy of Beatrice Murch (http://www.beatricemurchphotography.com).. **(B2)** It closes when the prey is touched, trapping the insects. **(C)** Modelled hα2M tetramer with each monomer in one color, for native **(C1)** and induced **(C2)** states. Black arrows point to the potentially large entrances to the central cavity, which would span ~40Å in diameter**. (D)** Native electrophoretic mobility studies, showing the differences gel mobilities for native **(D1)** and induced **(D2) h**α2M.

3) Receptor binding site: After induction by peptidases, hα2M exposes the RBD domain, which is recognized by the low-density lipoprotein receptor-related protein (LRP1) (Andersen et al., 2000) (Fig. 17) The complex peptidase-hα2M is internalized by endocytosis and degraded in the lysosomes, ensuring the clearance of the inhibitor with its cargo from the circulation in a fast process. In native hα2Ms, the RBD domains are still hidden and cannot interact with the receptor (Debanne et al., 1976; Strickland et al., 1990), being an exclusive mechanism for removing peptidases and all the interacting proteins that are bound afterwards to the induced state, like growth factor, cytokines and misfolded proteins. So induced hα2M also seems to serve as a system for modulating or preventing the harmful systemic or local effects of excess of cytokines like TGF-β, toxins such as ricin-A and misfolded pathogenic proteins by removing them from the circulation with LRP1 (Ghetie et al., 1991; LaMarre et al., 1991a; Wyatt et al., 2013a; Wyatt et al., 2014).

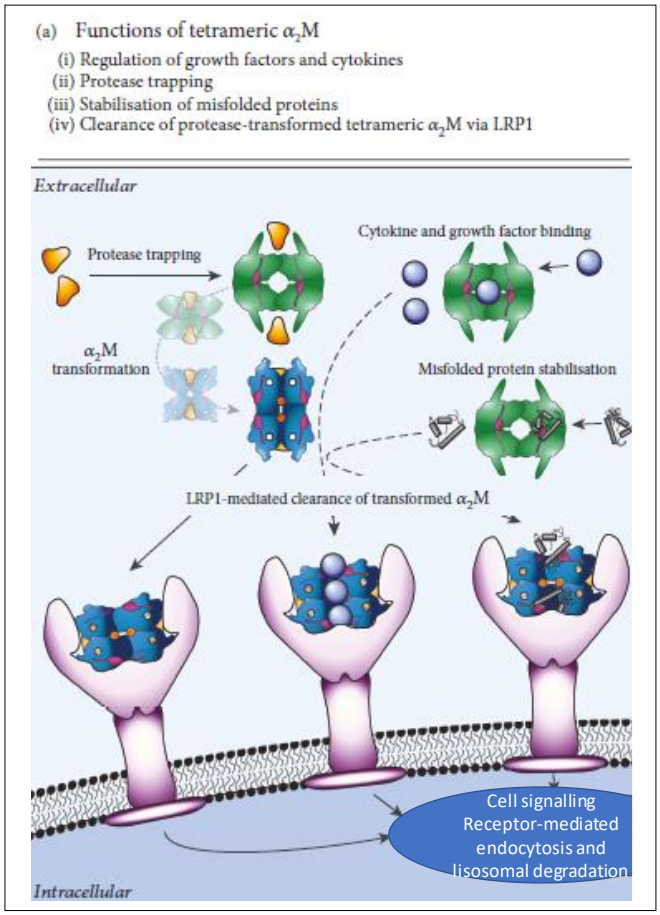


Figure 17. Schematic representation of interaction of LRP1 with peptidase-hα2M (Figure reproduced from Cater et al., 2019). Native hα2M (shown in green) results in a conformational change that exposes the binding site on hα2M for LRP1 (the RBD domain) when it is bound to proteases, which is responsible for the clearance of the peptidase-hα2M complex (shown in blue). hα2M also binds to many noncovalent ligands including cytokines, growth factors and misfolded proteins. Noncovalent binding of ligands occurs preferentially to the peptidase-hα2M conformation. Protease interaction is required to enable clearance of the complex via LRP1, when native α2M binds noncovalently to a non-protease substrate.

1.5.6 Biological interactions of hα2M with other molecules

hα2M regulate proteolysis in complex biological processes (Garcia-Ferrer et al., 2017), but in addition to binding and neutralization of peptidases, it modifies and modulate the activity of cytokines, hormones, growth factors, lipid factors and other proteins, and thus have a great impact on human physiology. A characteristic example is the interaction of hα2M with transforming growth factors-β (TGFβs), a family of ~25 kDa structurally homologous dimeric proteins. In mammals, the TGFβ family has three members (TGFβ1, TGFβ2 and TGFβ3), which share 70% sequence identity and similar three-dimensional structures (del Amo-Maestro et al., 2019). Their biological activity includes growth regulation, transcriptional activation of extracellular-matrix-related genes and chemotactic activity (LaMarre et al., 1991a; Liu et al., 2001). The predominant form of native latent TGFβs secreted by platelets and most cells in culture are the 235 kDa complex in which a 135 kDa modulator protein (the latent TGFβ binding protein, LTBP) is disulphide bonded to the latent TGFβ (Miyazono et al., 1988; Wakefield et at., 1988), the ~100 kDa disulphide latent TGFβ form expressed in recombinant cells (Zou and Sun, 2006) and the 750 kDa complex between mature TGFβ bound to α2M (Huang et al., 1988;O'Connor-McCourt and Wakefield, 1987; Wakefield et al., 1989) (Fig.18). They are primarily regulated by the non-covalently attached N-terminal latency-associated domain (LAP) (Annes et al., 2003), which acts as a pro-domain in the latent ~100 kDa pro-forms (pro-TGFβs) (Fig. 18B). Once in circulation, LAP (75 kDa disulphide homodimer) is removed and the mature TGFβ (25 kDa disulphide homodimer) availability is regulated by hα2M (Fig. 18C)., which sequesters most of these cytokines. through a currently unknown mechanism (Liu et al., 2001; O'Connor-McCourt and Wakefield, 1987; Webb et al., 1998). What is known is that hα2M positions E737, D742 and E753 within segment V723-T761 (numbering according to UniProt (UP) entry P01023) are involved in TGFβ1 binding (Arandjelovic et al., 2006) and that through them induced hα2M binds the cytokine with higher affinity than the native inhibitor (Arandjelovic et al., 2006).

TGFβ2 is the member of the TGFβ family that binds native and induced α2Ms with equivalent affinity (Crookston et l., 1994). It is also known that the TGFβ2 sequence from residues 40-47 (N^{40}ANFCAGA^{47}) and specially the residues Asn^{40}, Ala^{45}, and Ala^{47} of the mature protein (UP entry P61812) play an important role in determining affinity for native hα2M but not induced hα2M (Webb et al., 1994). It has also been hypothesized that TGFβ2 form complexes with induced hα2M through the interaction of the hydrophobic region around the Trp^{52} in TGFβ2 (Liu et al., 2001) and a hydrophobic region (residues 591–774) in induced hα2M (Gonias et al., 2000).The hydrophobicity is known to increase after the conversion of native hα2M to induced hα2M by treatment with peptidase or methylamine (Birkenmeier and Kopperschlager, 1994).

There is a controversy in the literature about the biological significance of TGFβ binding to hα2M. Cytokines which are bound to induced hα2M may be targeted to the α2M receptor/LRP (LaMarre et al., 1991), being a process that mediates cytokine clearance and regulates cytokine activity (Crookston et al., 1994). Whereas, cytokines which are bound to native hα2M may be stabilized (Philip and O'Conno-McCourt, 1991) and protected from protease cleavage (Matsuda et al., 1989), serving as a protective carrier protein (Feige et al., 1996).

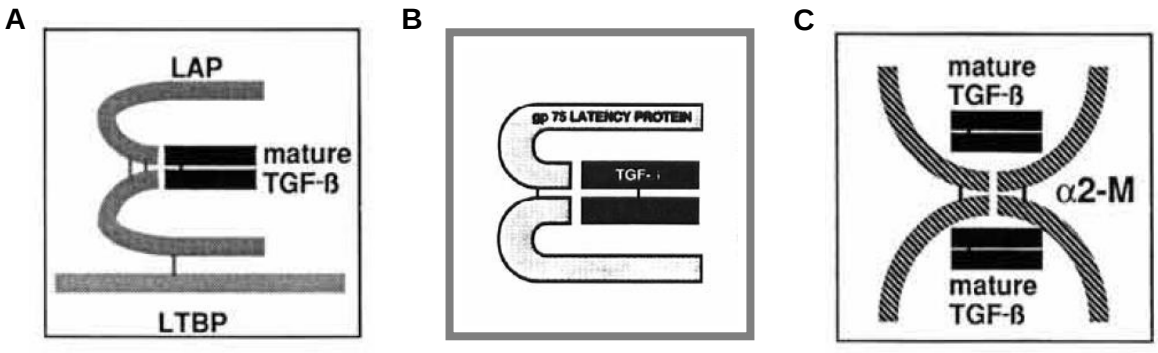


Figure 18. Schematic representation of the biologically latent forms of TGFβs (Figure reproduced from Wakefield et al., 1989)**. (A)** It is the secreted 235 kDa form by platelets, **(B)** the latent 100 kDa TGF-β form expressed in recombinant cells and **(C)** the 750 kDa complex between mature TGF-β and hα2M.

The functional and structural properties of hα2M are also exploited by pathogens such as *Streptococcus pyogenes* (group A streptococci), which forms stable interactions with native hα2M by using a surface protein, GRAB (Godehardt et al., 2004; Rasmusen et al., 1999). This 23 kDa protein consists of a Gram-positive membrane anchor motif, a variable number of 28-residue repeats, and a highly conserved N-terminal domain responsible for the interaction with hα2M. Two binding motifs of ten amino acids length, located in the N-terminal domain of GRAB, have been identified with the sequences PRIIPNGGTL (aa 41–50) and NAPEKLALRN (aa 56–65), respectively (numbering according to UP, entry Q7DAL7). The Arg^{42} and especially, the Arg^{64} seemed to play a critical role for the binding (Godehardt et al., 2004).

By recruiting native hα2M to the membrane, GRAB provides *S. pyogenes* with a mechanism to inhibit host peptidases, which protects bacterial surface structures and facilitates progressive dissemination in the infected tissue (Godehardt et al., 2004; Nyberg_et_al-2004) (Fig. 19). These interactions have only been preliminary characterized (Andersen et al., 1994; Goulas et al., 2014), and the mechanisms are still unknown.

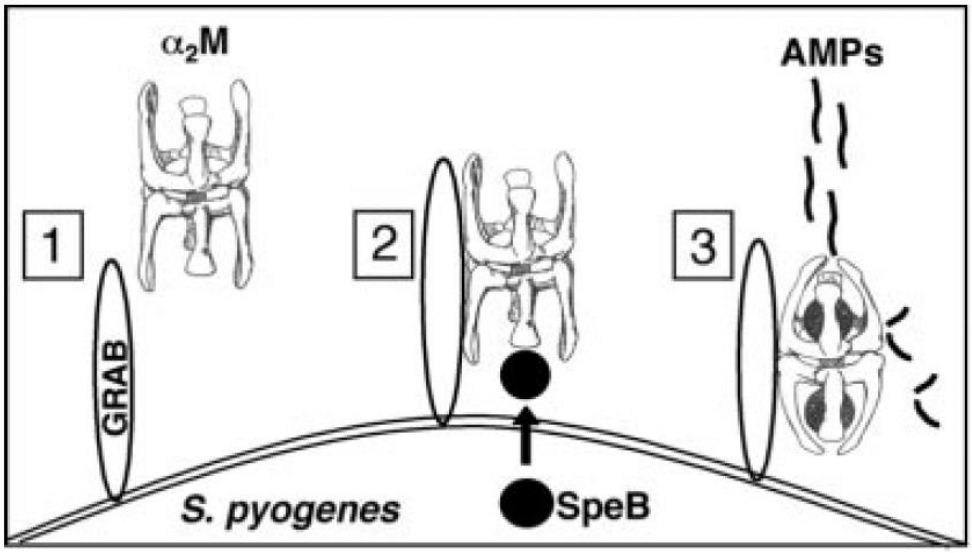


Figure 19. Schematic representation of the *Streptococcus pyogenes'* protective shield against antimicrobial peptides (AMPs) through the interaction of surface GRAB with native-hα2M (Figure reproduced from Nyberg_et_al-2004). **(1)** GRAB is expressed at the bacterial surface during early growth phase and bound to native hα2M during the inflammatory response in the infection. **(2)** During the infection, *S. pyogenes* peptidases, like SpeB, are also released and bound to the GRAB-native hα2M complex that is located in the surface. **(3)** This last interaction makes the kidnapping of small AMPs possible, such as LL-37 (4.5 kDa), that are able to enter induced hα2M.

Another interesting interaction is between the RBD domain of hα2M with its receptor, LRP1. This 600 kDa-protein is the receptor of multiple protein ligands for endocytic degradation. The ligand binding domains of LRP1 are called complement-like repeats (CR). Each CR is constituted by 40-42 residues, with three disulphide bridges and binds one calcium ion (Daly et al., 1995).

One of the best studied ligands of LRP1 is hα2M. It is known that the interaction of LRP1 with hα2M is exclusively with the induced state, which has exposed the RBD domain, and it is localized in the cluster 2 of LRP1 (Williams et al., 1994) (Fig. 20). The cluster 2 is composed by eight CR domains (CR3-CR10). The highest binding affinity among them is for the CR345 (Dolmer and Gettins, 2006). Biochemical studies of RBD in complex with CR are available, but not structural models which confirm all this information.

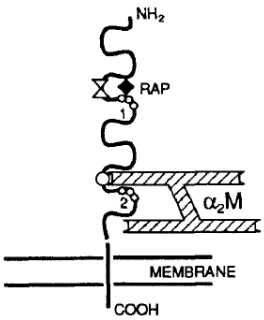


Figure 20. Schematic representation of the protein interaction between induced hα2M and cluster 2 of LRP1 (Figure reproduced from Williams et al., 1994).

Figure reproduced from “Chansons serrano80”.

1.6 Aims and general objectives of the study

The first goal of the thesis was to establish a system for heterologous expression of human glycoproteins with high yield, purity and homogeneity, incorporating the necessary flexibility to engineer the protein at will

All the current studies of the human subtilisin-kexin-isozyme 1 (hSKI1) are based on protein purified from total cell lysates with no structural information available. Thus, the major goals of this part of the thesis were:

- The establishment of an efficient heterologous expression system of human SKI1 for subsequent biochemical and biophysical characterization.
- To establish an adequate purification protocol to obtain pure protein for subsequent structural studies.

To shed light on the molecular features of the intricate working mechanism of the protein human α2-macroglobulin (hα2M), we were interested in working on:

- The isolation and purification of wild-type or recombinant hα2M variants.
- The biochemical characterization of them.
- To perform structural studies in order to improve he resolution of the wild-type hα2M (native and induced) by crystallising the recombinant hα2M variants (alone or in protein complexes).
- The expression of GRAB, pro-TGFβ2 and LRP1 receptor in bacteria (*Escherichia coli*) or eukaryotic cells like *Drosophila* Schneider 2 or human Expi293F™ cells.
- The biochemical, biophysical and structural characterizations of the interactions between wild-type hα2M (native and induced forms) with GRAB, pro-TGFβ2 and LRP1.
- And finally, to map the interactions by using the recombinant hα2M, GRAB, pro-TGFβ2 and LRP1.

The matrix metalloproteinase (MMP) family has mainly been studied in vertebrates. Here, we review the literature available on MMPs outside vertebrates and perform database searches for potential MMP CDs in invertebrates, plants, fungi, viruses, protists, archaea and bacteria.

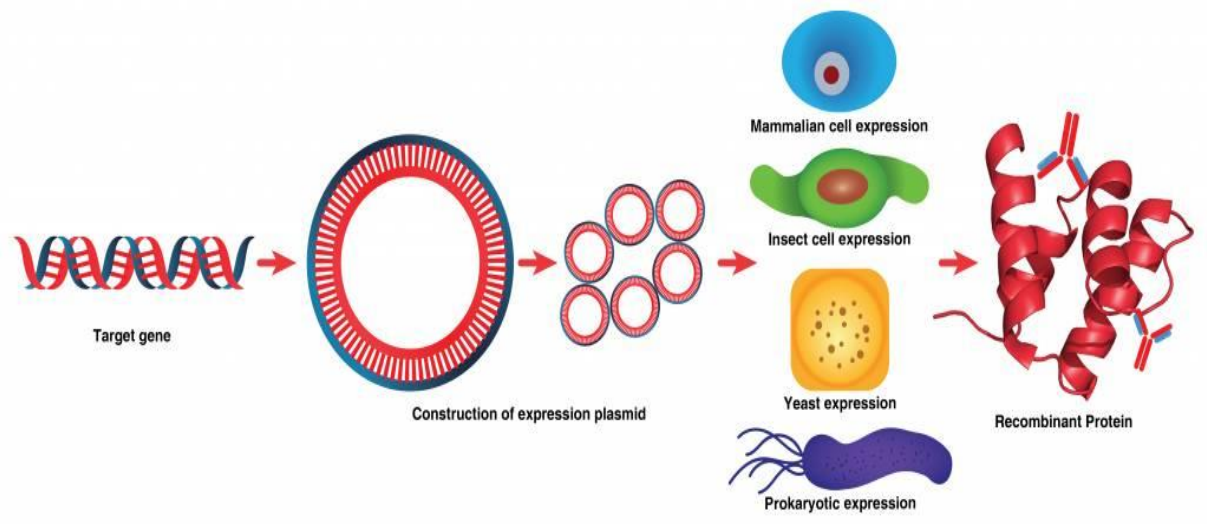


Figure reproduced from “cloud-clone.com”.

MATERIALS AND METHODS

2.1 *In silico* sequence analysis tools

In silico tools allow to analyze and predict the features of a particular DNA/protein to perform the best possible experimental design, first for creating DNA constructs and then, for designing protein purification and polishing process.

In this thesis, the following *in silico* sequence analysis tools have been used:

- For signal peptide prediction: - SignalP-5.0 server (http://www.cbs.dtu.dk/services/SignalP/)
  - Phobius prediction server (http://phobius.sbc.su.se/)
  - Topology Database of Transmembrane Proteins server (TOPDB) (http://topdb.enzim.hu/)
- For transmembrane domain prediction: - Prediction of transmembrane helices in proteins (http://www.cbs.dtu.dk/services/TMHMM/)
  - https://www.predictprotein.org
- For functional domain prediction: - Prosite Expasy server (https://prosite.expasy.org/)
- For hydropathicity level prediction: - https://web.expasy.org/cgi-bin/protscale/protscale.pl server and including "Hphob. / Kyte & Doolittle" in the search
- For disulphide bond prediction: - Scratch Protein Predictor server (http://scratch.proteomics.ics.uci.edu/explanation.html)
- For N-glycosylation prediction: - http://www.cbs.dtu.dk/services/NetNGlyc/
- For subcellular localization prediction: - DeepLoc-1.0 server (http://www.cbs.dtu.dk/services/DeepLoc/)
- For secondary structure prediction: - http://bioinf.cs.ucl.ac.uk/psipred
  - http://polyview.cchmc.org/
  - https://npsa-prabi.ibcp.fr/cgi-bin/secpred

Databases are other powerful tool to improve our knowledge about DNA/protein targets. The databases used in this research were:

- For peptidases (catalysed reaction and substrate): https://www.qmul.ac.uk/sbcs/iubmb/enzyme/EC3/4/.
- For peptidases (evolutionary scheme): http://www.ebi.ac.uk/merops/
- For peptidase inhibitors: https://www.ebi.ac.uk/merops/inhibitors/
- For any described protein: https://www.uniprot.org/uniprot
- For deposited structures: Protein data bank (PDB) server: https://www.rcsb.org/
- For the computation of various physical and chemical parameters for a given protein: https://web.expasy.org/protparam/

To plan and design the DNA clonings and easily analyze the results from DNA sequencing and N-terminal sequencings, we used the following software were:

- SnapGene (https://www.snapgene.com/).
- Geneious (https://www.geneious.com/)

2.2 Recombinant DNA techniques

To get specific and efficient amplification, DNA polymerase must be considered, pure DNA template must be used (without inhibitory contaminants (like detergents, SDS, RNA…) and primers should be accurately designed. The general rules for optimal primer design are (see Chapter 8 of Molecular Cloning: A laboratory manual; Sambrook and Russell, 2001):

- Base composition: G+C content between 40% and 60%.
- No use of self-complementary primers: As a general rule, no more than three consecutive nucleotides on one primer should be complementary to the other primer.
- Length of the primer complementary to the template: Between 18-25 nucleotides long and the primer pair should not differ in length by > 3 bp.
- Melting temperature (Tm): A primer pair should not differ by > 5°C. This can be checked by using the following servers: https://tmcalculator.neb.com/#!/main
  http://biotools.nubic.northwestern.edu/OligoCalc.html
- 5' Termini: To guarantee an efficient restriction enzyme cleavage, the primer should be extended by at least three additional nucleotides beyond the recognition sequence of the restriction enzyme. It is also useful to check the recommendations from the New England Biolab's catalogue about each restriction enzyme.
- 3'Termini: The 3' base of each primer should be G or C. However, primers ending in CG or CG sequences are not recommended because they assist the hairpin structures formation and may generate primer dimers.

In this thesis, constructs spanning different fragments of the genes coding for hSKI1 (UP Q14703), hα2M (UP P01023), GRAB from *Streptococcus pyogenes* serotype M1 (UP Q7DAL7) and for human pro-TGFβ2 (UP P61812) (for details on constructs, plasmids, vectors and primers, see Tables 9, 10, 11, 19, 22, Fig. 46, 53, 58, 66, 69 and in the section "List of expression vectors") were amplified with primers that introduced either restriction sites for enzyme digestion mediated directional cloning or vector sequence complementary overhangs for restriction-free (RF) cloning. First strategy is the "traditional" cloning, which can be hampered by a lack of appropriate restriction-sites and inefficient enzymatic

steps, and these are the advantages of the RF cloning (Stevenson et al., 2013). We also introduced punctual mutations by site-directed mutagenesis, which allows easily modify genes and like that, study the structural and functional properties of a protein (Yang et al., 2017).

The vectors used were pCri-4a, -7a, -8a, 9a (Goulas et al., 2014a) and pSP9 for bacterial expression; pMT/V5-His, pMT/BiP/V5-His (Invitrogen), pHroE2 (Excellgene) and pIEx (Novagen) for expression in *Drosophila melanogaster* Schneider 2 embryonic cells (S2; Gibco); and pXLG40 (Excellgene) and pCMV-Sport 6 (Thermo Scientific) for expression in Chinese hamster ovary (CHO; Excellgene) and human Expi293F™ cells (Gibco). PCR primers and enzymes were purchased from Sigma-Aldrich and Thermo Scientific, respectively. PCR was performed using Phusion^{R} High Fidelity DNA polymerase (Thermo Scientific) according to the manufacturer's instructions (Table 2) and following a standard optimization step by thermal gradient in each reaction.

A

Component	50 µL Reaction	Final Concentration
Nuclease-free water	to 50 µL	
5X Phusion HF	10 µL	1X
10 mM dNTPs	1 µL	200 µM
10 mM Forward primer	2.5 µL	0.5 µM
10 mM Reverse primer	2.5 µL	0.5 µM
Template DNA	variable	<250 ng
Phusion DNA Polymerase	0.5 µL	1.0 units/50 µL PCR

B

Step	Temp	Time
Initial Denaturation	98°C	30 sec
35 Cycles	98°C 45-72°C 72°C	10 sec 15 sec 15-30 sec/kb
Final Extension	72°C	10 min
Hold	4-10°C	

Table 2. Standard PCR protocol for Phusion^{R} High-Fidelity DNA Polymerase (New England Biolabs protocols). **(A)** PCR mix. **(B)** Thermocycling conditions.

2.2.1 Restriction enzyme directional cloning

The amplified DNA insert was purified with the OMEGA Biotek Cycle Purification Kit according to the manufacturer's instructions. The cloning vector (where the DNA insert is going to be introduced) and the clean PCR product were digested with the corresponding restriction enzymes for 1 hour at 37°C and then, purified from agarose gel with the OMEGA Biotek Gel Extraction Kit according to the manufacturer's instructions. After this, pure digested insert and vector were ligated O/N at RT with T4 DNA ligase with several insert DNA molar ratio over vector (Table 3)

Component	20 µL Reaction
Linear vector DNA	20- 100 ng
Insert DNA	1:1 to 5:1 molar ratio over vector
10X T4 DNA Ligase Buffer	2 µL
T4 DNA Ligase	1 µL
Nuclease-free water	to 20 µL

Table 3. T4 DNA Ligase protocol (Thermo Scientific protocols)

2.2.2 Site directed mutagenesis

Site-directed mutagenesis was performed by following a modified version of the one reported in the QuikChange II Site-Directed Mutagenesis Kit (from Agilent Technologies): standard PCR (Table 2) but with special primers (overlapping primers of 26-28 bp of length, Tm between 60-72°C, G+C content >40% and desired mutation in the middle of the primer, only introduced in one of the primers). The PCR reaction was followed by digestion for 2 hours at 37°C of 9 µL methylated parental DNA with 1 µL *Dpn*I.

2.2.3 Restriction free (RF) cloning

RF cloning was performed following the previously described protocol (Van den Ent and Löwe, 2006). In summary, a standard PCR (Table 2) in a total volume of 100 µL was purified with the Gel Extraction kit and eluted in a final concentration of 100 ng/µL, to be used as a primer pair in a second PCR reaction (where the template is the vector in which we are going to introduce the primer pair) (Table 4). Once completed, 9 µL of this PCR product was treated with 1 µL of *Dpn*I for 2 hours at 37°C to digest the methylated parental plasmid, using the resulting mix for transforming bacterial cells.

A

Component	50 µL Reaction	Final Concentration
Nuclease-free water	to 50 µL	
10X Reaction buffer	5 µL	1X
10 mM dNTPs	1 µL	200 µM
PCR (100 ng/µL)	2- 4 µL	400 ng
Template Vector (5 ng/µL)	5-10 µL	40 ng
Phusion DNA Polymerase	1 µL	2.0 units/50 µL PCR

B

Step	Temp	Time
Initial Denaturation	95°C	30 sec
35 Cycles	95°C 55-72°C 68°C	30 sec 1 min 2 min/kb
Hold	4-10°C	

Table 4. RF-PCR protocol for RF cloning (Van den Ent and Löwe, 2006). **(A)** RF-PCR mix. **(B)** Thermocycling conditions

2.2.4. Bacterial transformation with recombinant vectors

For all the types of clonings, chemical transformation of chemically competent *E.coli* DH5α cells (Novagen) were performed, according to Hanahan (Hanahan, 1983). Briefly, 50 µL of competent cells were transformed with 5 µL of each ligation(for directional cloning) or 10 µL from the *Dpn*I digestion (site directed mutagenesis and RF cloning), incubated for 30 min in ice, heat shocked for 60-90 sec at 42°C, followed by 5 min incubation on ice and subsequent incubation in lysogeny broth (LB) medium for 45 min at 900 rpm and 37°C, before being centrifuged for 5 min at 4000 x g. Finally, the resulting cell pellet from the centrifugation was plated into the appropriate antibiotic-supplemented LB-agar plate. After O/N incubation at 37°C, isolated colonies with regular sizes were picked from the plates and grown in antibiotic-supplemented LB for plasmid amplification (mini-preps). Positive clones were identified by colony PCR (sung a diluted mix of cells instead of template DNA) and confirmed by DNA sequencing (https://www.eurofinsgenomics.eu/en/custom-dna-sequencing/gatc-services/).

2.3 Recombinant protein expression

Once the target genes have been inserted into the desired expression vectors, their expression, solubility, yield and functionality were checked. The protein expression methods depend on the chosen protein expression system (Table 1) and in this thesis, three expression systems have been used (in bacterial, insect and mammalian cells) (Fig. 21, 22).

2.3.1 Protein expression in bacteria

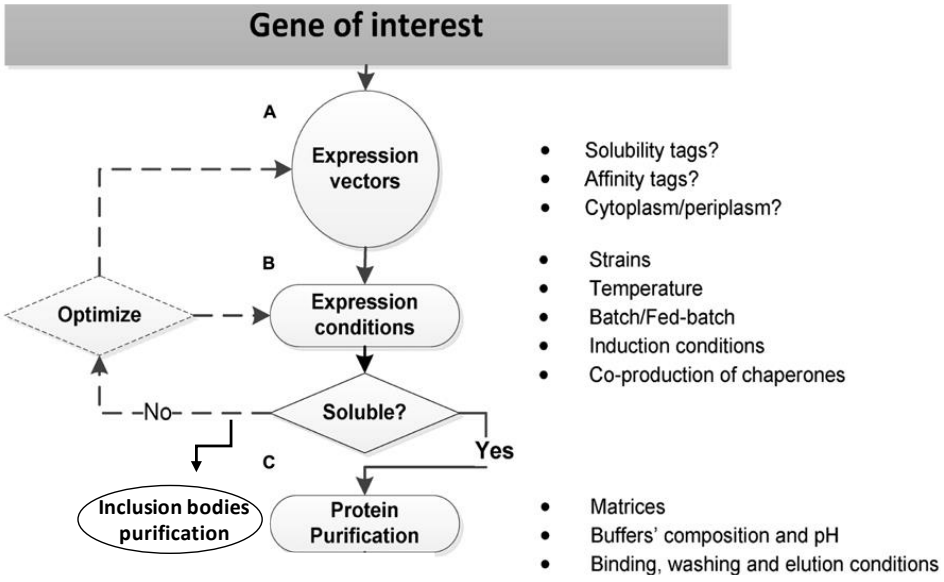


Figure 21. Workflow of the strategies for protein production in *E. coli* cells (Figure reproduced from Costa et al., 2014).

Several hSKI1 fragments were inserted into different pCri vectors (Goulas et al., 2014a) (see Table 9). Into pCri9a vector for periplasmatic expression, pCri7a vector for cytoplasmatic expression and pCri4a vector with a thioredoxin tag and a TEV cleavage site to assist in protein folding and purification. Apart from the *E.coli* BL21(DE3) strain, other engineered strains were used to improve protein solubilization, like Rosetta and Origami 2 (Baca and Hol, 2000; Sorensen et al., 2003). Also, several isopropyl-β-D-thiogalactopyranoside (IPTG) concentrations (0.1, 0.4 and 1 mM) for protein induction and expression conditions (6h at 37°C or O/N at 20°C or 18°C) were assayed.

A standard protocol to improve the obtention **soluble overexpression in bacterial cells** was the expression of GRAB, which was cloned into the pCri8a vector with two affinity tags strategies (pCri8a-H6-TEV-GRAB with N-terminal His6x-tag and pCri8a-H6-TEV-GRAB-STREP with an additional C-terminal Streptactin®II tag (Strep-tag; IBA Life Sciences) (see Fig. 61 and Table 22) and with a TEV recognition sequence. These constructs were transformed into *E. coli* BL21 (DE3) cells (Novagen, Hanahan et al., 1983), and cultures were grown in LB supplemented with 30 µg/mL kanamycin. After initial growth at 37°C to an OD600 ≈ 0.6, cultures were cooled to 20 °C, and protein expression was induced with 0.4 mM IPTG for 18–20 hours.

2.3.2 Protein expression in insect cells

During this thesis, three insect cell lines were used, High Five (Hi5) cells from *Trichoplusia ni* cells, Sf9 from *Spodoptera frugiperda* cells and Schneider 2 (S2) from embryonic *Drosophila melanogaster* cells. But we focus our efforts in the last one, where all the protocols were established (with the kind help of Philippe Leone, afmb; Marseille, France).

2.3.2.1 Generation of *Drosophila* Schneider 2 (S2) cells adapted to suspension

Drosophila S2 cells were originally grown in plates (monolaters) and in Sf900™II SFM medium (Gibco), supplemented with 10% FBS, 0.5 µg/mL of the antimycotic Fungizone, 100 units/mL of penicillin, and 100 µg/mL of streptomycin sulphate (Gibco). In an attempt to increase the protein yield, however, they were adapted to grow in suspension in the same medium but no-supplemented with FBS. Briefly, 5 mL of monolayer cells at 4 × 10^6 cells/mL were seeded in TS50 tubes at 180 rpm and 28°C (initially with 10% FBS in the medium) and diluted twice-three times per week to 4 × 10^6 cells/ml and then, the lysed cells were removed by centrifugation (800 x g for 5 min) (living cells remain in the pellet). When these cells were no longer precipitating and the cell growth was exponential, the FBS concentration was reduced (first at 5%, then at 2%, 0.5% and finally, 0.0%) and before every new batch, frozen cell aliquots were generated.

2.3.2.2 Maintenance of S2 cell cultures

S2 cells were cultured in Sf900™II SFM medium in TubeSpin bioreactor tubes (TS50 for 5 to 10 mL cultures and TS600 for 100 to 200-mL cultures; Techno Plastic Products AG) as previously described (Xie, Q. et al., 2011). Cells were passaged three times per week to a final density of 4×10^6 cells/mL. Cells were incubated at 28°C in a shaker (Brunswick Scientific Innova) under agitation at 220 rpm.

2.3.2.3 Generation of S2 cells adapted to different media

Drosophila S2 cells were also adapted to different insect media like TC100 and IPL-41 (both from Gibco), and these adaptations were performed like in the section 2.3.2.1. Thus, cells were subcultured at 4×10^6 cells/mL in the new condition until being adapted (>90% viability), first, 75% old medium and 25% new medium; then, 50%:50%; next, 25%:75%; and finally, 0%:100%.

2.3.2.4 Plasmid transfection of S2 cells

For protein expression, S2 cells were transfected with two kinds of expression vectors (Fig. 22):

- For inducible-stable gene transfection: pMT/V5-His and pMT/BiP/V5-His (Invitrogen) vectors.
- For transient gene expression: pHroE2 (Excellgene) and pIEx (Novagen) vectors.

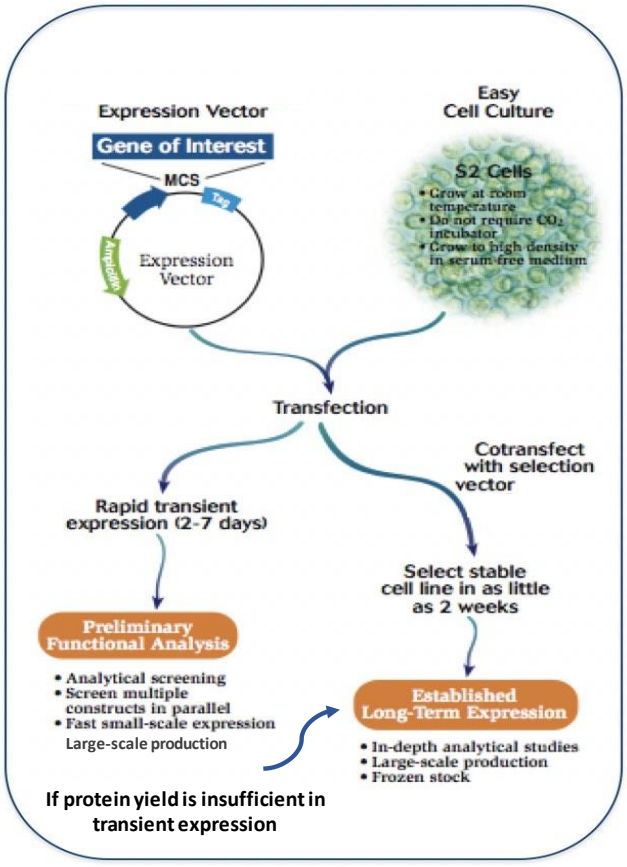


Figure 22. Workflow of the *Drosophila* expression system (Fig. adapted from Pasteur Institute, Expression platform).

Regarding the plasmid transfection, three methods were assayed:

- Transfection of S2 cells with calcium phosphate
- Transfection with Lipofectamine™ 2000 (Thermo Scientific)
- Transfection with linear 25 kDa polyethylenimine (PEI) (Polysciences Europe GmbH

2.3.2.4.1 Calcium phosphate transfection of S2 cells

2.5M $CaCl_2$ solution was prepared in Milli-Q water, filtered with a 0.22 µm filter devise and stored at 4°C. HBS* solution (1X) (* standard HBS solution *plus* glucose at a concentration of 1 mg/mL) was prepared in Milli-Q water, pH 7.0, filtered with a 0.22 µm filter devise and stored at -20°C. 0.1M $CuSO_4$ solution was made in Milli-Q water, filtered with a 0.22 µm filter devise and stored at RT. The DNA was purified with the GeneJET Plasmid Maxiprep Kit (Thermo Scientific) and stored at −20°C in sterile Milli-Q water at 1 mg/mL.

For transfection in 6-well plates (9.6 cm^2 dishes), the day before transfection (D_{-1}) S2 cells were split in 1.5×10^6 cells/mL per well (1 ml at 1.5×10^6 cells/mL + 1 mL of fresh Sf900™II medium supplemented with 10% FBS *plus* antibiotics and antimycotics). The transfection day (D_0) with all the reagents prewarmed and homogenized, 8 µg plasmid DNA were added dropwise to 120 µL HBS* in an Eppendorf tube (per well) and vortexed for 3 sec, and then 8 µL $CaCl_2$ were also added dropwise to the mixture and vortexed for 3 sec. The final mixture was incubated for 30 min at RT and then poured drop-by-drop on the cells (one well). After transfection (D_{+1}), $CuSO_4$ was added to induce expression when it was necessary (inducible promoter). In the case of the pMT/GNBP3-hSKI1 (18-998)-H8 construct, protein induction was performed by adding of 10 µL $CuSO_4$ (500 µM final concentration). Expressions were harvested in the same medium for three-five days for protein expression (in the case of pMT/GNBP3-hSKI1 (18-998)-H8 construct, expression was harvested for three days).

2.3.2.4.2 Lipofectamine transfection of S2 cells

S2 cells were diluted to 1.5×10^6 cells/mL per well (1 ml at 1.5×10^6 cells/mL + 1 mL of fresh Sf900™II medium supplemented with 10% FBS *plus* antibiotics and antimycotics) at D_{-1} for transfection in 6-well plates (9.6 cm^2 dishes). Plasmid transfection was performed the next day (D_0), in polystyrene tubes with 7.5 µL Lipofectamine™ 2000 (Thermo Scientific) and 2.5 µg plasmid DNA per transfection well. The mixture was incubated for 30 min at RT and then poured drop-by-drop on the cells (one well). At D_{+1}, $CuSO_4$ was added up to 200 µM (final concentration) to induce expression when it was necessary (inducible promoter).

2.3.2.4.3 Polyethylenimine transfection of S2 cells

Linear 25 kDa polyethylenimine (PEI; Polysciences Europe GmbH) was prepared in Milli-Q water at a concentration of 1 mg/mL and pH 7.0. The solution was filter-sterilized and stored at −20°C. Plasmid DNA was produced in *Escherichia coli* DH5α cells, purified with the GeneJET Plasmid Maxiprep Kit and stored at −20°C in sterile Milli-Q water at 1 mg/mL.

For plasmid transfection of S2 cells in suspension, cells were diluted to 4×10^6 cells/mL at D_{-1}. At D_0, cells were centrifuged (200 x g for 5 min) and resuspended in prewarmed fresh medium (Sf900™II medium supplemented with antibiotics and antimycotics) to a cell density of 15×10^6 cells/mL ("prewarmed transfection volume"). Then, in a sterile eppendorf tube with all the reagents prewarmed and homogenized, a mixture of 0.6 µg plasmid DNA and 2 µg PEI per 1×10^6 cells per prewarmed transfection medium was pre-incubated for 15–20 min at RT and then added dropwise to the cell cultures. These were further incubated for 1 hour at 28°C and 220 rpm, subsequently diluted with prewarmed fresh medium to 5×10^6 cells/ml and harvested after five-seven days for protein purification. This was the method followed to transfect all the pIEx constructs of this thesis (Tables 10, 19 and 22).

2.3.2.5 Generation of polyclonal cell lines in S2 cells

S2 cells were grown at 28°C in tissue culture flasks for the semi-adherent monolayer cell cultures (with SF900II medium supplemented with 10% FBS, antimycotic and antibiotics) or in 50-ml flasks (with SF900II medium supplemented with antimycotic and antibiotics) shaken at 220 rpm in an orbital shaker for suspension cell cultures. S2 cells were co-transfected with the recombinant DNA (pMT/V5-His and pMT/BiP/V5-His constructs, with inducible promoter) and pCoPuro (antibiotic selection vector) at a ratio of 8:1 (construct: selection vector) using the calcium phosphate protocol previously described. After two days of growth in normal harvesting conditions, the transfected cells were selected with 10 µg/ml puromycin (Thermo Scientific) in the medium for 3 weeks and then sustained with 1 µg/ml puromycin. Then, frozen stocks were generated. Viable cell densities of these cultures were kept above 10^6 cells/mL by continuous passage. For the expression in stable cell cultures either in monolayer or in suspension, $2\text{-}5 \times 10^6$ cells/mL were induced with $CuSO_4$ at a final concentration of 500 µM. The induced cells were harvested 3 days post-induction. This was the protocol to generate the stable transfected S2 cell lines: pMT/BiP -hSKI1 (18-998)-H8, pMT/GNBP3-hSKI1 (18-998)-H8 and the controls, pMT/GFP-H8 and pMT/GNBP3-H8.

2.3.2.6 Freezing S2 cells

When S2 cells were between 1.0-2.0 × 10^7 cells/mL in a 75 cm2 flask (for monolayer cell cultures) or in 12 mL (for suspension culture) and the viability between 95-99%, were centrifuged at 800 x g for 5 min , medium was saved and the cell pellet was resuspended at a cell density of 1.1 × 10^7 cells/mL in 1 mL freezing medium. Freezing medium was 40% conditionated medium (the saved media from the cell centrifugation)) *plus* 40% fresh medium *plus* 10% FBS *plus* 10% DMSO. As soon as the DMSO was added, the 1 mL vial was frozen, first in a container (Mr Frosty™ from Thermo Scientific) at -8°C for 24h, and then, transferred to liquid nitrogen for long term storage.

2.3.2.7 Initiating S2 cell culture from frozen stock

Frozen vial was clean with ethanol and then, resuspended in 10 mL of fresh medium (SF900II medium). This tube was centrifuged at 1000 x g for 5 min and the resulting cell pellet was resuspended in 10 mL of fresh medium with 2% FBS. When cells were stabilized (3-4 days later, with >90% viability), the FBS was totally removed from the medium.

2.3.2.8 Cell counting and viability test

The viability of S2 cells was determined by the Trypan Blue test. In short, a dilution of cells in 1X PBS *plus* 0.4% Trypan Blue solution (Thermo Scientific) was prepared. Then, a hemacytometer chamber was filled with 10 µL of the dilution. Non- viable and viable cells were counted (non-viable cells were blue, viable cells were unstained) under the microscope in four 1 x 1 mm squares of one chamber and determined the average number of cells per square. Cell viability was also determined (it must be at least 95% for healthy log-phase cultures):

$$\% \text{ viable cells} = [1.00 - (\text{Number of blue cells} \div \text{Number of total cells})] \times 100$$

To calculate the number of viable cells per mL of culture, the formula used was the following (correcting the dilution factor):

$$\text{Number of viable cells} \times 10^4 \times 1.1 = \text{cells/mL culture}$$

2.3.3 Protein expression in mammalian cells

In this thesis, three different mammalian cell lines have been used: HEK293T (in monolayer), CHO (suspension) and Expi293F (suspension) cells. All three mammalian cell lines were transfected with

pure plasmid DNA , purified with the GeneJET Plasmid Maxiprep Kit and stored at −20°C in sterile Milli-Q water at 1 mg/mL; and linear 25 kDa PEI, which was prepared in Milli-Q water at a concentration of 1 mg/mL and pH 7.0. The solution was filter-sterilized and stored at −20°C.

2.3.3.1 Cell culture growth and PEI transfection of Human embryonic kidney 293T (HEK293T) cells

HEK293T cells were cultured in monolayers in DMEM medium supplemented with 2mM Glutamine, 10% FBS, 0.5 μg/mL of the antimycotic Fungizone, 100 units/mL of penicillin, and 100 μg/mL of streptomycin sulphate (all reagents from Gibco) (complete DMEM medium).

These cells were transfected with linear 25 kDa PEI, following the basis of the transfection protocol provided by Nadia Sukusu Nielsen (collaboration with the University of Aarhus). Cells were briefly diluted at D_{-3} to 4.5×10^6 cells/mL in complete DMEM medium per 100 mm plates (56.7 cm^2 dishes). Three days later (D_0), the culture medium was changed 1-4 hours prior to transfection (complete DMEM medium without FBS). After this time, with all the reagents and plasmid DNA at RT and homogenized, 18 μg plasmid DNA were mixed with 180 μL PEI (in a 10:1 ratio (PEI:DNA)) in a sterile tube and up to 2 mL (final transfection volume) with transfection medium (DMEM medium without FBS, Phenol Red, antimycotic and antibiotics) per transfection plate. This mixture was incubated for 15 mins at RT and then, added dropwise to the cell cultures. After 4-5 hours, the transfection medium was removed and new complete DMEM medium without FBS was added. Transfected cells were harvested after four days.

2.3.3.2 Cell culture growth and PEI transfection of Chinese hamster ovary (CHO) cells

CHO cells cultured in suspension at 37°C, 5% CO_2 and 180 rpm in a Multitron Cell Shaker Incubator (Infors HT) in proCHO5 medium (Excellgene, unknown recipe), diluted twice per week at 0.5×10^6 cells/mL.

These cells were transfected with linear 25 kDa PEI, following the basis of the transfection protocol provided by Excellgene (Monthey, Switzerland). Cells were briefly diluted to 2×10^6 cells/mL in proCHO5 medium at D_{-1}. At D_0, cells were centrifuged (800 x g for 5 min) and resuspended in transfection medium (M40 with linear 25-PEI, unknown recipe) to a cell density of 10×10^6 cells/mL (transfection volume). Then, 13 μg plasmid DNA per mL of transfection volume were added straight into the transfection volume/cell culture. Everything was incubated for 3 hours at 31°C, 5% CO_2 and

180 rpm. Then, the transfected cell culture was 2.5X diluted with proCHO5 medium. These were further incubated at 37°C, 5% CO_2 and 180 rpm and harvested after 4-10 days.

2.3.3.3 Cell culture growth and PEI transfection of Expi293F™ cells

Expi293F cells were cultured in suspension at 37°C, 8% CO_2 and 150 rpm in a Multitron Cell Shaker Incubator in FreeStyle™ F17 expression medium (Gibco) with 0.2% Pluronic F-68 (Gibco) *plus* 8 mM L-glutamine (Gibco) and supplemented with 0.5 µg/mL of the antimycotic Fungizone, 100 units/mL of penicillin, and 100 µg/mL of streptomycin sulphate (Gibco) (complete FreeStyle™ F17 medium).

For cell growth, 125 mL or 1000 mL polycarbonate Erlenmeyer flasks (FPC0125S and FPC1000S, respectively; Tri Forest Labware) for 25 to 30-mL and 100 to 250-mL cultures were used, respectively were used. Cells were diluted three times per week to a final density of 0.3–0.5 × 10^6 cells/mL.

For PEI transfection, cells were diluted to 0.7 × 10^6 cells/mL in complete FreeStyle™ F17 medium at D_{-1}. At D_0, cells were centrifuged (200 x g for 5 min) and resuspended in prewarmed fresh complete FreeStyle™ F17 medium at a cell density of 1 × 10^6 cells/mL (expression volume). Then, in a sterile tube with all the reagents prewarmed and homogenized, 1 mg of plasmid DNA and 3 mg of PEI in 20 mL of Opti-MEM Medium (Gibco) per liter of expression volume were mixed and incubated at RT for 15–20 min and then, added dropwise to the cell cultures. Expi293F transfections were harvested after three days at 37°C, 8% CO_2 and 150 rpm for protein purification. This was the method followed to transfect all the CMV-Sport6 constructs of this thesis (Tables 11, 19 and 22). For protein deglycosylation, 5 µM of kifunensine (Sigma Aldrich) was added after adding the transfection mixture and before the harvesting at 37°C.

2.3.3.4 Generation of Expi293F™ cells adapted to different media

Expi293F cells were originally cultured in Expi293™ expression medium (Gibco) and adapted to other cheaper media. These other media were FreeStyle293™ medium, a mixture of Expi293™ and FreeStyle293™ with a proportion 1:3 (Expi:FreeStyle) and FreeStyle™ F17 medium. In short, the adaptations were performed by subculturing the cells at 0.3 × 10^6 cells/mL in the normal medium (from now, old medium) and when they reached a cell density of >2 × 10^6 cells/mL with >90% viability (after 3–4 days), they were diluted to 0.6 × 10^6 cells/mL in a new condition (condition A: 2/3 old medium *plus* 1/3 new medium). When these cells were adapted to the condition A, which means they had reached a cell density >3 × 10^6 cells/mL with >90% viability, they were diluted to 0.6 × 10^6 cells/mL in a new

condition (condition B: 1/3 old medium *plus* 2/3 new medium). Once again, when the cells were adapted to the condition B (cell density >3 × 10^6 cells/mL with >90% viability), they were diluted to 0.6 × 10^6 cells/mL in the last condition (condition C: 100% new medium). Finally, once these cells were stabilized in the condition C= new medium, they were normally subcultured at 0.3 × 10^6 cells/mL and frozen cell aliquots were generated.

2.3.3.5 Freezing Expi293F™ cells

When Expi293F™ cells were between 0.5-2.0 × 10^6 cells/mL and viability between 95-99%, were centrifuged at 200 x g for 5 min ,the medium was saved and the cell pellet was resuspended at a cell density of 1.0 × 10^7 cells/mL in 1 mL freezing medium. Freezing medium was 45% conditionated medium (the saved media from the cell centrifugation)) *plus* 45% fresh medium *plus* 10% DMSO. As soon as the DMSO was added, the 1 mL vial was frozen, first in a container (Mr Frosty™ from Thermo Scientific) at -80°C for 24h, and then, transferred to liquid nitrogen for long term storage.

2.3.3.6 Initiating Expi293F™ cell culture from frozen stock

Frozen vial was clean with ethanol and then, resuspended in fresh medium. Then, this tube was centrifuged at 200 x g for 5 min to remove the DMSO and the resulting cell pellet was resuspended in fresh medium. When cells reached a cell density of >2 × 10^6 cells/mL with >90% viability (after 3–4 days), they were diluted to 0.5 × 10^6 cells/mL and normally subcultured.

2.3.3.7 Cell counting and viability test

The protocol followed was the same than for S2 cells (section 2.3.2.8).

2.4 Protein purification

Nowadays, the most common method for protein purification is the chromatography and there are different types of chromatographies depending on the property of the purification strategy (Table 5).

Method	Protein property
Affinity chromatography (AC)	Specific ligand recognition (biospecific or nonbiospecific)
Immobilized metal ion affinity chromatography (IMAC)	Metal ion binding
Ion exchange chromatography (IEX)	Charge
Gel filtration (GF) / Size exclusion chromatography (SEC)	Size
Hydrophobic interaction chromatography (HIC) and Reversed phase chromatography (RPC)	Hydrophobicity
Chromatofocusing	Isoelectric point

Table 5. Chromatographic purifications according to the protein properties used (Table reproduced from Protein Purification Handbook from Amersham Biosciences).

2.4.1 Recombinant protein purification

Protein purification steps were regularly performed at 4°C if not otherwise stated.

For the proteins produced in bacteria (Tables 9, 22), cells were collected by centrifugation at 6.000 × g for 30 min, washed in buffer A (50 mM Tris-HCl, 250 mM sodium chloride, pH 7.5) and resuspended in the same buffer *plus* Complete EDTA-free Peptidase Inhibitor Cocktail Tablets and DNase I (both from Roche Diagnostics). Cells were lysed with a cell disrupter (Constant Systems) at a pressure of 1.35 kbar, cell debris were removed by centrifugation at 30,000 × g for 1 hour, and the supernatants and pellets were separately kept and analysed.

For the proteins produced in insect and mammalian systems (Tables 10,11, 19, 22), cells were removed by centrifugation at 2.800 × g for 20 min and the supernatants were filtered to remove the remaining cell debris with 0.45 µm surfactant free cellulose acetate (SFCA) membrane filters (Thermo Scientific), followed by extensive dialysis with 3.5 kDa slide-A-lyzer™ Dialysis flask (Thermo Scientific) against the buffer used for the IMAC, the buffer A (50 Mm Tris-HCl, 250 mM sodium chloride, pH 7.5) (250 mL against 2X 5L; first overday and then, overnight). The filtered and dialyzed supernatants were used for subsequent purification steps.

2.4.1.1 Isolation of extracellular secreted protein

Supernatants containing the proteins of interest were incubated for 20 min (expression in insect cells) or 1 hour (expression in bacterial and mammalian cells) with nickel-nitrilotriacetic acid resin (Ni-NTA; Invitrogen) *plus* 20 mM imidazole, which was subsequently loaded onto an open column for batch purification (Bio-Rad), washed extensively with buffer A *plus* 20 mM imidazole, and eluted with buffer A *plus* 300 mM imidazole (direct Ni-NTA). A second incubation of the flow-through (FT) with the resin was usually performed to recover all the target protein (new resin for insect supernatants and the same resin for the mammalian expressions, because the insect media have a quelating component).

Instead of Ni-NTA, Talon^{R} resin (Takara) (using Co reactive core instead of Ni) was sometimes used for recovering hSKI1 proteins (Tables 10 and 11 from supernatants (capture less protein but also, less contaminants, thus it is more specific). In this case, the binding and washing of the resin were performed with buffer A *plus* 5 mM imidazole and eluted with buffer A *plus* 150 mM imidazole.

Another affinity resin used was the Streptactin®XT Superflow Suspension resin (IBA Life Sciences) for the purification of Strep-tagged GRAB and Strep-tagged pro-TGFβ2 (Table 22). The binding and washing steps were performed with 100 mM Tris·HCl (pH 8.0) and 150 mM sodium chloride buffer; and the final elution step with the same buffer *plus* 50 mM biotin.

In the case of the GRAB protein (expressed in bacteria in pCri8a vector with TEV cleavage; Table22), eluted samples from the IMAC were then dialyzed overnight (using the 3.5 kDa dialysis membranes (Biotech RC Dialysis Membranes, Fisher)) against buffer A *plus* 1 mM 1,4-dithio-DL-threitol (DTT) in the presence of His_{6x}-tagged TEV at a peptidase:protein weight ratio of 1:100 and 1 mM DTT. The resulting cleavage left additional residues (glycine-alanine-methionine) at the N-terminus of the target proteins due to the cloning strategy (see Table 22). Digested samples were passed several times through the Ni-NTA resin previously equilibrated with buffer A *plus* 20 mM imidazole to remove His_{6x}-tagged molecules and the FT containing untagged GRAB was collected (reverse Ni-NTA).

For the CR345 protein expressed in bacteria (pCri8a vector) (Table 22), eluted samples were dialyzed twice (overday and overnight) against buffer A but without DTT (CR has several disulphide bonds) and in presence of His6x-tagged TEV at a peptidase:protein weight ratio of 1:10 (because the efficiency of the cleavage is less without DTT). The resulting cleavage left additional residues (glycine-alanine-methionine-glycine) at the N-terminus of the target proteins due to the cloning strategy (see Table 22) Digested samples were passed several times through Ni-NTA resin previously equilibrated with buffer A *plus* 20 mM imidazole to remove His6-tagged molecules and the FT containing untagged CR345 was collected.

In all cases, proteins eluted from direct and reverse Ni-NTA chromatographies were dialyzed overnight against buffer B (20 mM Tris-HCl, 5 mM sodium chloride, pH 7.5) using the 3.5 kDa dialysis membranes, concentrated with Vivaspin centrifugal devices (Sartorius)and further purified by IEX on a TSKgel DEAE-2SW column (TOSOH Bioscience) equilibrated with buffer B. A gradient of 2–30% buffer C (20 mM Tris-HCl, 1 M sodium chloride, pH 7.5) was applied over 30 mL, and samples were collected and pooled. Subsequently, each pool was concentrated by ultrafiltration and subjected to SEC in Superdex 75 10/300 (GRAB, CR and pro-TGFβ2), Superdex 200 10/300 (BTMD hSKI1, N-hα2M and C-hα2M) or Superose 6 10/300 (full-length recombinant hα2M) columns (GE Healthcare Life Sciences) in buffer D (20 mM Tris-HCl, 150 mM sodium chloride, pH 7.5).

2.4.1.2 Protein isolation from periplasmic fraction of *E.coli* (by cold osmotic shock)

Cell pellet from hSKI1 construct expressed in pCri9a vector for periplasmic expression (Table 9) was resuspended in 25 mL (per L of culture) of sucrose buffer (20% (wt/vol) sucrose, 100 mM Tris-HCl, pH 8.0 and 1 mM EDTA). Next, the supernatant (periplasmic fraction) was removed by centrifugation at 10.000 × g for 10 min and saved. The pellet was resuspended in 25 mL (per L of culture) of 5 mM $MgCl_2$ and stirred for 20 min. After the incubation, the supernatant (osmotic shock fraction) was recovered by centrifugation at 10.000 × g for 10 min and also saved. This last pellet (rest of membranes) was saved for protein extraction from inclusion bodies. The periplasmic and osmotic shock fractions were pooled and filtered through 0.45 µm membrane filters and subsequently dialyzed overnight with the 3.5 kDa dialysis membranes against buffer A (Ni-NTA loading buffer). Dialyzed fractions were purified by affinity chromatography using Ni-NTA resin, with buffer A *plus* 20 mM imidazole for binding and washing; and with buffer A *plus* 300 mM imidazole for elution. Eluted proteins were analyzed by 10% SDS-PAGE.

2.4.1.3 Protein isolation from inclusion bodies

Two slightly different protocols were used in this thesis:

- **<u>Inclusion bodies purification protocol A:</u>**

Cell pellets (from the first centrifugation) and final membranes (from the periplasmic purification) containing the proteins of interest (hSKI1 constructs expressed in bacteria cells, Table 9) were resuspended in 20-30 mL of EDTA buffer (20 mM Tris-HCl, pH 7.5-8.0, 1 mM EDTA and 1% Triton X100) using a Potter-Elvehjem type Glass- Teflon homogenizer. The lysate was incubated for 10 mins. Then,

the insoluble fraction was collected by centrifugation at 25.000 × g for 15 min and saved. This pellet was gently resuspended in 20-30 mL of urea buffer (100 mM Tris- HCl, pH 7.5-8.0, 6M urea) with a Potter and incubated for 30 min at RT (urea crystallizes at low temperature). Subsequently, the insoluble fraction was collected by centrifugation at 25.000 × g for 10 min at 15°C. This second pellet was resuspended in 20-30 mL of guanidine hydrochloride buffer (100 mM Tris- HCl, pH 7.5-8.0, 6M GdmCl) using the Potter. Lysate was incubated for 20 min at RT and afterwards, the supernatant was recovered by centrifugation at 25.000 × g for 10 min. The three supernatant fractions and the last pellet were analyzed by 10% SDS-PAGE and the fractions with the protein of interest were purified by Ni-NTA with buffer 50 mM Tris- HCl, pH 7.5-8.0 *plus* 4M urea/ 3M GdmCl (for supernatants from urea and GdmCl fractions, respectively) and *plus* 20 mM imidazole for binding and washing, or 300 mM imidazole for elution. Then, eluted proteins were refolded.

- **Inclusion bodies purification protocol B:**

Cell pellets containing the proteins of interest (hSKI1 (188-477) and hSKI1 (210-477) proteins, Table 9) were resuspended in 50 mL of DTT buffer (50 mM Tris-HCl, pH 8.0, 500 mM sodium chloride and 1 mM DTT) and were broken with a cell disrupter at a pressure of 1.91 kbar. Then, supernatants were removed by centrifugation at 30,000 × g for 30 min. Remaining pellets were resuspended in 30 mL of DTT buffer *plus* 2% Triton X100, using a Potter homogenizer. Insoluble fraction was collected by centrifugation at 30,000 × g for 30 min and remaining pellets were again resuspended in 30 mL of DTT buffer *plus* 2% Triton X100, using a Potter. Then, supernatants were removed by centrifugation at 30,000 × g for 30 min. These resulting new pellets were resuspended in DTT buffer *plus* 8M urea and Complete EDTA-free Peptidase Inhibitor Cocktail Tablets, using the Potter, and incubated for 2h at RT with continuous mix. Afterwards, insoluble fractions were collected by centrifugation at 30,000 × g for 30 min and at RT. Pellets were resuspended with the Potter in DTT buffer *plus* 6M GdmCl. Lysates were incubated for 20 min at RT and then, the supernatants were recovered by centrifugation at 30.000 × g for 30 min. The three supernatants fractions and the last pellet were analyzed by SDS-PAGE and the fractions with the protein of interest were diluted to avoid quelating the resin (because the buffer contained DTT) and purified by Ni-NTA with buffer 50 mM Tris- HCl, pH 7.5-8.0 *plus* 4M urea/ 3M GdmCl (for supernatants from urea and GdmCl fractions, respectively) and 20 mM imidazole for binding and washing, or 300 mM imidazole for elution. Then, eluted proteins were refolded.

2.4.1.4 Protein refolding

After each protocol of protein extraction from inclusion bodies, a refolding protocol is always necessary because the proteins are completely denatured (because of the nature of the inclusion bodies and the strong chaotropic agents used during the extraction process). In this thesis, three protocols were used.

- **Refolding protocol A:**

Proteins eluted from the Ni-NTA (urea and GdmCl extractions from the inclusion bodies purification protocol A) were separately dialyzed overnight against refolding buffer A (50 mM Tris-HCl, pH 7.5-8.0, 4M urea/ 3M GdmCl, 5% Glycerol, 0.5M L-arginine) *plus* 0.5 mM reduced Glutathione (GSSG) and 3M oxidized Glutathione (GSH) (refolding buffer A.A) and refolding buffer A *plus* 0.5 mM GSH and 3M GSSG GSH (refolding buffer A.B) using the 3.5 kDa dialysis membranes. The precipitated-unfolded proteins were removed by centrifugation at 2.000 × g for 10 min. Soluble proteins were dialyzed overnight against refolding buffer B (50 mM Tris-HCl, pH 7.5-8.0, 5% Glycerol, 0.5M L-arginine) *plus* 0.5 mM GSSG and 3M GSH (refolding buffer B.A) and refolding buffer B *plus* 0.5 mM GSH and 3M GSSG (refolding buffer B.B). Then, the precipitated proteins were removed by centrifugation at 2.000 × g for 10 min and fractions (soluble-dialyzed and insoluble-precipitated) were analyzed by 10% SDS-PAGE.

- **Refolding protocol B (multi-step refolding protocol):**

This protocol is a continuation of protocol A, but with an additional dialysis step to remove the L-arginine from the buffer. Therefore, after dialysis in buffer B.A (with more soluble protein), the precipitated-unfolded protein was removed by centrifugation at 2.000 × g for 10 min and a last overnight dialysis was performed with the supernatant in buffer C (50 mM Tris-HCl, pH 7.5-8.0, 5% Glycerol and 4 mM β-Mercaptoethanol). Then, the precipitated protein was removed by centrifugation at 2.000 × g for 10 min and fractions (soluble-dialyzed and insoluble-precipitated) were analyzed by 10% SDS-PAGE.

- **Refolding protocol C:**

Proteins eluted from the Ni-NTA (urea and GdmCl extractions from the inclusion bodies purification protocol B) were diluted up to 0.3 mg/mL and afterwards, they were passed through a Sephadex G25 Pd10 desalting column (GE Healthcare Life Sciences) previously equilibrated with the following buffer:

50 mM Tris-HCl, pH 8.0, 500 mM sodium chloride, pH 8.0. Eluted proteins were analyzed by 10% SDS-PAGE.

2.4.2 Wild-type human α2-macroglobulin purification from blood

All protein purification steps were performed at 4°C if not otherwise stated.

Wild-type full-length hα2M was isolated from blood plasma from individual donors and purified essentially as described previously (Goulas et al., 2014b; Imber and Pizzo, 1981; Sottrup-Jensen et al., 1980). Briefly, plasma was subjected to sequential precipitation steps with 4–12% PEG 4,000, and the final precipitate containing hα2M was reconstituted in 20 mM sodium phosphate at pH 6.4. Partially purified hα2M was captured with a zinc-chelating resin (G-Biosciences), washed with buffer E (50 mM sodium phosphate, 250 mM sodium chloride, pH 7.2) *plus* 10 mM imidazole and eluted in the same buffer *plus* 250 mM imidazole and 100 mM EDTA. The protein was first passed through a PD10 desalting column previously equilibrated with 20 mM HEPES, pH 7.5 and then subjected to an IEX step in a Q Sepharose column (2.5 × 10 cm; GE Healthcare Life Sciences), previously equilibrated with 15% buffer F (20 mM HEPES, 1 M sodium chloride, pH 7.5). A gradient of 20–30% buffer F was applied for 150 min and fractions were collected. Collected samples were dialyzed overnight against buffer G (20 mM sodium phosphate, 5 mM sodium chloride, pH 7.4) and further purified by IEX in a TSKgel DEAE-2SW column, previously equilibrated with buffer G. A gradient of 7–20% buffer H (20 mM sodium phosphate, 1 M sodium chloride, pH 7.4) was applied over 30 mL, and samples were collected and pooled. Subsequently, each pool was concentrated and subjected to a final polishing step by SEC in a Superose 6 10/300 column in buffer I (20 mM sodium phosphate, 150 mM sodium chloride, pH 7.4).

2.5 Determination of protein identity, purity and concentration

Protein identity and purity were assessed by 10% sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) (Haider et al., 2011) stained with Coomassie Brilliant Blue or with silver staining (Pierce™ Silver Stain Kit from Thermo Scientific), by Western blot, peptide mass fingerprinting of tryptic protein digests, N-terminal sequencing through Edman degradation and mass spectrometry. The latter three were carried out at the Protein Chemistry Service and the Proteomics Facilities of the Centro de Investigaciones Biológicas (CIB; Madrid, Spain). Ultrafiltration steps were performed with Vivaspin 15 and Vivaspin 500 filter devices of 10 to 50 kDa cut-off (Sartorius Stedim Biotech).

Protein concentrations were estimated by measuring the absorbance at 280 nm in a spectrophotometer (NanoDrop) and applying the respective theoretical extinction coefficients.

Concentrations were also measured by the BCA Protein Assay Kit (Thermo Scientific) with bovine serum albumin fraction V (BSA; Sigma-Aldrich) as a standard.

Induced hα2M was differentiated from native hα2M by native-PAGE, because of their different mobilities.

2.5.1 Native polyacrylamide gel electrophoresis

The native gels were made according to the standard Ornstein-Davis buffer formulation (basic gels for proteins with a pI smaller than 8.3) (Davis, 1964; Ornstein, 1964) (Table 6, recipe for one gel).The running buffer for 1L was 3 g Tris; 14.4 g Glycine, pH 8.3. The formula for the 5X sample buffer was: 1M Tris pH 6.8; 1% Bromofenol Blue and 50% Glycerol. The gels were run at 150V and, in the case of the mobility studies of α2M, for at least 4 hours.

Component	10% Resolving gel buffer	10% Stacking gel buffer
Acrylamide/Bis-acrylamide (37:1 ratio)	1 mL	0.67 mL
1.5 M Buffer pH 9.0	1.5 mL	
0.5 M Buffer pH 6.8		1 mL
Milli Q water	3.5 mL	2.3 mL
10% PSA	50 µL	30 µL
TEMED	7 µL	7 µL

Table 6. Formula for one 10% native-PAGE.

2.5.2 Sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE

The recipe for three-four gels of 10% SDS-PAGE gels (with 0.75 mm) is shown in detail in Table 7. gel buffer contains (for 1L): 181,5 g Tris Base, pH 8.45 and 1.5 g SDS. Running buffer is using anode and cathode. For 10X anode (for 1L): 242 g Tris Base, pH 8.9 and 144 g Glycine. For 10X cathode (for 1L): 121 g Tris Base, pH 8.45, 179 g Tricine and 5 g SDS. These gels were run at 25-30mA.

Component	10% Resolving gel buffer	10% Stacking gel buffer
Acrylamide/Bis-acrylamide (37:1 ratio)	6.66 mL	1.33 mL
Gel buffer	6.66 mL	2.6 mL
50% Glycerol	4 mL	
Milli Q water	2.68 mL	6.07 mL
10% PSA	66 µL	80 µL
TEMED	6.6 µL	8 µL

Table 7. Formula for three-four10% SDS-PAGE.

2.5.3 Western blot analysis

Protein samples were separated by 10% SDS-PAGE, transferred to Hybond ECL nitrocellulose membranes (GE Healthcare Life Sciences), and blocked for two hours under gentle stirring at room temperature with 50 mL of blocking solution (phosphate buffered saline; PBS) plus 0.1% Tween 20 and 5% BSA. His6-tagged proteins were detected by immunoblot analysis using the monoclonal His-HRP Conjugated Antibody (Santa Cruz Biotechnology) diluted 1:5,000 in PBS plus 0.1% Tween 20. Strep-tagged proteins were detected with the Streptavidin-Peroxidase Conjugated Antibody from Streptomyces avidinii (Sigma-Aldrich) diluted 1:1,000 in PBS plus 0.1% Tween 20 and 1% BSA. Complexes were detected using an enhanced chemiluminescence system (Super Signal West Pico Chemiluminescent; Pierce) according to the manufacturer's instructions. Membranes were exposed to Hyperfilm ECL films (GE Healthcare Life Sciences).

2.6 Biochemical assays

2.6.1 Detection of free sulfhydryl groups

Detection of free sulfhydryl groups was performed with the Fluorometric Thiol Assay Kit (ab112158 assay; Abcam) following the manufacturer's instructions and using glutathione as a standard for the dose response curve. The fluorescent signal was measured in a microplate fluorimeter (Infinite M200, TECAN) at λex = 490 nm and λem = 520 nm in 96-well plates containing 100 µL reaction volumes (50 µL of assay reaction mixture *plus* 50 µL of glutathione-standard or test samples) in duplicate. Fluorescence was measured after preincubation of wild-type hα2M (0.39 µM) or C-hα2M obtained

from human cells (1.6 µM), with or without treatment with methylamine for 10, 20, 30, 45 and 60 min, at room temperature.

2.6.2 Peptide mass fingerprint

The samples sent to the Protein Chemistry Service and the Proteomics Facilities of the Centro de Investigaciones Biológicas (CIB) (Madrid, Spain) for peptide mass fingerprinting analysis were first run in a 10% SDS-PAGE and separate it as much as possible from other bands into the sample, and also run without other proteins in parallel to avoid contamination. Then, the gel was stained for 5 min and destained for 5-10 min with fresh Coomassie Blue staining (0.1% solution in 40% methanol/10% acetic acid) and destaining solution (50% methanol), respectively. Each sample was excised with a clean razor blade and placed in 1.5 ml Eppendorf with 50 µL buffer to keep it wet for shipment.

2.6.3 N-terminal Edman degradation

Each sample sent to the Protein Chemistry Service and the Proteomics Facilities of the Centro de Investigaciones Biológicas (CIB) (Madrid, Spain) for their N-terminal Edman degradation analysis, was previously run in a 10% SDS-PAGE close to the protein ladder to control its running. These samples were salts and detergents free and very pure. The gels used were polymerized overnight before running the sample to eliminate any N-terminal blockage due to unpolymerized acrylamide in the gels. Also, all the solutions were new to avoid contamination. After the blotting in a PVDF membrane, it was stained with Coomassie Brilliant Blue staining for 5 minutes followed by destaining (5-10 minutes) in a methanol solution and rinsed thoroughly with Milli-Q water. After drying at RT, the bands of interest were excised with a clean razor blade, taking care not to cut off empty membrane without protein, it was then finally placed in 1.5 ml Eppendorf for shipment.

2.6.4 Methylamine induction of human α2-macroglobulin

Induced hα2M was obtained by treating 1 mg/mL native hα2M in buffer D (20 mM Tris-HCl, 150 mM sodium chloride, pH 7.5) with 200 mM methylamine hydrochloride for one hour at room temperature. Subsequently, the sample was dialyzed against the same buffer D and analyzed by native-PAGE.

2.7 Proteolytic assays

2.7.1 Tested substrates

Proteolytic activity against fluorogenic peptide substrates of sequence: Abz-Lys-Asp-Glu-Ser-Tyr-Arg-K(dnp) (FRET1), Abz-Thr-Val-Leu-Glu-Arg-Ser-K(dnp) (FRET2), Abz-Asp-Tyr-Val-Ala-Ser-Glu-K(dnp) (FRET3), Abz-Tyr-Gly-Lys-Arg-Val-Phe-K(dnp) (FRET4) and Abz-Val-Lys-Phe-Tyr-Asp-Ile-K(dnp) (FRET5) (Goulas et al., 2011) (where Abz is ortho-aminobenzoic acid and K(Dnp) is Ne-2,4-dinitrophenyl-lysine with free carboxyl group) (Cotrin et al., 2004). These reactions were performed per triplicate at room temperature and monitored for up to 24 h in a microplate fluorimeter (Infinite M200, TECAN) (λex= 260 nm and λem= 420 nm). FRETS were used at 10 µg/mL in 200 µL-reaction volumes with 0.2 µg/µL of hSKI1 proteins (from 18-947 and from 18-998, expressed in Expi293FTM cells).

2.7.2 Zymography studies

To this aim, 10% SDS-PAGE gel was prepared containing 0.2% (w/v) gelatin. Protein sample at 0.1 µg/µL (protein came from a crystallization drop) was subjected to electrophoresis at 4°C with equal volume of SDS-PAGE sample buffer without β-mercaptoethanol. Gels were washed twice in 20 mM Tris-HCl pH 7.4, 150 mM sodium chloride, 10 mM calcium chloride, and 2.5% (v/v) Triton X-100 for 15 min and then incubated in the same buffer without detergent for 16 h at 37 °C under gentle shaking. Gels were stained with Coomassie Brilliant Blue.

2.7.3 Proteolytic activity inhibition assays

Inhibition assays against synthetic protein substrates were performed in a microplate fluorimeter (Infinite M200, TECAN) in 200 µL reaction volumes with the fluorescence-based EnzCheck Assay Kit containing BODIPY FL-casein (*λ*ex = 505 nm and *λ*em = 513 nm) as fluorescein conjugate (Invitrogen) at 10 µg/mL in buffer D. Inhibition was measured after preincubation of a two-fold molar excess of wild-type (1 µg/µL) or four-fold molar excess of recombinant hα2M (0.3 µg/µL) with trypsin (0.015 µg/µL) for 15 min at RT. The substrate was added to the Reaction mixture and the residual tryptic activity was measured over a period of two hours.

Inhibition assays against natural protein substrates were performed by analysing the following reaction by 10% SDS-PAGE. 10 µL reaction containing wild-type hα2M (5 µg) and trypsin in the same molar ratio were incubated for 5 min at RT and then, 1 mg/mL of casein was added to the reaction.

The mixture was incubated for 30 min at RT and afterwards, inhibited by adding Pefabloc (10 µg) (Pefabloc^{R} SC PLUS, from Sigma-Aldrich).

2.8 Protein binding studies *in vitro*

2.8.1 Size exclusion chromatography

Protein interactions between GRAB or pro-TGFβ2 or CR with native wild-type hα2M, induced wild-type hα2M and recombinant hα2M proteins were preliminary studied by SEC, where 250 µL samples of the complexes at a molar ration of 4:1 were injected in a Superose 6 10/300 column (for complex with wild-type hα2M) or Superdex 200 10/300 column (for complex with recombinant hα2M proteins) after their incubation for 15 min at RT or overnight at 4°C or 1 hour at 37°C. Control proteins were also injected separately. Experiments with recombinant hα2M were performed in buffer D (20 mM Tris-HCl, 150 mM sodium chloride, pH 7.5) and experiments with wild-type hα2M, in buffer I (20 mM sodium phosphate, 150 mM sodium chloride, pH 7.4). Peaks were analyzed by 10% SDS-PAGE, and then, the samples from the peak with the "complex" were concentrated by Vivaspin filter devices and reinjected to verify the interaction of the complex.

2.8.1.1 Multiangle laser light scattering coupled to size exclusion chromatography

Multi-angle laser light scattering (MALLS) coupled to size exclusion chromatography allows the determination of absolute molar mass and size of selected sample fractions eluting from the column, being a very useful technique to identify peak content and monodipersity, as well as helping to determine protein complex subunit stoichiometries in solution

We performed MALLS in a Dawn Helios II apparatus (Wyatt Technologies) coupled to a SEC Superose 6 10/300 column (SEC-MALLS) equilibrated in buffer I at 25 °C at the joint IBMB/IRB Crystallography Platform, Barcelona Science Park (Catalonia, Spain) to analyze binding of GRAB or pro-TGF-β2 to native or induced wild-type hα2M at a molar ratio of 4:1. ASTRA 7 software (Wyatt Technologies) was used for data processing and analysis, for which a dn/dc value (refractive index increment) typical for proteins(0.185 mL/g) was assumed. All experiments were performed in triplicate.

2.8.2 Pull-down assays

The interactions between native wild-type hα2M /MA-induced wild-type hα2M/ C-hα2M and GRAB were also detected by pull-down with Strep-Tactin® and Ni-NTA. Briefly, GRAB-Strep was first bound to the Strep-Tactin® resin and then, native wild-type hα2M (with no-tags) was passed through. After being extensively washed, native hα2M was still bound to GRAB. Same experiment was performed for MA-induced hα2M. The interaction between C-hα2M and GRAB was slightly different. C-hα2M-His was first bound to the Ni-NTA and then, cleaved GRAB (without His-tag) was passed through. After being extensively washed, GRAB was still bound to C-hα2M. All interactions were analyzed by 10% SDS-PAGE and Coomassie Blue or Silver stainings.

2.8.3 Surface plasmon resonance and kinetic data analysis

The binding kinetics (association and dissociation) and affinity (complex formation at the equilibrium) of GRAB or pro-TGFβ2 (ligands) with native wild-type hα2M, induced wild-type hα2M, recombinant N-hα2M or recombinant C-hα2M (analytes) were studied by surface plasmon resonance with a Biacore™ T200 Biosensor System (GE Healthcare Life Sciences) at the Scientific and Technological Centers of the University of Barcelona (Catalonia, Spain). To bind ligands provided with a Strep-tag, Streptactin®XT (IBA LifeSciences) was immobilized at 25°C on the surface of the four flow cells of a sensor chip CM5 series S (GE Healthcare Life Sciences) at 3.000 response units (RU) through amine coupling, as described previously (Sinha-Datta et al., 2015). Subsequently, Strep-tagged GRAB (at 9.7 nM) or pro-TGFβ2 (at 19.0 nM) in HBNS buffer (10 mM HEPES, 150 mM sodium chloride, pH 7.4) were immobilized at low RU density on different flow cells of the chip by virtue of the strong interaction between the Strep-tag and streptactin at 5 µL/min for 24 sec at 37 °C. To monitor association, the immobilized ligands were then exposed to the analytes at different concentrations in HBNS (4–600 nM for native and induced wild-type hα2M; 75–1,200 nM for N-hα2M and C-hα2M), which were injected at 30 µL/min for 120–240 sec at 37 °C. Thereafter, HBNS was injected for analyte dissociation from the immobilized ligands for 90–300 sec. To dissociate bound ligands and regenerate the chip surface, 3 M guanidine hydrochloride was injected at 30 µL/min for 30 sec after each cycle. These experiments were double referenced by keeping the first flow cell without ligand, and by an injection step at analyte concentration zero. The affinity analysis was performed by plotting binding responses in the steady-state region of the sensorgrams (Req) against analyte concentrations to determine the overall equilibrium dissociation constant (KD). Sensorgrams were analyzed with the BIAEVALUATION program v. 3.0 (GE Healthcare Life Sciences) and fitted to a 1:1 Langmuir interaction

model. The likelihood of fitting was assessed through the χ2 statistical parameter (Sinha-Datta et al., 2015)..

In a separate qualitative experiment, ligands GRAB (at 120 nM) and pro-TGFβ2 (at 950 nM) were premixed with the analytes at different concentrations (2–150 nM for native and induced wild-type hα2M; 38–600 nM for N-hα2M and C-hα2M) and incubated for one hour at 37°C. Subsequently, the mixtures were injected at 15 µL/min at 37°C according to a published multicycle method (Zhao et al., 2017). The binding was measured through the increase in RU after injection of the premixes and the stability of the resultant complexes through their elution with buffer HBNS at a flow rate of 30 µL/min. Ligand solutions without analyte were used as negative controls of complex formation and the sensor surface was regenerated after each sample injection.

2.8.4 Protein labelling (Sulfo-NHS-AMCA)

GRAB and pro-TGFβ2 were labelled with fluorogenic sulfosuccinimidyl-7-amino-4-methylcoumarin-3-acetate (Sulfo-NHS-AMCA; Thermo Scientific) according to the manufacturer's instructions with a 10–15 molar excess of reagent over protein in buffer I (20 mM sodium phosphate, 150 mM sodium chloride, pH 7.4).for 1 hour at room temperature. Thereafter, the proteins were extensively dialyzed against buffer I to remove non-reacted dye. To assess binding, labelled GRAB or pro-TGFβ2 were mixed with wild-type hα2M (native and induced) or recombinant fragments N-hα2M and C-hα2M at a 4:1 molar ratio, incubated in buffer I for two hours at 37 °C, and then analyzed by 10% native PAGE (Haider et al., 2011). Gel fluorescence was visualized in a gel reader (G:BOX F3 Gel Doc System, Syngene) and the fluorescence was measured (λex = 345–350 nm and λem = 440–460 nm). Negative controls (unlabelled proteins) were included in each experiment. After fluorescence detection, native gels were stained with Coomassie Brilliant Blue (Thermo Scientific) to detect the negative controls.

2.8.5 BS^3 (bis(sulfosuccinimidyl)suberate) crosslinking

The protein interactions between GRAB or mature TGFβ2 with native wild-type hα2M were detected by crosslinking with BS^3 (Thermo Scientific) according to the manufacturer's instructions. Briefly, GRAB-Strep and pro-TGFβ2-Strep were preincubated with native wild-type hα2M for 30 min at RT (twenty five-molar excess of ligand) in 50 µL of 100 mM sodium phosphate, 150 mM sodium chloride, pH 7.4 buffer. Then, the BS^3 was added from 0.125-2 mM (final concentration) and the reaction was incubated for 30 min at RT. After this incubation, the reaction was stopped with 1 M Tris

for 15 min at RT. And finally, the results were analyzed by 10% SDS-PAGE and Western-blot against the Strep-tag.

2.9 Structural characterization

2.9.1 Crystallization trials

Protein crystallization assays and successful conditions are explained in section 3.4.2.10. All pure protein samples were concentrated with Vivaspin centrifugal devices and then, filtered with 0.22 µL centrigugal filters (Ultrafree^{R} -MC-GV from Millipore) before proceeding with the crystallization experiments.

Crystallization trials were first performed by screening of the conditions at different temperatures, protein concentrations, protein-precipitant ratios, buffers and additives at the IBMB/IRB Automated Crystallography Platform. This platform contains three robots: a TECAN Freedom EVO to prepare reservoir solutions and a Cartesian Microsys 4000 XL (Genomic Solutions) and a Phoenix/RE (Art Robbins) for 100 nL-nanodrop dispensing. Then, the scale up of the best conditions to bigger volumes and condition range was performed by the sitting-drop vapor-diffusion method (Adachi et al., 2003).

This method consists of setting up protein drops of 0.5, 1 or 2 µL *plus* the precipitant (in the same microliter range; 1+1; 1+2), which is the same solution of the 500 µL reservoir, in a 24-well Cryschem crystallization dishes (Hamptom Research) (Fig. 23).

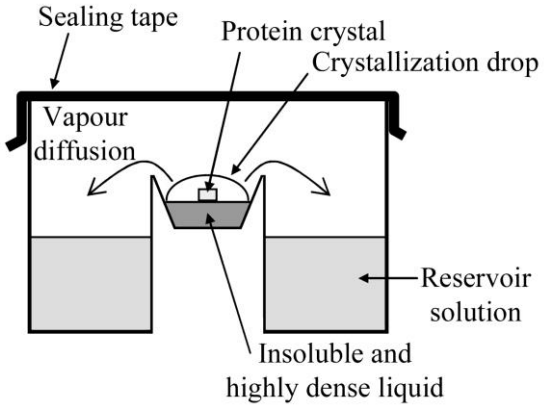


Figure 23. Scheme representation of the sitting-drop vapor-diffusion method (Figure reproduced from Adachi et al., 2003).

2.9.2 Cryoprotection, crystallization of mature TGFβ2 and diffraction data collection.

Cryoprotection conditions were are explained in section 3.4.2.10 for each case.

Octahistidine-tagged pro-TGFβ2, mostly cleaved in the linker between LAP and mature domains of pro-TGFβ2, at ~5 mg/mL in 10 mM Tris·HCl pH 8.0, 150 sodium chloride was incubated O/N at 4°C with human induced α2-macroglobulin (hα2M) at ~7 mg/mL in 10 mM Tris·HCl pH 8.0, 150 mM sodium chloride. The hα2M was previously purified from blood and reacted with methylamine as reported before. The hα2M/pro-TGFβ2 reaction mixture was subjected to crystallization assays by the sitting-drop vapor diffusion method at the Automated Crystallography Platform of Barcelona Science Park. Reservoir solutions were mixed by a Tecan robot and crystallization drops of 100 nL were dispensed by a Cartesian Microsys 4000 XL (Genomic Solutions) robot or a Phoenix nanodrop robot (Art Robbins) on 96 × 2-well MRC nanoplates (Innovadyne). Crystallization plates were stored at 20°C or 4°C in Bruker steady-temperature crystal farms, and successful conditions were scaled up to the microliter range in 24-well Cryschem crystallization dishes (Hampton Research).

Diffraction data processing statistics are provided in Table 26.

2.9.3 Structure solution and refinement

The structure of mature TGFβ2 was solved by likelihood-scoring molecular replacement with the PHASER (McCoy A. J et al., 2007) program and the coordinates of a GF protomer crystallized in a different space group and unit cell (PDB 2TGI; Daopin et al., 1992). Subsequently, an automatic tracing step was performed with ARP/ wARP (Langer et al., 2008), which outputted a model that was completed through successive rounds of manual model building with the COOT program (Emsley et al., 2010) and crystallographic refinement with the PHENIX (Afonine et al., 2012) and BUSTER/TNT (Smart et al., 2012) programs. The latter included TLS refinement.

Structure figures were made with the CHIMERA program (Pettersen et al., 2004). Structures were superposed with the SSM program (Krissinel and Henrick, 2004) within COOT. A search for structural similarity against the PDB was performed with DALI (Holm and Laakso, 2016). The final model of human TGFβ2 GF was validated with the wwPDB Validation Server (https://wwwwwpdb.org/validation) (Berman et al., 2003) and is available at the PDB at https://www.rcsb.org (access code 6I9J).

2.10 Phylogenetic studies

The literature available on matrix metalloproteinases (MMPs) outside vertebrates was extensively reviewed through PubMed at https://www.ncbi.nlm.nih.gov/pubmed. The search for potential

members in invertebrates, plants, fungi, viruses, protists, archaea and bacteria was performed with the amino acid sequence of structurally validated MMP catalytic domains within UniProt (www.uniprot.org) or the National Center for Biotechnology Information (blast.ncbi.nlm.nih.gov/Blast.cgi) using standard parameters. The taxonomy convention follows the Catalogue of Life (http://www.catalogueoflife.org/col; (Roskov et al., 2017)).

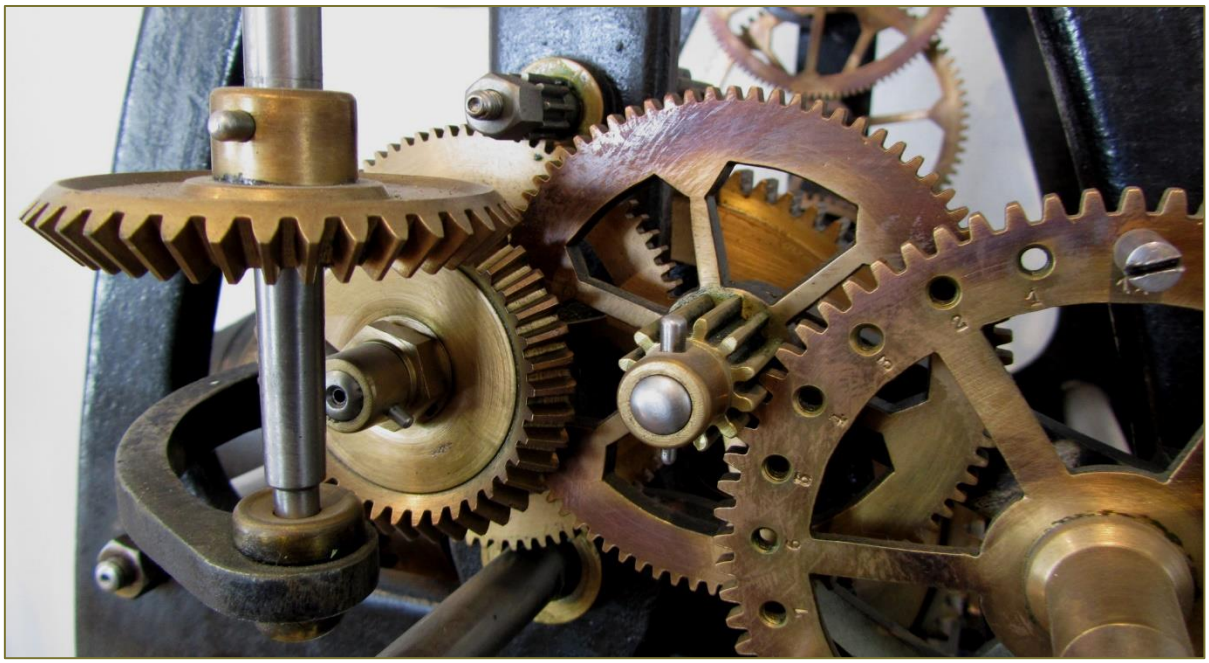
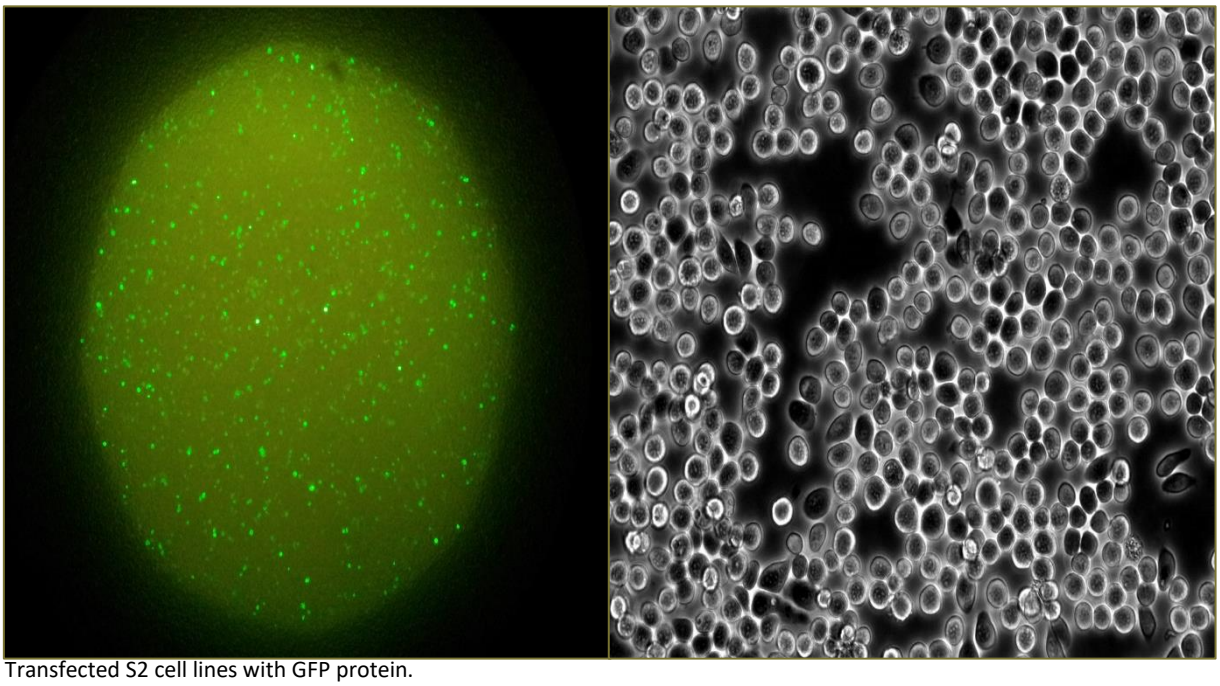


Figure reproduced from “photocmmunity.de” (Michael Ruppel)

RESULTS



Transfected S2 cell lines with GFP protein.

RESULTS

3.1 Establishment of eukaryotic expression systems for recombinant protein production

3.1.1 OBJECTIVES

The goal of this part of the thesis was to establish one or several efficient expression systems for human recombinant proteins production at our laboratory. Our laboratory had some expertise on protein expression, purification and crystallization of bacterial expression systems; but the background in eukaryotic expression systems was more limited. Thus, the expression systems should be established step by step to understand the importance of every element. The goal was to develop an easy, affordable, reproducible and high-performing eukaryotic protein expression system for the laboratory.

3.1.2 RESULTS AND DISCUSSION

3.1.2.1 Establishment of cell culture and transfection conditions for protein expression in the *Drosophila* Schneider 2(S2) cell system

3.1.2.1.1 Adaptation of S2 cells to suspension

S2 cells are originally a semi-adherent monolayer cell culture that grow in plates with 10% fetal bovine serum (FBS) added into insect medium at 28°C. Growth in plates is always a good choice for the initial steps of protein detection and only needs a temperature-controlled incubator with regulated CO_2 supply. On the other hand, protein yields using this method are low because adherent cell cultures yield lower densities when compared to suspension cultures. We were interested in establishing a new eukaryotic expression system at the laboratory in order to express large amounts of well folded human recombinant proteins, not only for biochemical assays, but also for crystallization. The latter studies having higher demand of pure protein. For this reason, the first point of the implementation of the chosen insect expression system was the adaptation of the cells to suspension in order to increase cell densities. Like that, we could potentially get higher expressed protein yields. This meant to optimize:

- Cell culture conditions (temperature and speed shaking) considering the orbital shaker incubator (Brunswick Scientific Innova).
- Cell culture volume depending on the tube shape (in our case: tubeSpin bioreactor tubes, TS).

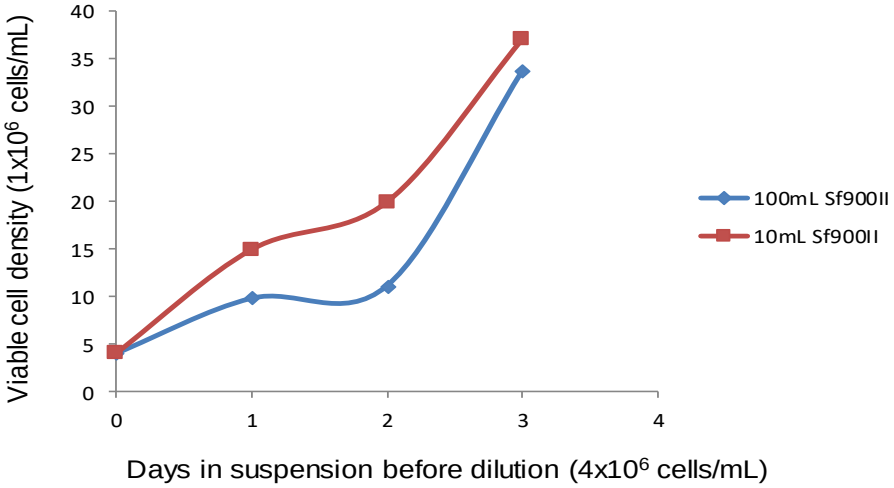


Figure 24. Cell growth comparison depending on the cell culture volume (S2 cells) at 220 rpm and 28°C. X axis represents the days that cells are growing before being split up to 4 × 10^6 cells/mL. Y axis represents the viable cell density.

S2 cells which had been adapted to suspension grew better and in higher densities using TS50 for 5 to 10 mL cultures and TS600 for 100-200 mL cultures under an agitation of 220 rpm and temperature at 28°C. Small scale cultures seemed to grow better than large ones (Fig. 24), but the handling,and thus, the risk of contamination, are also higher. From this point on, we decided to use only 10 mL cultures for maintenance and 200 mL cultures for protein production. We also decided to respect the volume limits at 220 rpm to avoid cell precipitation (because of low speed and/or bigger culture volumes) and foam (high speed), that caused cell death.

3.1.2.1.2 Determination of the appropriate growth conditions and protein expression medium for S2 cells

The usual media for the maintenance of insect cell cultures, including Sf9 cells and other similar cell lines, are IPL-41 and TC-100 (both from Gibco). These media are very economical (55€ / 500mL) and the recipes are available. However, this kind of media need to be supplemented with 10% FBS .

The serum-free insect cell culture media is another alternative. SF900™ II SFM (Gibco) (79,50€ / 500mL) are designed for large-scale production of recombinant proteins and suitable for suspension of monolayer culture. This medium is ready to use, not requiring the addition of any serum, glutamine or surfactants. The formulation is, however, not available.

Taking all that into account, we decided to adapt S2 cells to IPL-41 and TC-100 media to compare cell growth and expression protein yields.

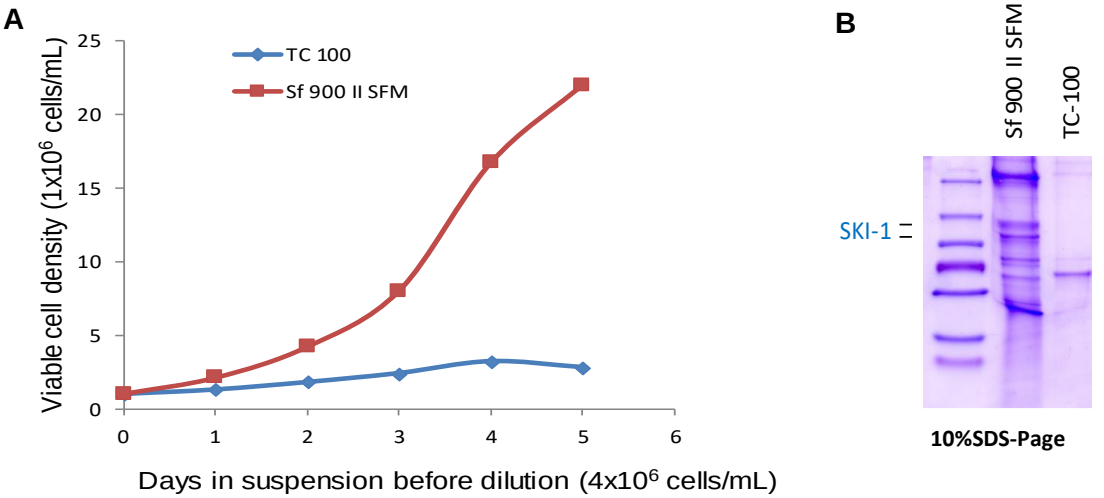


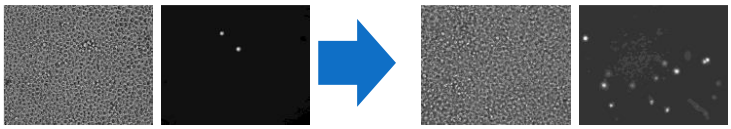
Figure 25. Cell growth and expression protein yield comparison depending on the insect cell culture medium (S2 cells) at 220 rpm and 28°C. (A) X axis represents the days that cells are growing before being split up to 4 x 10^6 cells/mL. Y axis represents the viable cell density**. (B)** Pre-adapted S2 cells for 4 weeks to TC100 and Sf900II media were transfected with pMT/GNBP3-hSKI1 (18-998)-H8 and harvested after 5 days. Protein analysis was performed by SDS-PAGE and stained by Coomassie Brilliant Blue.

Among these media, the highest cell density (~22 x 10^6 cells/mL) was reached with Sf900II medium (Fig. 25A). IPL-41 medium could not even support high density growth of S2 cells in suspension. The TC-100 medium could support it, but the cell growth was low and no protein expression (Fig. 25B). In conclusion, among the media tested, the only one suitable for high expression performance was the SF900II medium and consequently, it was selected for further experiments to optimize the transfection of S2 cells.

3.1.2.1.3 Optimization of calcium phosphate transfection of S2 cells

We optimized and assayed the calcium phosphate transfection protocol in adherent cultures (9.6 cm^2 plates, 6-well plates) and the efficiency of transfection was determined by measuring the percentage of cells emitting green fluorescence light (λex = 395 and 475nm; λem = 508 and 503 nm, respectively) at day 3 post-transfection, due to the presence of successful transfected GFP (transfection with the pMT/GFP-H8 construct). Variables assayed were:

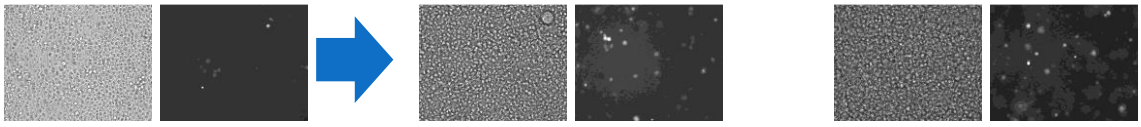
- Premix of DNA with calcium chloride or HBS* solution (* standard HBS solution *plus* glucose):



5%FBS
4ug DNA
CaCl2: DNA premix

5%FBS
4ug DNA
PO4: DNA premix

- FBS concentration (0,2-10%):

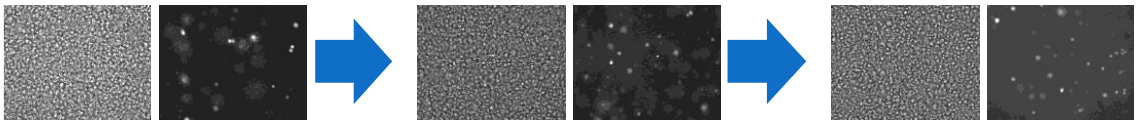


0,2%FBS
8ug DNA
PO4: DNA premix

5%FBS
8ug DNA
PO4: DNA premix

10%FBS
8ug DNA
PO4: DNA premix

- DNA amount (4-16 µg):



10%FBS
4ug DNA
PO4: DNA premix

10%FBS
8ug DNA
PO4: DNA premix

10%FBS
16ug DNA
PO4: DNA premix

Figure 26. Step by step calcium phosphate transfection of S2 cells transfected with pMT/GFP-H8 construct. Steps: reagent order (DNA with calcium chloride or HBS* solution), FBS concentration (0.2-10%) and DNA amount (4-16 µg). Same cellular area observed with light phase contrast (left) and epifluorescence (GFP excitation, right). Squares in blue indicate the best condition for each trial.

The best conditions for using calcium phosphate transfection were: a first premix of DNA with *HBS and then $CaCl_2$; DNA amount from 8 to 16 µg and FBS concentration of 10% (v/v) (Fig. 26).

3.1.2.1.4 Optimization of lipofectamine transfection of S2 cells

We also assayed the transfection with Lipofectamine™ 2000 (Invitrogen) in S2 cells (9.6 cm^2 plates, 6-well plates) and the efficiency of transfection was determined similarly, by measuring the percentage of positive GFP cells at day 3 post-transfection.

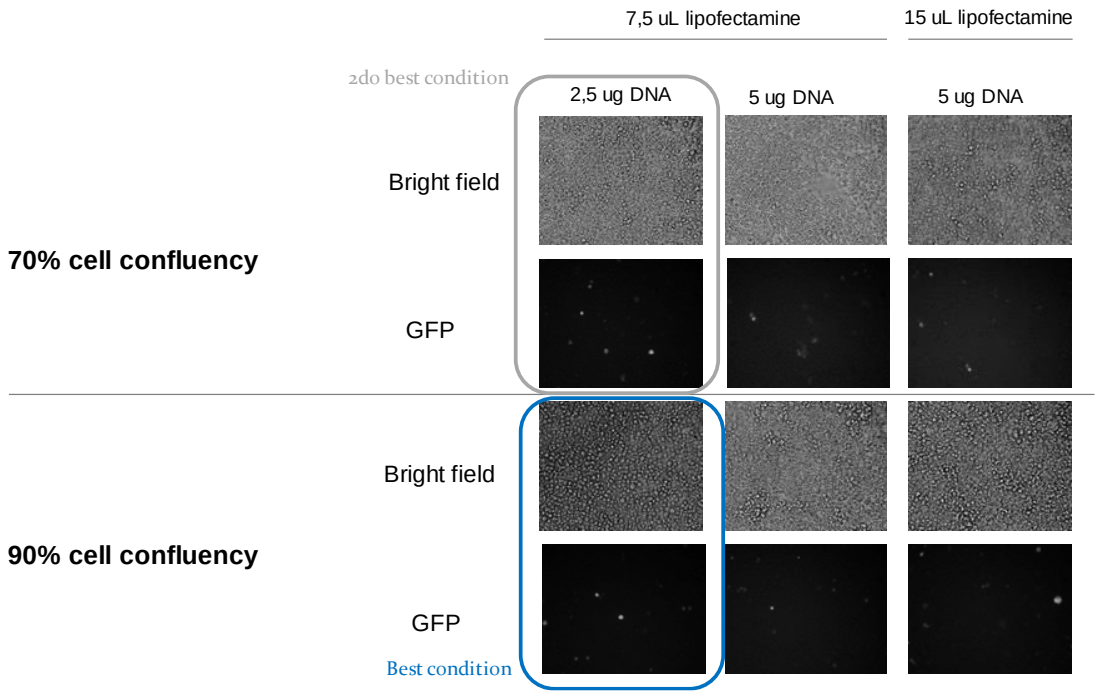


Figure 27 Optimization of lipofectamine transfection of S2 cells transfected with pMT/GFP-H8 construct . Steps: cell confluency, lipofectamine and plasmid DNA amounts**.** Same cellular area without (upper) and with GFP excitation (lower). Squares indicate the best conditions (7.5 µL lipofectamine and 2.5 µg DNA for 90 and 70% cell confluency).

Slightly higher transfection efficiencies were obtained at 90% cell confluency, using 7.5 µL of lipofectamine and 2.5 µg of plasmid DNA in Sf900 medium supplemented with 10% FBS (pMT/GFP-H8) (Fig. 27). In any case, transfection efficiency was significantly lower than calcium phosphate transfection and also, the Lipofectamine had a higher cost, so we discarded this method.

3.1.2.1.5 Optimization of polyethylenimine (PEI) transfection of S2 cells

Although we observed efficient DNA delivery into S2 cells by calcium phosphate transfection, we also tested the cationic polymer PEI, following the protocol previously described (Shen, 2014) but setting different parameters for polyplex formation and cell harvesting. First parameters assayed: optimal cell density (8, 12 and 15 x 10^6 cells/mL) before adding the polyplex (DNA + PEI) and cell culture influence (addition the polyplex into fresh or conditioned medium).

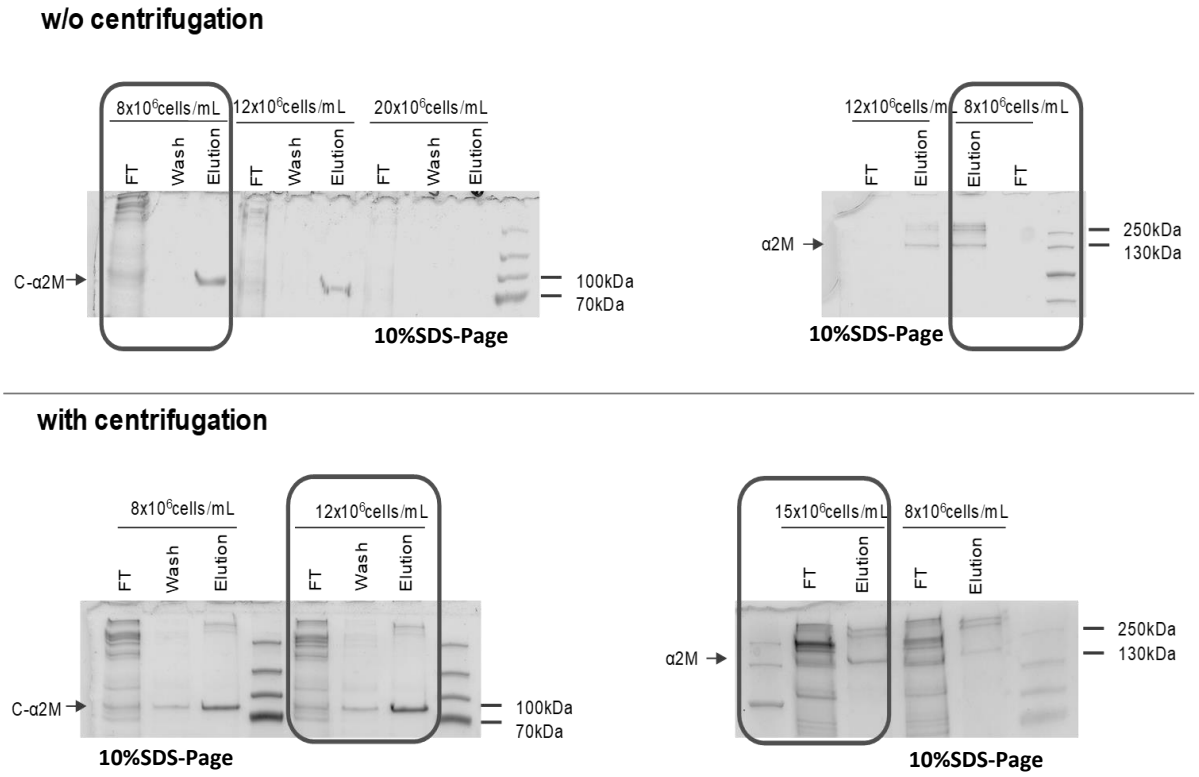


Figure 28. Optimal cell density evaluation for PEI transfection (S2 cells). Two different protocols were assayed (with and without centrifugation of the cells) using two different proteins (C-terminal and full-length α2M) and different cellular densities (8, 12 and 15 x 10^6 cells/mL) before the polyplex addition and harvesting after 5 days post-transfection. Results were analysed by SDS-PAGE and stained by Coomassie Brilliant Blue.

In order to improve the PEI transfection, we started with a standard protocol and we followed a gradual modification approach modifications (with or without centrifugation of the cell culture before adding the DNA-PEI complex) to verify the importance of clean medium to avoid the kidnap of polyplex and set up the optimal cell density (Fig. 28). We observed transfection for both tested proteins with the two different approaches. Although the transfection efficiency at lower cell concentration (8 x 10^6 cells/mL) was similar in both cases, at higher cell concentrations a significant improvement was observed when media was replaced immediately before transfection, confirming the importance of fresh media for PEI transfection at high cell densities

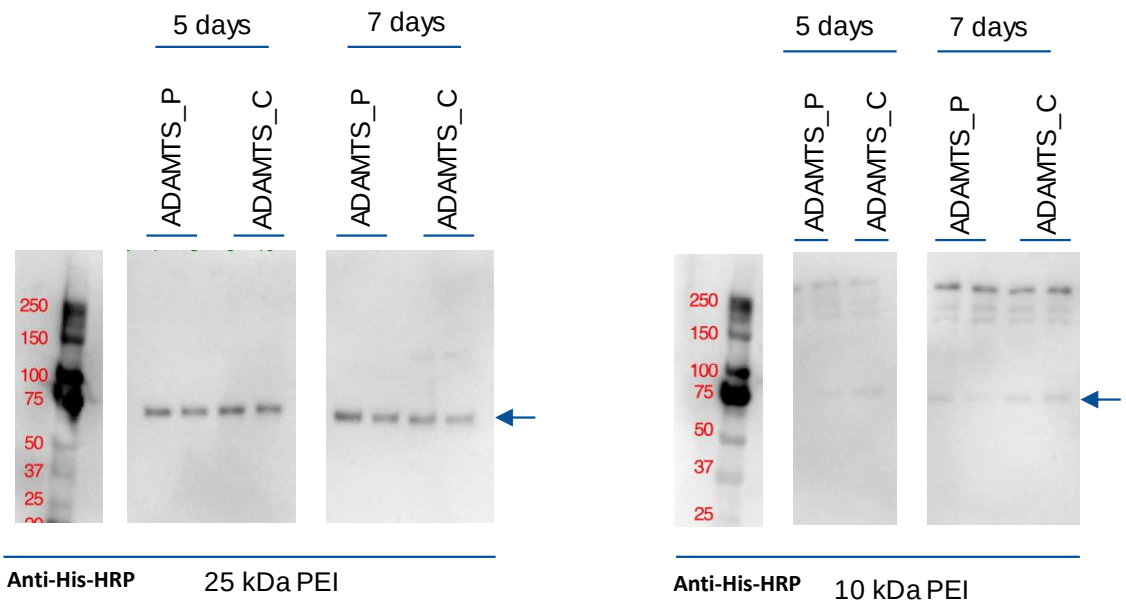


Figure 29. Optimization trials with variables the size of the PEI and the protein expression period (S2 cells). Two PEIs of different molecular weight (25 and 10 kDa) were assayed with two different proteins (ADAMTS -prodomain and ADAMTS -catalytic) at 5 and 7 harvesting days. Results were analysed by Western blot using a His-tag antibody.

Other checked transfection parameters were the PEI size (25 and 10 kDa) and the harvesting days post-transfection (5 and 7 days) (Fig. 29). For the two transfected proteins, the best PEI was the linear 25 kDa. Using this PEI size, similar results were found for both 5 and 7 transfection days. However, there was protein only at day 7 in the case of 10 kDa PEI and the expressed protein yields were very low (the experimental setup was exactly the same in both cases and the results are based on the band density on the gels).

We concluded that DNA delivery into S2 cells by PEI-mediated transfection was a very interesting option for transient transfection on a large scale. It was affordable and the previously proposed protocol (Shen, 2014), simple and reproducible. Linear 25 kDa PEI was definitively better than the 10 kDa one and that the cell centrifugation and seeding in new pre-heated medium was important to reach high cell densities during the transfection and thus, higher protein yields expressed. Cell densities 12-15x10^6 cells/mL for transfection have shown to be the best ones following this protocol. Regarding the harvesting days, they seemed not to make big differences between 5 and 7 days post-transfection (at least with these proteins). However, we decided to assay this parameter independently for each tested protein, because this can be a vary from protein to protein. Although it is not important for robust-stable proteins, it could be critical for the stability of degradation prone proteins.

3.1.2.1.6 Generation of polyclonal cell lines in S2 cells

S2 cells, derived from embryonic *Drosophila* cells, are one of the few *Drosophila* cell lines used for heterologous gene expression. The *Drosophila* expression system (DES) is designed for recombinant gene expression using the *Drosophila* constitutive actin5C promoter (Chung and Keller 1990) or the inducible MT promoter (Bunch et al. 1988) in clonal cell lines (Cherbas and Cherbas 2007). Based on these studies, we decided to develop stable overexpressing cell lines in order to increase the expression yields of difficult protein targets that are not successfully expressed by transient expression.

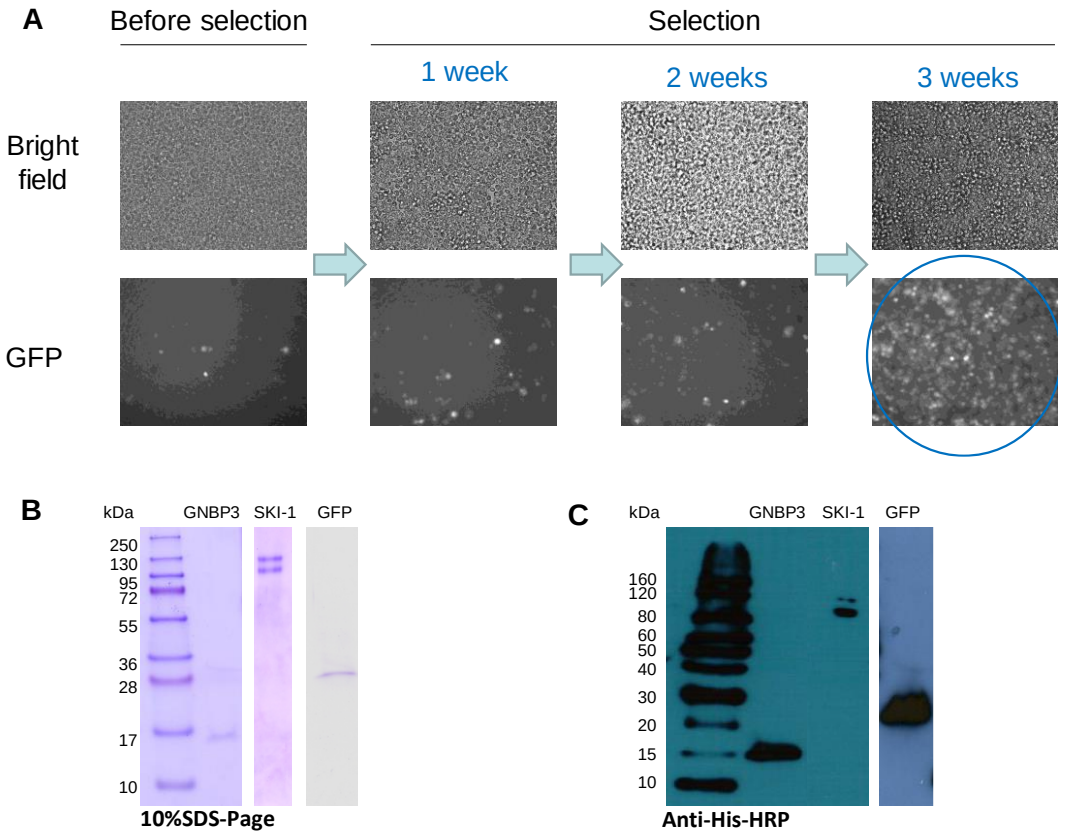


Figure 30. Stable transfected cells: selection, protein expression and purification trials using the *Drosophila* expression system. (A) Overexpression of GFP (pMT/GFP-H8) in transfected *Drosophila* cell cultures with puromycin selection for 3 weeks. Fluorescence microscopy at 72h post-induction. Blue circle indicates the successful selection process. **(B)** Expression and purification by affinity chromatography of GFP (intracellularly), GNBP3 and SKI-1 (extracellularly). **(C)** Western blot analysis of overexpressed and purified S2 proteins.

First, puromycin was tested as a selection agent for insect cell pool generation. S2 cells were incubated with various concentrations of puromycin. A concentration of 10 µg/mL of puromycin completely killed non-transfected S2 cells within 5 days and it was decided that a puromycin concentration of 10 µg/mL would be used for cell pool generation.

As a proof of concept, we expressed two standard proteins, GFP and Gram-negative binding protein 3 (GNBP3) and the ectodomain construct of human Subtilisin/kexin-isozyme 1 (hSKI1). S2 cells were grown in tissue culture flasks for the semi-adherent monolayer cell cultures with SF900II medium supplemented with 10% FBS or in 50-ml flasks shaken at 220 rpm in an orbital shaker for suspension cell cultures. S2 cells were co-transfected with the recombinant DNA (pMT/GFP-H8 or pMT/GNBP3-H8 or pMT/GNBP3-hSKI1(18-998)-H8) and pCoPuro (selection vector) at a ratio of 8:1 using the calcium phosphate protocol previously described. After two days of growth in normal harvesting conditions (no selection antibiotic into the medium), the transfected cells were selected with 10 µg/ml puromycin for 3 weeks and then sustained with 1 µg/ml puromycin (Fig. 30A). Viable cell densities of these stock cultures were kept above of 10^6 cells/mL by continuous passage. For the expression in stable cell cultures either in monolayer or in suspension, 2-5x10^6 cells/mL were induced with **$CuSO_4$** at a final concentration of 0.5 mM. The induced cells were harvested 3 days post-induction and the expression levels assessed by SDS-PAGE (Fig. 30B) and Western blot using an antibody against the His-tag (Fig. 31C).

With this assay, we proved the feasibility of generating stable insect cell pools for sustained recombinant protein production. For the following investigation, we decided to use it for difficult targets, not satisfactorily expressed by transient transfection.

3.1.2.2 Establishment of cell culture and transfection conditions and protein expression in the HighFive (Hi5) cell system

3.1.2.2.1 Adaptation of Hi5 cells to suspension and different media

HighFive (Hi5) cell line is one of the preferred insect cell lines for recombinant protein production using BEVS. This is mainly due to its higher volumetric productivity and better post-translational modification capacity for secreted glycoproteins (Drugmand et al. 2012; Granados et al. 1994). Because of this, we considered that it was important to test this cell line with our approach (cells in suspension and transient transfection by PEI).

Hi5 cells were first adapted to Sf900II medium supplemented with 10% FBS. The FBS was then gradually reduced up to 0.5% and finally 0%. The Hi5 cells were in parallel adapted to suspension with trypsin treatment for 4 weeks to avoid the formation of cell clumps. Once the suspension of Hi5 cell

lines were generated and stabilized, their cell growth was studied for 4 days and with triplicate experiments (Fig. 31).

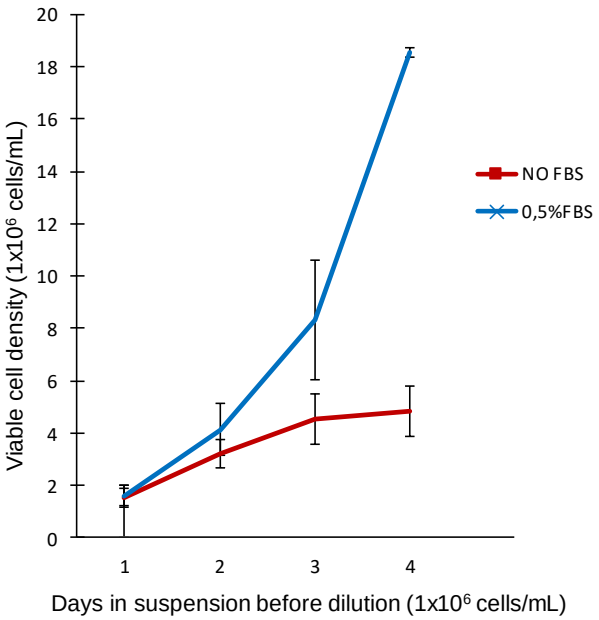


Figure 31. Growth curves of Hi5 cells in varying FBS concentration, using Sf900II medium at 220 rpm and 28°C. X axis represents the days that cells are growing before being split up to 1 x 10^6 cells/mL. Y axis represents the viable cell density.

Hi5 can not grow in a free FBS medium, so we concluded that cells cultured in 0.5% FBS were the best adapted to suspension, despite they tended to grow better in plates in the form of semi-adherent monolayers. We continued with the 0.5% FBS-suspension line for further experiments, even though not being optimal for big scale protein production from suspension cultures.

3.1.2.2.2 Transient transfection of Hi5 cells mediated by polyethylenimine

Hi5 cells kept in suspension were transfected following the PEI transfection protocol from Excellgene (given during the internship, www.excellgene.com). The two types of PEI (25 and 10 kDa) were assayed in parallel at 3 and 4 harvesting days with two different proteins (ADAMTS -prodomain and -catalytic fragments). The transfection efficiency was followed by Western blot using an anti-His antibody. (Fig. 32).

Transfections of Hi5 cells were successful in all the combinations, but the 25 kDa PEI was the most efficient independently of the protein expressed and the harvesting days.

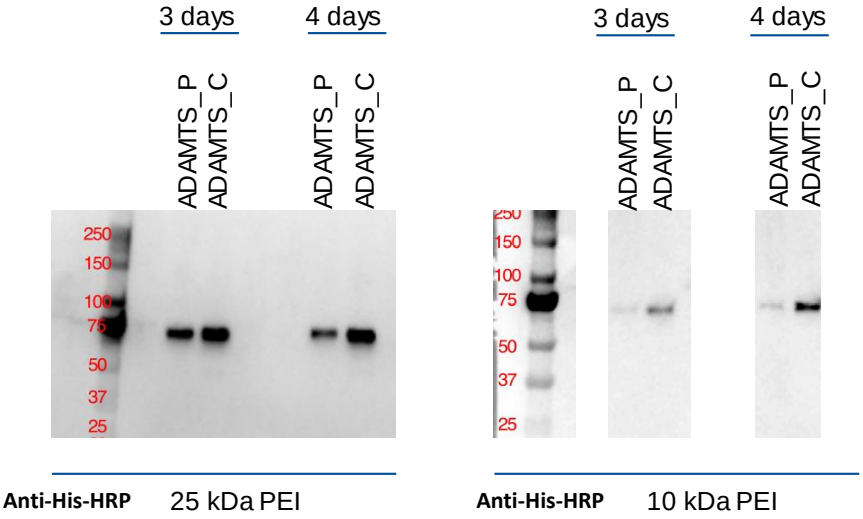


Figure 32. Optimization of protein expression trials by using the Hi5 cell line with variables the PEI and harvesting days. 25 and 10 kDa PEI were assayed with two different proteins at 3 and 4 harvesting days post-transfection. Results were analysed by Western blot using an anti-His antibody.

3.1.2.3 Establishment of cell culture conditions and transfection for protein expression in mammalian cells

3.1.2.3.1 Trial of transient transfection and recombinant protein expression in HEK293T cells

The HEK293T cell line, originally referred as 293tsA1609neo, is a highly transfectable derivative of human embryonic kidney 293 cells and contains both the SV40 T-antigen and the neomycin resistance gene. It grows in monolayers; the culture medium requirements are affordable (DMEM + 2mM Glutamine + 10% FBS) and it is commonly used for transient transfection. For these reasons, we thought that this could be an interesting alternative for the expression of some of our more challenging human protein targets, not successfully expressed with the insect systems. In addition, it has been demonstrated to be better for glycoprotein production.

Following the basis of the transient transfection protocol provided by Nadia Sukusu Nielsen (collaboration with the University of Aarhus) we cloned in the pCMV-Sport6 and expressed several difficult human proteins. Briefly, 18 µg plasmid DNA were mixed with 25 kDa PEI in a 10:1 ratio (PEI:DNA) per transfection plate (56.7 cm^2 plates), incubated for 15 mins and then added dropwise to the cell cultures. After four harvesting days, the culture was collected and centrifuged to separate the

supernatants from the cells. Subsequently, supernatant was purified by Ni-NTA and protein expressions were analysed by Western blot and SDS-PAGE.

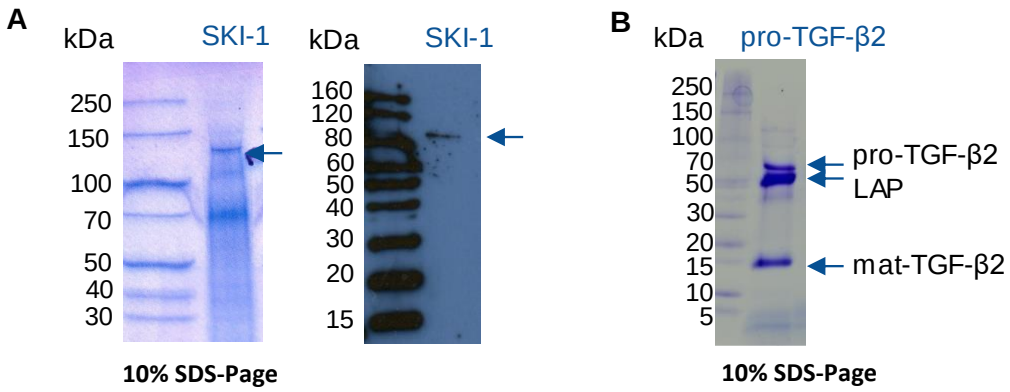


Figure 33. Protein expression optimisation trials with HEK293T cells by PEI transfection in plates. (A) Overexpression of hSKI1 in transfected HEK293T cells. Purified protein was analysed by Western blot using an antibody against the His-tag and SDS-PAGE stained with Coomassie Brilliant Blue. **(B)** Overexpression of pro-TGFβ2 in transfected HEK293T cells. Purified protein and fragments (full-length, LAP domain and mature protein) were detected by SDS-PAGE stained with Coomassie Brilliant Blue.

The protein expression was positive for both targets (Fig.33). hSKI1 has been always expressed using stably transfected S2 lines, but never before with transient transfection and even less with mammalian cells, so this was an alternative that had to be considered.

In the case of pro-TGFβ2, the protein expression in S2 cells was complicated and very low, while following this new approach, the protein yield (~1.2 mg per liter of cell culture) was good enough for biochemical and structural studies.

Based on these results we decided to stablish the last expression system of this thesis, to be able to express difficult human glycosylated proteins in sufficient yields for subsequent studies..

3.1.2.3.2 Establishment of a mammalian protein expression system with cells in suspension (Expi293F cells)

Expi293F human cells are derived from the 293-cell line and are maintained in suspension culture. They are typically grown in shaker incubators to high density in presence of 5-8% CO_2. This human cell line is also highly transfectable and its protein yields are even better compared to standard 293 cell lines in transient protein expression. However, the Expi293™ expression medium (Gibco) is sold at a high price (306€ / 1L), so we decided to substitute it for another medium with high protein expression levels and if it was possible, without compromising the protein yield. Two adapted Expi293 cell lines

were generated in the first round of media comparison (adapted cells to FreeStyle293™ medium and to a mixture of Expi293™ and FreeStyle293™ with a proportion 1:3 (Expi:FreeStyle).

FreeStyle™293 Expression Medium was chosen because it is used for the maintenance of the parental cells of Expi293F cells. Composition should therefore be similar, and its price deserved the attempt (174€ / 1L). Nevertheless, the cell growth at day 7 resulted to be less than the half of cell growth of cells in Expi medium (Fig.34A).

Fortunately, the combination 1:3 of both media (Expi:FreeStyle) (210€ / 1L) allowed a growth rate close to that obtained with the original medium, for at least the three first days (good enough for a normal transfection) (Fig.34A).

The protein yields were also compared, following a PEI-transient transfection protocol already published (Lukassen et al., 2016). They seemed to be very similar (Fig.34B), so we decided to use from then on the combination 1:3 (Expi:FreeStyle) for maintenance and transfection of Expi293F.

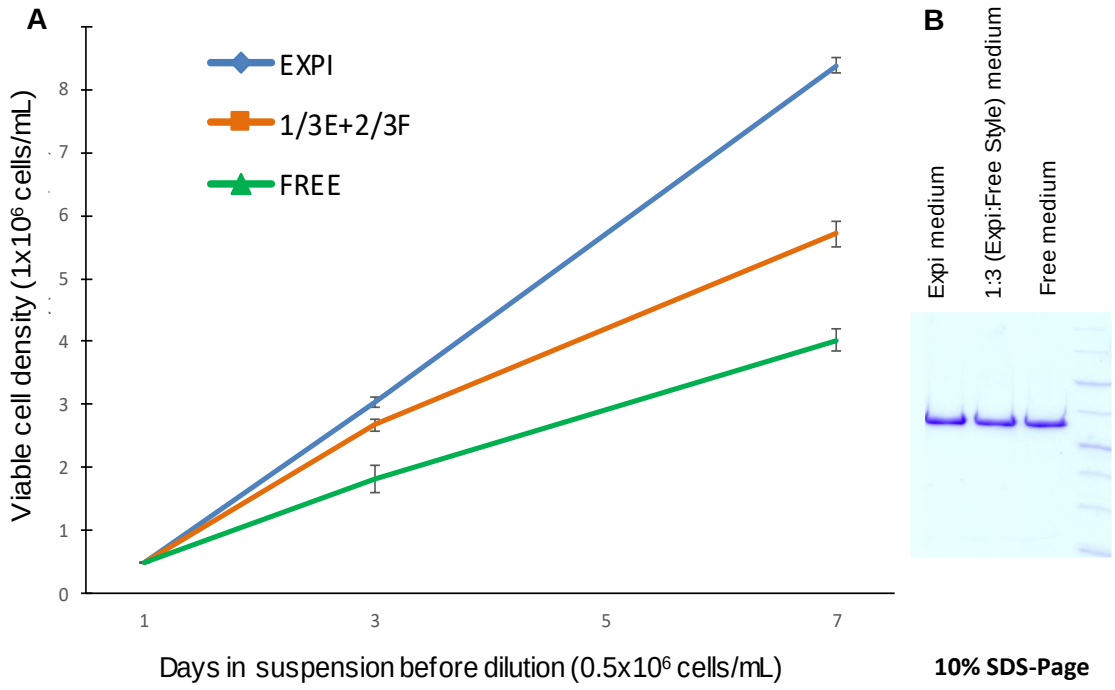


Figure 34. Comparison of cell growth and transfection yields with different mammalian expression media (Expi293F™ cells, first round). (A) X axis represents the days that cells are growing before being split up to 0,5x10⁶ cells/mL. Y axis represents the viable cell density**. (B)** Expi293F cells and adapted Expi293F cells to two different media (1:3 Expi293:FreeStyle293 and FreeStyle293) were transfected with TGFβIP pDNA and harvested after 3 days. Results were analysed by SDS-PAGE and stained with Coomassie Brilliant Blue.

We performed a second round of media optimization and in this case we compared the growth curve and protein expression with the Expi293F cells in their original medium, adapted Expi cells to the 1:3 (Expi:FreeStyle) medium and adapted Expi cells to FreeStyle™ F17 expression medium (Gibco). This medium had a similar cost (179€ / 1L) as FreeStyle™ 293.

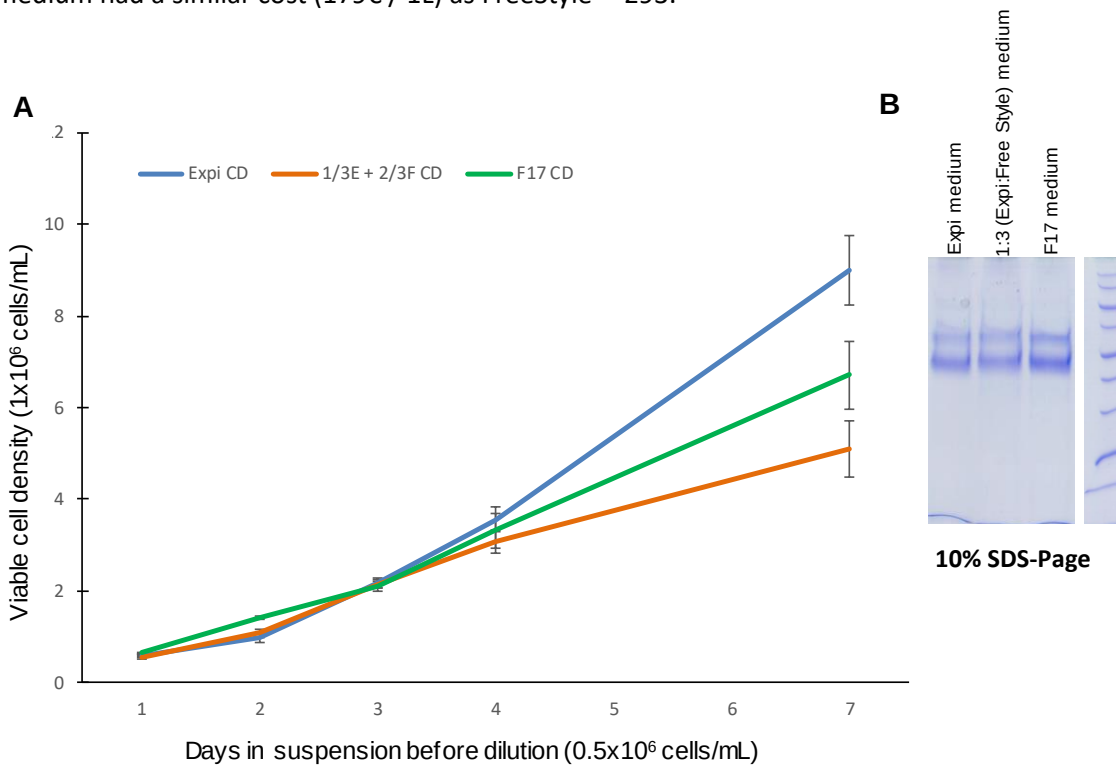


Figure 35. Comparison of cell growth and transfection yields with different mammalian expression media (Expi293F™ cells, second round (A) X axis represents the days that cells are growing before being split up to 0,5x10⁶ cells/mL. Y axis represents the viable cell density**. (B)** Expi293F cells and adapted Expi293F cells in two different media (1:3 Expi293:FreeStyle293 and FreeStyle F17 media), were transfected with pro-TGF-β2 pDNA (pS6-TGFβ2-H8) and harvested after 3 days. Results were analysed by SDS-PAGE and stained with Coomassie Brilliant Blue.

The growth rates of the three analysed media were very similar up to day 4 (Fig. 35A) and considering the regular transfections (harvested after 3 days), any of these media could be equally used for transfection. Regarding the protein expression analysis, the highest protein yield was for the Expi cells adapted to FreeStyle F17 medium (Fig. 35B). F17 medium was therefore chosen to replace the very expensive Expi Expression medium. The formula was also supplemented with 0.2% Pluronic F-68 to avoid clump formation.

3.1.2.3.3 Trial of transient transfection and recombinant protein expression in CHO cells

CHO cells have been the most commonly used mammalian host for large-scale commercial production of therapeutic proteins. Recent advances in cell culture technology for CHO cells have achieved significant improvement in protein production, so we considered it appropriate to test the system during the internship at the company Excellgene (expert on Eukaryotic protein expression) with some difficult targets, never expressed in mammalian cells human (catalytic and BTMD hSKI1 proteins).

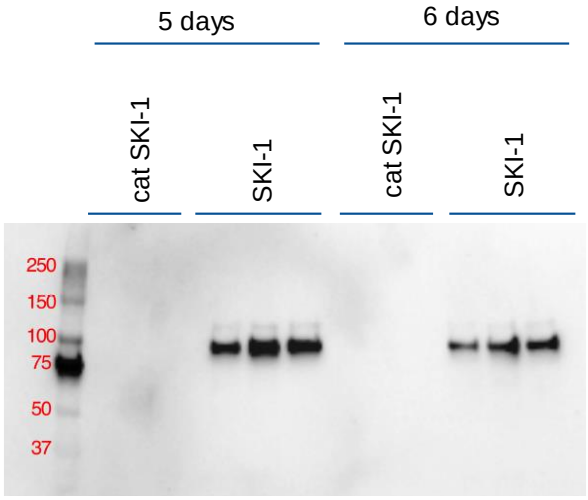


Figure 36. PEI transfection optimization trials for protein expression in CHO cells. 25 kDa PEI was assayed for expression of cat-SKI1 and SKI-1 in the CHO expression system. Proteins were collected and purified after 5 or 6 days. Results were analysed by Western blot using an antibody against the His-tag.

The protein transfection and expression were successful for the ectodomain construct of hSKI1 (Fig. 36), especially after 5 days of harvesting. This was done using the Excellgene's transfection protocol and their suspension of CHO cells in a new medium (M40), not yet commercially available. We concluded that this mammalian expression system could be a good alternative for the Expi293F system. They would work in parallel, because they have interesting differences, like glycosylation introduction. These post-translational modifications can have big differences in the processing and recognition of proteins. Another exciting characteristic of these cells are the well-known stable protein production with yield reaching more than 10 g/L.

3.1.3 CONCLUSIONS AND FUTURE PERSPECTIVES

On this part of the thesis, two eukaryotic expression systems have been established. One system based on insect cell cultures (*Drosophila melanogaster*, S2) and the another on mammalian cell cultures (human cells, Expi293F). Many more have been tested (Hi5, HEK293T, CHO cells) in order to choose the most efficient system for recombinant protein production.

In the case of insect cells, Hi5 and S2 lines were initially studied. Both cell cultures were adapted to suspension to reach higher cell densities and to achieve increased protein yields. However, only S2 cells were able to survive in lack of FBS into the medium without clump formation for long periods. This was fundamental to reduce the price of the expression system. Therefore, S2 cells were optimised further. Culture conditions were studied including, optimal culture volume (10 mL for maintenance and 200 mL for transfection in TubeSpin bioreactor tubes) and harvesting conditions (at 200 rpm and 28°C). Several affordable insect expression media were analysed (IPL-41, TC-100 and Sf900II) with the best for protein expression in suspension being the Sf900II (79,50€ / 500mL).

There were also assayed several transfection reagents and up to three transfection protocols which were optimized. In the case of calcium phosphate transfection, it was used for both kind of transfections (transient and inducible, to generate the stable transfected S2 cell lines). Lipofectamine™ 2000 reagent resulted in being expensive for big scale protein production and less efficient than calcium phosphate on our hands. But on the other side, polyethylenimine, and specially the 25 KDa, demonstrated to be perfect for our interests. Inexpensive large transient transfections were performed following the modified Xiao Shen's protocol. S2 cells were centrifuged and resuspended in prewarmed fresh medium to a cell density of 15×10^6 cells/mL. A mixture of 0.6 µg DNA and 2 µg PEI per 1×10^6 cells and per prewarmed transfection volume was pre-incubated for 15–20 min at room temperature and then added dropwise to the cell cultures. These were further incubated for 1 hour at 28 °C and 220 rpm, subsequently diluted with prewarmed fresh medium to 5×10^6 cells/ml and harvested after seven days for protein purification.

Regarding mammalian cell cultures, three cell lines were mainly tested: HEK293T, Expi293F and CHO cells. All of them were transiently transfected with 25 kDa PEI, following different protocols attending to the features of each cell line. First assays were performed in monolayer-HEK293T cells, pursuing the protocol provided by Nadia Sukusu Nielsen (University of Aarhus) and successful results were obtained, which opened the door to establish another eukaryotic expression protein system. Next experiments were executed and scaled up in suspension human cells, called Expi293F. These cells were originally grown in Expi expression medium, but they were adapted to different media (FreeStyle™293, mixture 1:3 Expi:FreeStyle and FreeStyle™ F17) to reduce their expense. After two rounds of optimization, the

best medium in terms of affordability and protein yield was the F17. About protein expression, Expi293F cells were transfected at 1×10^6 cells/mL with a mixture of 1 mg of DNA and 3 mg of PEI in 20 mL of Opti-MEM Medium per liter of expression medium and after 3 harvesting days post-transfection, high protein yields were obtained.

Another promising result was the achieved in CHO cells by PEI transfection at Excellgene. Their glycosylation introduction in human recombinant proteins was slightly different and the stable line generation was well-known. Therefore, this interesting mammalian expression system must be highly considered as the next implementation in our laboratory.

Figure reproduced from “gelifesciences.com”.

RESULTS

3.2 Functional and structural studies on the human subtilisin kexin-isozyme 1 (hSKI1)

3.2.1 OBJECTIVES

Hamster SKI1/S1P was the first SKI1 expressed protein (Cheng et al., 1999). Full-length (aa 1-1052) and before the transmembrane domain (aa 1-983) constructs of this protein were expressed by stable transfected cell lines in mutant CHO cells, deficient in S1P (SRD-12B cells), using roller bottles. They identified three fragments for SKI1, which were analysed by immobilized metal affinity chromatography (IMAC) from pellet (two fragments) and supernatant (the third one). With this experimental approach, a protein yield of around 0.4-0.8 mg/L cell culture for the soluble hamster SKI1 was obtained.

The discovery of human SKI1 (hSKI1) happened in 1999 by the cloning of a human subtilisin-related c-DNA, which was identical to the hamster S1P (Seidah et al., 1999). In this case, the expression of the full-length hSKI1 was by recombinant Vaccinia virus infection and, in parallel, generation of stable transfected HEK293 cells. However, they did not report any protein yield and all the activity assays were performed in the total cell culture, followed by immunoprecipitation with specific antiserums.

Touré and co-workers (Touré et al., 2000) also tried the infection of mammalian cells (BSC40 and LoVo cells) with recombinant Vaccinia virus and stable transfected HEK293 cells. In this case, hSKI1 was properly purified but unfortunately, the limited quantity of purified protein made it not possible to carry out full kinetic studies.

It was not until 2006, when Bodvard and his collaborators. managed to produce 0.4-0.8 mg/L cell culture of partially purified hSKI1 expressed in Sf9 cells by BEVS (Bodvard et al., 2006).

So, regarding the literature, most of the characterizations of the human SKI1 protein came from total cell lysates or partially purified protein and no structural information had been reported.

In this context, the major goals of this part of the thesis were:

- The generation and expression of different constructs of human SKI1 in as many protein systems as possible, as they were needed for its biochemical and biophysical characterization.
- To establish a good purification protocol to obtain sufficient amount of pure protein and in this way, to attempt crystallization and determination of the structure.

3.2.2 RESULTS AND DISCUSSION

3.2.2.1 Analysis of hSKI1 sequence

Human SKI1 gene (MBTPS1) encodes a 1052-residue protein with an estimated molecular weight of 148 kDa for the zymogen, which is processed progressively into two other membrane forms of approximately, 120 and 106 kDa in the ER and then, into a secreted form of 98 kDa (Seidah and Prat, 2012).

According to Uniprot (entry Q14703), SKI1 zymogen comprises a signal peptide (SP) of 17 amino acids, followed by a prodomain (up to aa 186), a subtilase domain (aa 190-472) with three active sites (Asp[218], His[249] and Ser[414]), a transmembrane domain (aa 999-1021) and finally, a predicted basic cytosolic tail (aa 1022-1052).The N-terminal prodomain suffers two rounds of processing (cleavages B'/B in *cis* = (**RKVF**[133]↓ **RSLK**[137]↓; and cleavages C'/C in *trans* = **RRAS**[166]↓LSLGSGFWHATGRHSS**RRLL**[186]↓) and the active form of SKI1 is autocatalytic shedded into KHQKLL[953]↓to be secreted (in bold positions P4 and P2 of the consensus motif). Up to five N-glycosylations linked to asparagines are also described (236, 305, 515, 728 and 939 positions) (Fig. 9).

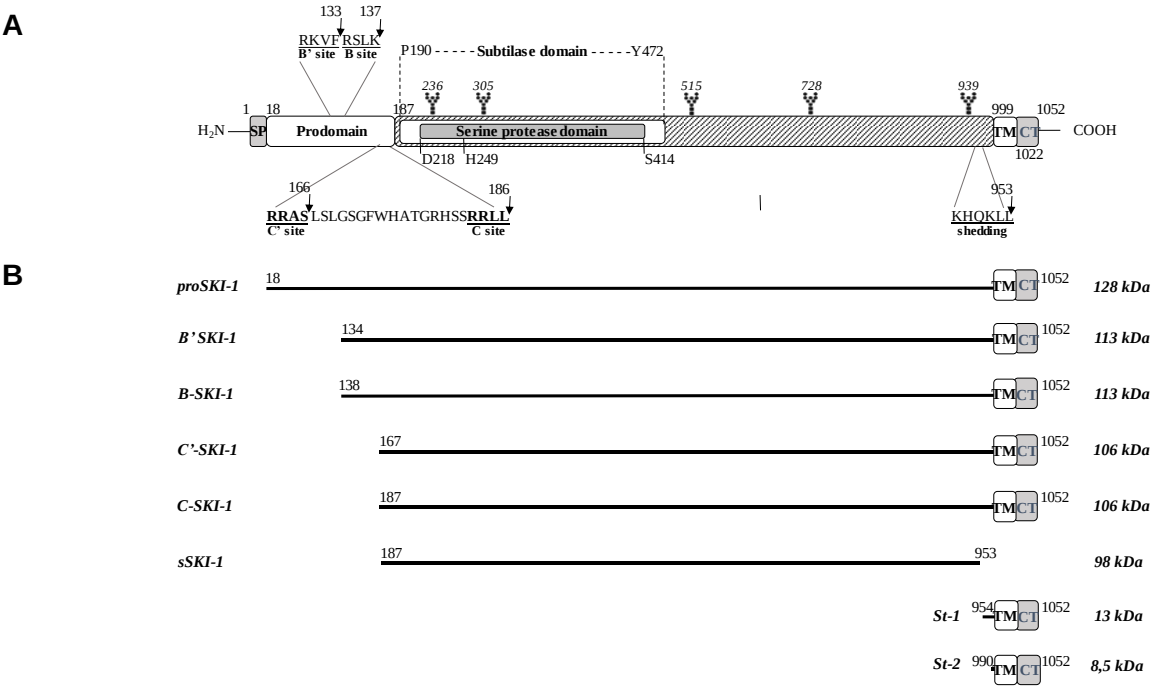


Figure 9. Schematic representation of hSKI1 protein and the processed forms of SKI1 after the multistep maturation process. The residue numbers correspond to UP, entry Q14703. **(A)** Functional domains of zymogen SKI1 are the signal peptide (SP), prodomain (PD), transmembrane (TM) and cytosolic tail (CT). N-linked glycosylation sites are highlighted with a sugar chain. Cleavages and shedding motifs are shown in magnification and underlined. The exact cleavage is also pinpointed with an arrow. **(B)** Scheme depicting the basic domain structure of the different processed SKI1 forms and stubs that result after the B'/B and C'/C cleavages and the subsequent shedding. Amino acid length and molecular weight are given for each processed form of SKI1.

3.2.2.1.1 Signal peptide prediction

Previous data were corroborated by informatic tools. Several analyses were performed for the prediction of the signal peptide, first using: SignalP-5.0 (http://www.cbs.dtu.dk/services/SignalP/) and Phobius prediction (http://phobius.sbc.su.se/) servers (Almagro Armenteros et al., 2019). Predicted signal peptide cleavages for both servers were between residues 22 and 23 (Fig. 37). A second round of analysis was performed to confirm this result (http://topdb.enzim.hu/), because up to now, the published signal peptide cleavage of hSKI1 was between 17 and 18 residues (Fig.38).

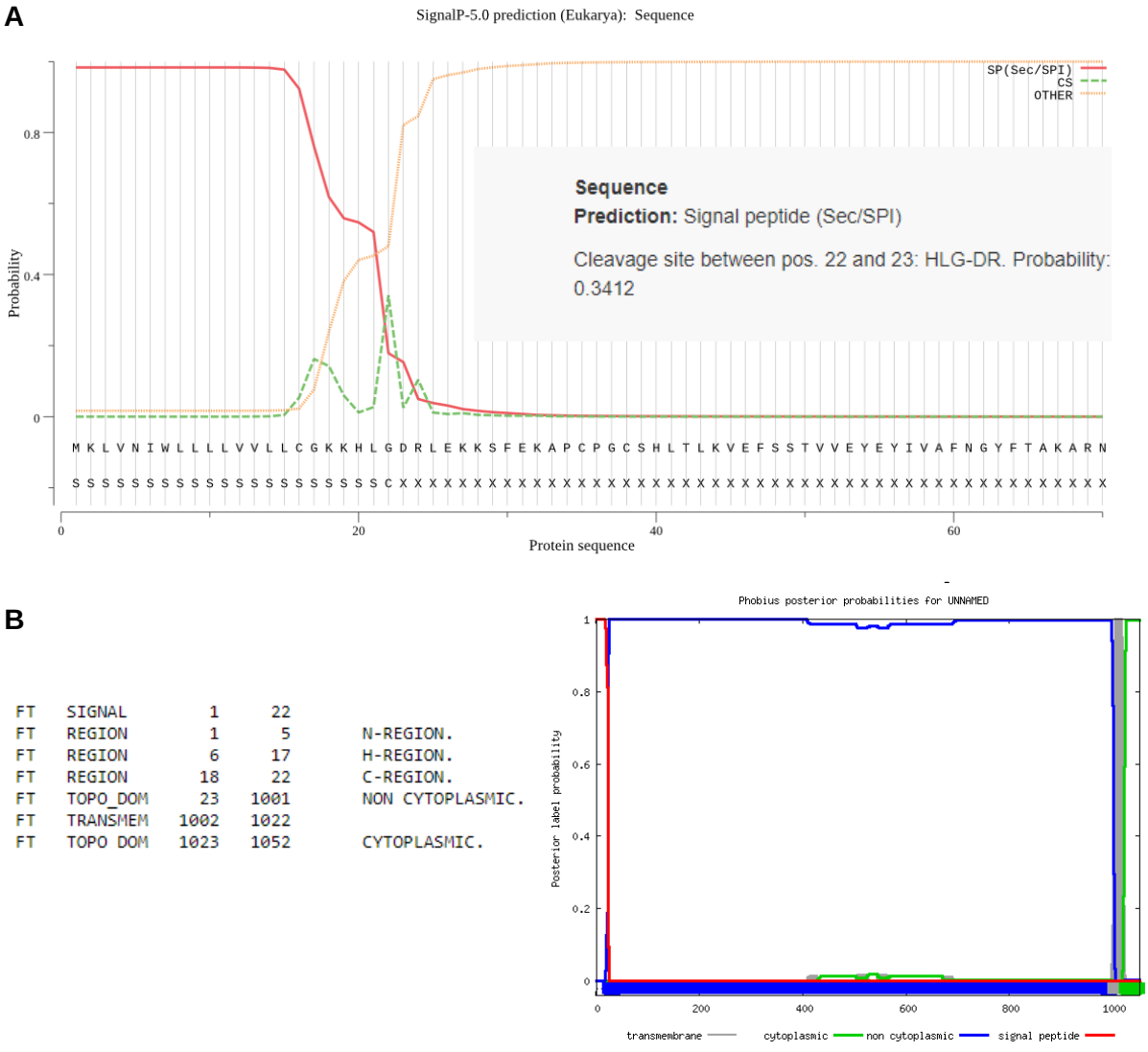


Figure 37. Signal peptide prediction of hSKI1. (A) Prediction by SignalP**.** Three likelihood probabilities are reported on this plot, the SP(Sec/SPI) in red ("standard" secretory signal peptides transported by the Sec translocon and cleaved by Signal Peptidase I (Lep)), the CS (the cleavage site) in green and OTHER in orange (the probability that the sequence does not have any kind of signal peptide). Y axis represents the probability of the three factors (SP(Sec/SPI), CS and OTHER). X axis represents the protein sequence with two lines, upper with the amino acid sequence and bellow, its correspondence with S (signal peptide), C (cleavage) or X (non-signal peptide). **(B)** Prediction by Phobius. The plot shows the probabilities for signal peptide in red, non-cytoplasmic in blue, cytoplasmic in green and transmembrane in grey and localizes them into the protein sequence. Being Y and X axis, probability and protein sequence, respectively.

Membrane	Outside: Cysternal space	Endoplasmic reticulum membrane	Inside: Cytoplasm

Topology

Number of transmembrane segments: 1
Reliability of this topology: 85

	Begin	End	Localisation
1	1	17	Signal
2	18	1001	Cysternal space
3	1002	1022	Endoplasmic reticulum membrane
4	1023	1052	Cytoplasm

Figure 38. Signal peptide prediction of hSKI1 (second round analysis). Using the Topology Database of Transmembrane Proteins server (TOPDP) a signal peptide sequence from aa 1 to 17, a luminal domain from aa 18 to 1001, a transmembrane domain to aa 1002 to 1022 and a cytosolic tail from aa 2023 to 1052 were identified.

3.2.2.1.2 Transmembrane domain prediction

The used servers to determine the length and position of the transmembrane domain were: Phobius (Fig.37) (TM domain from aa 1002 to 1022), TOPDP (Fig. 38) (TM domain from aa 1002 to 1022), http://www.cbs.dtu.dk/services/TMHMM/ (Fig. 39A) (TM domain from aa 999 to 1021) and https://www.predictprotein.org (Fig 35B) (TM domain from aa 1000 to 1021).

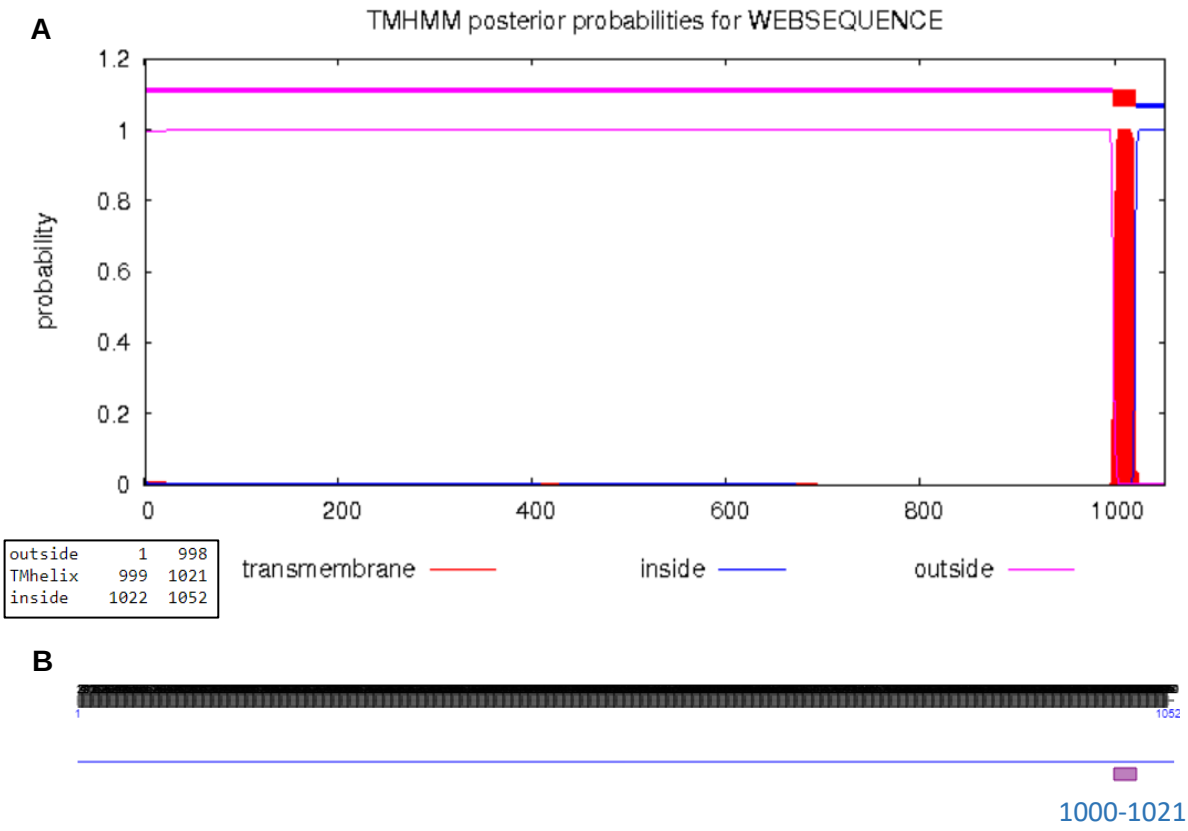
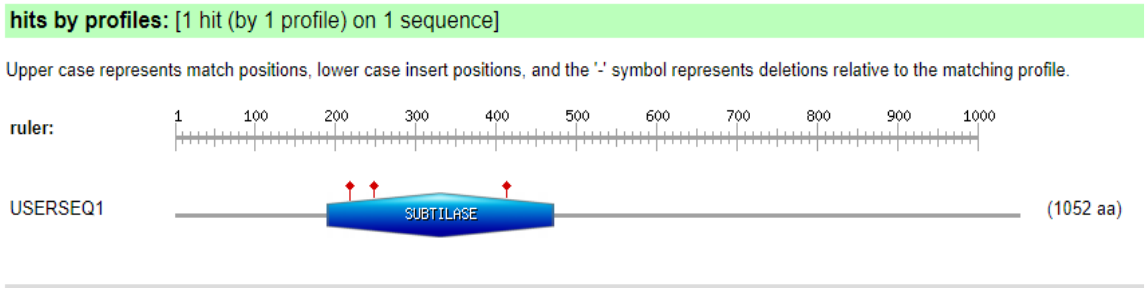


Figure 39. Transmembrane prediction of hSKI1. Several servers were used to identify the transmembrane domain position: TMHMM **(A)** and Predict Protein **(B).** Both programmes reported a transmembrane domain in the same region, but they disagreed about the length.

3.2.2.1.3 Functional domain prediction

The functional domains of hSKI1 were predicted by Prosite Expasy server (https://prosite.expasy.org/). A subtilase/peptidase S8 domain profile was identified from Pro^{190} to Tyr^{472} with three active sites in Asp^{218}, His^{249} and Ser^{414} positions (Fig. 40).

A



B

PS51892 **SUBTILASE** *Serine proteases, subtilase domain profile* :

190 - 472: **score** = 32.087

```
PRQVAQT--LQADVLWQMGYTGANVRVAVFDTGLSEKHPHFK-------NVKERTNWTNE
rtlDDGLGHGTFVAGVIASMR----ECQGFAPDAELHIFRVFTNNQVSYTSWFLDAFNYA
ILKKIDVLNLSIGGPDF--MDHPFVDKVWELTANNVIMVSAIGNDGPLYGTLN---NPAD
QMDVIGVGGIDFEDNIARFSSRGMTtwelpggygrmkPDIVTYGAGVRGS--GVKGGCRA
LSGTSVASPVVAGAVTLLVSTVQkreLVNPASMKQALIASARRLPGV-NMFEQGHGKLDL
LRAY
```

Predicted features:

DOMAIN	190	472	Peptidase S8	[condition: none]	
ACT_SITE	218		Charge relay system	[condition: D]	[group: 1]
ACT_SITE	249		Charge relay system	[condition: H]	[group: 1]
ACT_SITE	414		Charge relay system	[condition: S]	[group: 1]

Figure 40. Functional domain prediction of hSKI1. (A) This figure represents the match positions for the functional domain with a blue box and the identified active sites with red lollipops. **(B)** The protein sequence, highlighting the active sites in red, and the predicted features are also shown.

3.2.2.1.4 Hydropathicity level prediction

The grand average of hydropathicity (GRAVY) was calculated to attempt to predict the behavior of hSKI1 for the purification. It is calculated by adding the hydropathy values of each amino acid residues and dividing by the total number of residues in the sequence (Kyte and Doolittle, 1982). Positive scores indicate good hydrophobicity.

The GRAVY value for hSKI1 was -0.319, using the https://web.expasy.org/cgi-bin/protscale/protscale.pl server and including “Hphob. / Kyte & Doolittle” in the search. This result means that full-length hSKI1 is not especially hydrophobic despite its transmembrane domain. A plot with the hydropathy value of each amino acid in the sequence was also generated (Fig. 41).

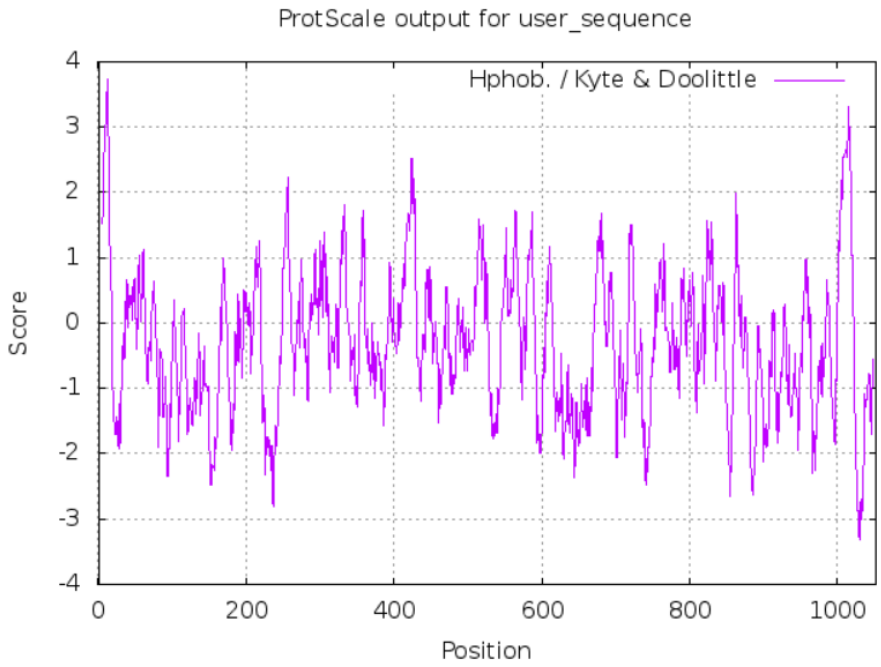


Figure 41. Hydropathicity value prediction of hSKI1. This plot represents the hydropathy values of each amino acid residues in the amino acidic sequence.

3.2.2.1.5 Disulphide bond prediction

The Scratch Protein Predictor (http://scratch.proteomics.ics.uci.edu/explanation.html) predicted the formation of disulphide bonds in the cysteines at the following positions: 16, 35, 38, 494, 501, 684, 791, 849, 859, 927 (Fig. 42).

Predicted disulfide bonds(cysteine pairs) ordered by probability in descending order:

Bond_Index	Cys1_Position	Cys2_Position
1	494	501
2	16	35
3	849	859
4	38	927
5	684	791

Figure 42. Disulphide bond prediction of hSKI1 protein. Predicted disulphide bonds ordered by probability in descending order.

These data were not reported in the literature, so they should be considered and corroborated during the protein purification.

3.2.2.1.6 N-glycosylation prediction

The used server to identify the N-glycosylations was: http://www.cbs.dtu.dk/services/NetNGlyc/. The prediction for hSKI1 was seven potential N-glycosylations, but only five above threshold (Fig. 43).

A

```
Name:  Sequence          Length:  1052
MKLVNIWLLLLVVLLCGKKHLGDRLEKKSFEKAPCPGCSHLTLKVEFSSTVVEYEYIVAFNGYFTAKARNSFISSALKSS      80
EVDNWRIIPRNNPSSDYPSDFEVIQIKEKQKAGLLTLEDHPNIKRVTPQRKVFRSLKYAESDPTVPCNETRWSQKWQSSR     160
PLRRASLSLGSGFWHATGRHSSRRLLRAIPRQVAQTLQADVLWQMGYTGANVRVAVFDTGLSEKHPHFKNVKERTNWTNE     240
RTLDDGLGHGTFVAGVIASMRECQGFAPDAELHIFRVFTNNQVSYTSWFLDAFNYAILKKIDVLNLSIGGPDFMDHPFVD     320
KVWELTANNVIMVSAIGNDGPLYGTLNNPADQMDVIGVGGIDFEDNIARFSSRGMTTWELPGGYGRMKPDIVTYGAGVRG     400
SGVKGGCRALSGTSVASPVVAGAVTLLVSTVQKRELVNPASMKQALIASARRLPGVNMFEQGHGKLDLLRAYQILNSYKP     480
QASLSPSYIDLTECPYMWPYCSQPIYYGGMPTVVNVTILNGMGVTGRIVDKPDWQPYLPQNGDNIEVAFSYSSVLWPWSG     560
YLAISISVTKKAASWEGIAQGHVMITVASPAETESKNGAEQTSTVKLPIKVKIIPTPPRSKRVLWDQYHNLRYPPGYFPR     640
DNLRMKNDPLDWNGDHIHTNFRDMYQHLRSMGYFVEVLGAPFTCFDASQYGTLLMVDSEEEYFPEEIAKLRRDVDNGLSL     720
VIFSDWYNTSVMRKVKFYDENTRQWWMPDTGGANIPALNELLSVWNMGFSDGLYEGEFTLANHDMYYASGCSIAKFPEDG     800
VVITQTFKDQGLEVLKQETAVVENVPILGLYQIPAEGGGRIVLYGDSNCLDDSHRQKDCFWLLDALLQYTSYGVTPPSLS     880
HSGNRQRPPSGAGSVTPERMEGNHLHRYSKVLEAHLGDPKPRPLPACPRLSWAKPQPLNETAPSNLWKHQKLLSIDLDKV     960
VLPNFRSNRPQVRPLSPGESGAWDIPGGIMPGRYNQEVGQTIPVFAFLGAMVVLAFFVVQINKAKSRPKRRKPRVKRPQL    1040
MQQVHPPKTPSV
.....                                                                                 80
...........N.....................................................N.....................160
.....................................................................N....              240
.....                                                                                 320
.....                                                                                 400
.....                                                                                 480
................................N.....                                                 560
.....                                                                                 640
.....                                                                                 720
.....                                                                                 800
.....                                                                                 880
.....................................................N.....                          960
.....                                                                                1040
............                                                                         1120
```

B

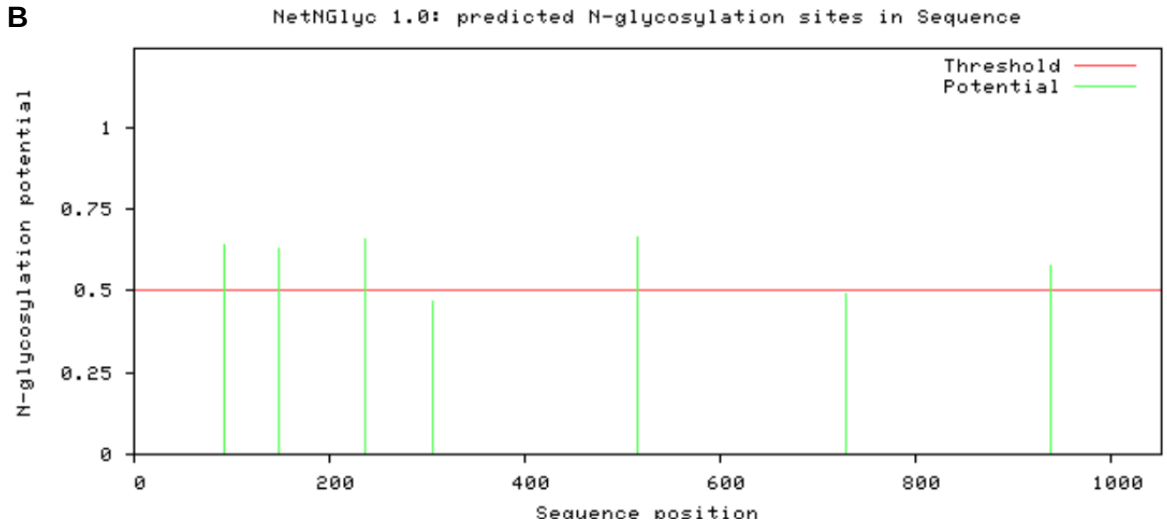


Figure 43. N-glycosylation prediction of hSKI1 protein. (A) hSKI1 amino acid sequence was analysed for N-glycosylations. Asn-Xaa-Ser/Thr in the sequence were highlighted in blue and predicted asparagines to be N-glycosylated were highlighted in red. **(B)** Representation of high probability (red line) to be N-glycosylated the potential asparagines (green line). Y axis represents the N-glycosylation potential and X axis the sequence position.

3.2.2.1.7 Subcellular localization prediction

The software DeepLoc-1.0 was used for the prediction of subcellular localizations of hSKI1 protein (http://www.cbs.dtu.dk/services/DeepLoc/). It was predicted to be a membrane protein with a very high probability (0.999) (Fig. 44)

Figure 44. Subcellular localization prediction of hSKI1 protein. (A) Subcellular location**, (B)** hierarchical tree and **(C)** position importance graphics and plots that result after performing an analysis of the protein sequence.

Another used server was the Scratch Protein Predictor with its tool ABTMpro. It predicted that the transmembrane protein was alpha helical with high probability (0.965087).

3.2.2.1.8 Secondary structure prediction

To study the secondary structure of hSKI1, several analysis were performed with more than one server: http://bioinf.cs.ucl.ac.uk/psipred; http://polyview.cchmc.org/; https://npsa-prabi.ibcp.fr/cgi-bin/secpred (GOR IV). Then, results were matched to generate a final graph with a sequencing image creator, called Geneious (https://www.geneious.com/) (Fig. 45).

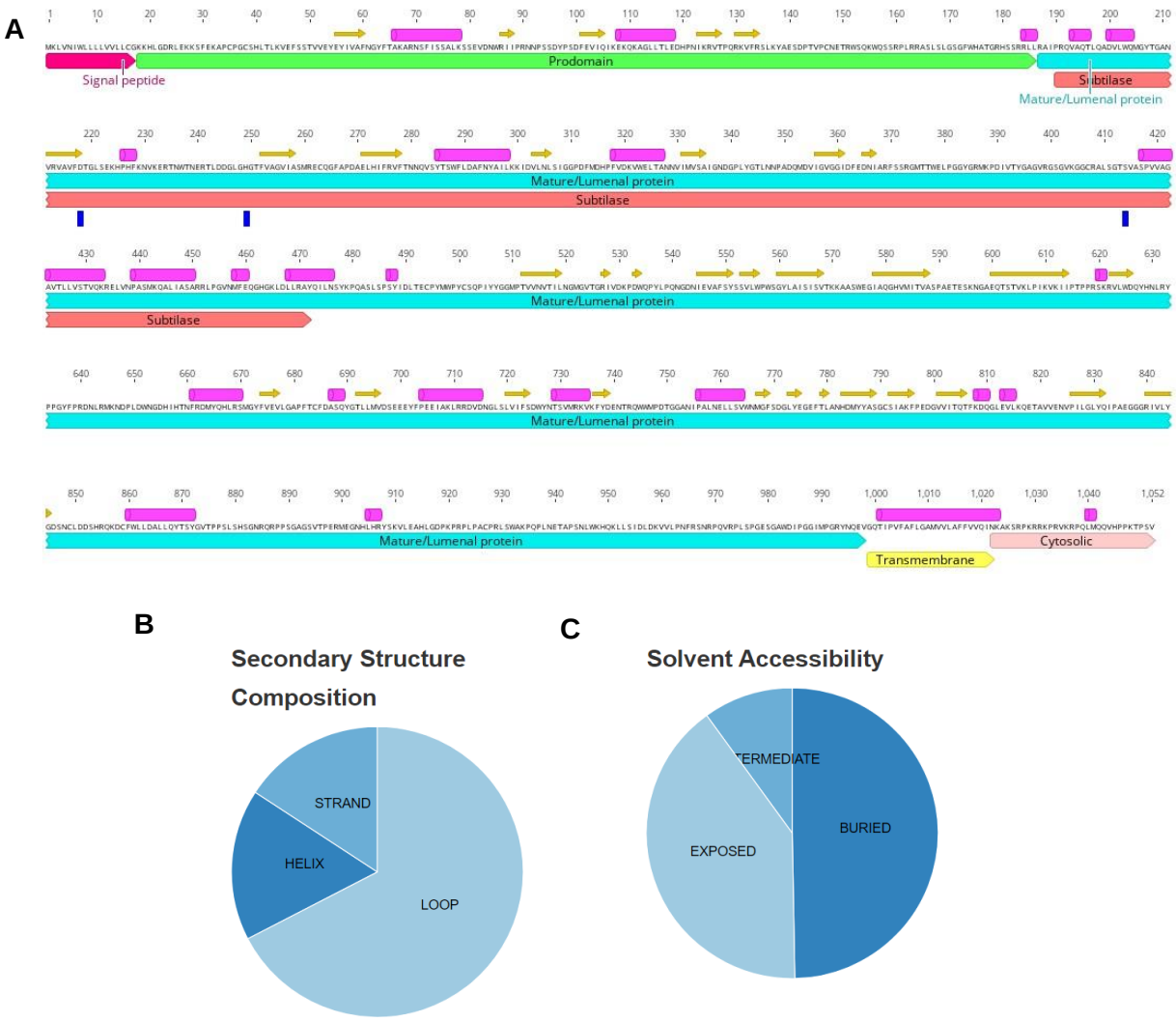


Figure 45. Secondary structure prediction of hSKI1 protein. (A) Yellow arrows represent β-strands and pink cylinders represent α-helices**. (B)** Secondary structure composition. **(C)** Solvent accessibility.

Prediction analysis were performed in order to confirm the data previously obtained from Uniprot. However, several discrepancies were found, like for example, the signal peptide length. There are two possibilities: cleavage peptide positions 22 and 23 and the reported at Uniprot, between positions 17 and 18.

On the other hand, there is a complete agreement about the transmembrane properties of hSKI1 and its subcellular location in organelles. Its alpha-helix transmembrane domain is predicted to place around positions 999 and 1021. Despite of this domain, it is not an especially hydrophobic protein.

Another discovery has been the potential disulphide bonds, never before described in the literature. This should be of course also be considered for protein purification because an inherent problem in the process of disulphide bond formation is oxidation of cysteines, which can cause misfolding, aggregation and finally, result in low yields during protein production.

In summary, all these predictions have been considered when making the final experimental design, for the creation of the DNA constructs of hSKI1, as well as for the purification and polishing protocols.

3.2.2.2 Design and generation of hSKI1 constructs

hSKI1 was cloned from human c-DNA and the sequence was validated by sequencing with its internal primers (see Table 8).

Multiple constructs spanning different hSKI1 fragments were then generated and inserted into different expression vectors for testing protein expression, biochemical and biophysical characterization and finally, crystallization (further details about these plasmids can be found in the "List of expression vectors" and in the corresponding pr).

The used protein expression systems with their respectively generated constructs were:

- The bacteria expression system with the pCri vectors (Goulas et al., 2014a) (see Table 9, in yellow the expressed constructs in inclusion bodies).
- The insect expression system with transient and inducible protein expression (see Table 10, in green the expressed constructs with soluble protein).
- The mammalian expression system with transient protein expression (see Table 11, in green the expressed constructs with soluble protein).

In order to get soluble expression of hSKI1, several cloning strategies were followed:

a) To remove the own hSKI1 signal peptide (from 1 to 17 residues), which naturally sends the protein to the ER, and to replace it for another one, like for a periplasmic or secreted expression.
b) To remove the transmembrane domain (ended constructs in V^{998}), that is recognized by the organelles (ER and Golgi apparatus).
c) To express the minimal expression unit of hSKI1 to facilitate the proper folding as soluble protein. Constructs up to residue 429 were cloned for serine protease expression and up to residues 473 and 477 for subtilase domain expression (the Q^{473} is straight after the described end of the subtilase domain and the S^{477} is its end avoiding disturbing the secondary structure). The expression of the catalytic domain was also assayed with (from k^{18}) and without (from A^{188} and A^{210}) prodomain.
d) Lastly, a hSKI1 construct was produced lacking the shedding site (before residue 953), to generate the soluble SKI1.

Apart from the engineering of hSKI1 fragments, the chosen vectors have several tags to facilitate protein purification. These purification strategies are going to be explained in detail bellow.

Table 8. Internal sequencing primers of hSKI1 and sequencing primers of the main expression vectors

Primer name	Parental DNA	Primer sequence (from 5' to 3')	Comments
SKI1_seq_rev 500	Human SKI1 c-DNA	GAATTCCACCTTCAAAGTCA	Reverse primer to sequence the N-terminal end of hSKI1 DNA. Primer annealing in SKI1 prodomain.
SKI1_seq_fwd 781	Human SKI1 c-DNA	CAAACATCAAACGGGTCACG	Forward primer to sequence hSKI1 DNA. Primer annealing in SKI1 prodomain.
SKI1_seq_fwd 1561	Human SKI1 c-DNA	CAGGAGGCTACGGTCGCATG	Forward primer to sequence hSKI1 DNA. Primer annealing in SKI1 subtilase domain.
SKI1_seq_fwd 1731	Human SKI1 c-DNA	GCATTGCTCAGGGCCATGTC	Forward primer to sequence hSKI1 DNA. Primer annealing in SKI1 lumenal domain.
SKI1_seq_fwd 2401	Human SKI1 c-DNA	TCAGGGATATGTACCAGCAT	Forward primer to sequence hSKI1 DNA. Primer annealing in SKI1 lumenal domain.
SKI1_seq_fwd 3121	Human SKI1 c-DNA	AAGGAAACCATCTTCATCGG	Forward primer to sequence hSKI1 DNA. Primer annealing in SKI1 lumenal domain, before the transmembrane domain.
T7 promoter (fwd)	pCri vectors	TAATACGACTCACTATAGGG	Forward primer to sequence the N-terminal end of the clonings performed and inserted into pCri vectors (Goulas et al., 2014a). These vectors are for protein expression in BL21 (DE3) bacteria cells, which have a copy of T7 RNA polymerase and kanamycin resistance. They also make easier the protein purification with different tags (TEV, tobacco-etch virus peptidase; TRX, thioredoxin; Strep- and His-tags…).
T7 terminal (rev)	pCri vectors	GCTAGTTATTGCTCAGCGG	Reverse primer to sequence the C-terminal end of the clonings performed and inserted into pCri vectors.
MT_fwd	pMT/V5-His A and pMT/BiP/V5-His A vectors	CATCTCAGTGCAACTAAA	Forward primer to sequence the N-terminal end of the clonings performed and inserted into pMT/V5-His and pMT/BiP/V5-His vectors. These vectors are designed for transient and inducible protein expression in *Drosophila* cells (from Invitrogen). When these expression vectors are cotransfected with the selection vector (pCoHygro. pCoBlast or pCoPure) allow selection of stable cell lines with the appropriate antibiotic. In this case, expression of the protein of interest is achieved by adding the inductor (with a *Drosophila* metallothionein (MT) promoter) (Bunch et al., 1988). The protein will be secreted because of the *Drosophila* BiP secretion signal (Kirkpatrick et al., 1995) in the case of the pMT/BiP/V5-His vectors; and intracellular in the case of pMT/V5-His vectors. It also encodes an ampicillin resistance gene for selection of transformants in *E. coli.*
BGH_rev	pMT/V5-His A and pMT/BiP/V5-His A vectors	TAGAAGGCACAGTCGAGG	Reverse primer to sequence the C-terminal end of the clonings performed and inserted into pMT/BiP/V5-His and pMT/V5-His vectors.

IE1_fwd	pIEx vector	TGGATATTGTTTCAGTTGCAAG	Forward primer to sequence the N-terminal end of the clonings performed and inserted into pIEx vector. This is vector is designed for transient protein expression in *Spodoptera*-derived insect cells. pIEx vector encodes the signal peptide of adipokinetic hormone (AKH) to allow secretion of the protein of interest, C-terminal Strep- and His-tags sequences for protein purification and ampicillin resistance gene for selection of transformants in *E.coli*.
AS-Stag_rev	pIEx vector	GTCCATGTGCTGGCGTTC	Reverse primer to sequence the C-terminal end of the clonings performed and inserted into pIEx vector.
pIEx5_rev	pIEx vector	TTATTAGTGATGGTGATG	Reverse primer to sequence the C-terminal end of the clonings performed and inserted into pIEx vector.
CMV_fwd	pCMV-Sport 6 vector	CGCAAATGGGCGGTAGGCGTG	Forward primer to sequence the N-terminal end of the clonings performed and inserted into pCMV-SPORT 6 vector. This is vector is designed for transient protein expression of c-DNA (containing Kozak and ATG) in mammalian cells (Invitrogen). pCMV-SPORT 6 vector encodes a cytomegalovirus (CMV) promoter and an ampicillin resistance gene for selection of transformants in *E.coli*.
M13_fwd	pCMV-Sport 6 vector	TGTAAAACGACGGCCAGT	Reverse primer to sequence the C-terminal end of the clonings performed and inserted into pCMV-SPORT 6 vector.
Further information and feature maps of the vectors in the section “List of expression vectors”.			

Table 9. Constructs, primers and plasmids of hSKI1 for expression in bacteria							
Plasmid name	**Protein**	**Parental DNA**	**Forward -primer ***	**Reverse-primer ***	**Protein sequence ****	**Tags *****	**Comments**
pSP9-hSKI1 (18-429)-H8	K^{18}-S^{429} hSKI1	Human c-DNA pSP9 vector	GGTTTCGCTACCGTAGCGCAG GCCAAGAAACATCTGGGCGAC AGACT	GTGGTGGTGGTGTCCTGATCC GCTCACTAACAAGGTGACAGC AC	**K^{18}-S^{429}** +GSG+**H_{8x}**	C-t H_{8x}	Prodomain and serine protease domains of hSKI1 (from K^{18} to S^{429}) in bacteria cells. The gene was inserted by restriction free cloning (between the OmpA signal peptide and the 8xHis tag) into the pSP9 vector.
pSP9-hSKI1 (18-473)-H8	K^{18}-Q^{473} hSKI1	Human c-DNA pSP9 vector	GGTTTCGCTACCGTAGCGCAG GCCAAGAAACATCTGGGCGAC AGACT	GTGGTGGTGGTGTCCTGATCC CTGATAGGCTCTGAGCAGATC G	**K^{18}-Q^{473}** +GSG+**H_{8x}**	C-t H_{8x}	Prodomain and subtilase domains of hSKI1 (from K^{18} to Q^{473}) in bacteria cells. As pSP9-hSKI1 (18-429)-H8 construct for the cloning.
pSP9-hSKI1 (18-998)-H8	K^{18}- V^{998} hSKI1	Human c-DNA pSP9 vector	GGTTTCGCTACCGTAGCGCAG GCCAAGAAACATCTGGGCGAC AGACT	GTGGTGGTGGTGTCCTGATCC CACCTCCTGGTTGTAGCGGCC AG	**K^{18}-V^{998}** +GSG+**H_{8x}**	C-t H_{8x}	BTMD (before transmembrane domain; from K^{18} to V^{998}) hSKI1 in bacteria cells. As pSP9-hSKI1 (18-429)-H8 for the cloning.
pCri9a-hSKI1 (18-429)-H6	K^{18}-S^{429} hSKI1	Human c-DNA pCri9a vector	GCCCAGCCGGCGATGGCCATG GGCAAGAAACATCTGGGCGAC AGACTGG	GTGGTGGTGGTGTCCTGATCC GCTCACTAACAAGGTGACAGC AC	MG+ **K^{18}-S^{429}** + GSG+**H_{6x}**	C-t H_{6x}	Prodomain and serine protease domains of hSKI1 (from K^{18} to S^{429}) in bacteria cells. The gene was inserted by restriction free cloning (between the periplasmic signal peptide and the 6xHis tag) into the pCri9a vector.
pCri9a-hSKI1 (18-473)-H6	K^{18}-Q^{473} hSKI1	Human c-DNA pCri9a vector	GCCCAGCCGGCGATGGCCATG GGCAAGAAACATCTGGGCGAC AGACTGG	GTGGTGGTGGTGTCCTGATCC CTGATAGGCTCTGAGCAGATC G	MG+ **K^{18}-Q^{473}** + GSG+**H_{6x}**	C-t H_{6x}	Prodomain and subtilase domains of hSKI1 (from K^{18} to Q^{473}) in bacteria cells. As pCri9a-hSKI1 (18-429)-H6 construct for the cloning.
pCri9a-hSKI1 (18-998)-H6	K^{18}- V^{998} hSKI1	Human c-DNA pCri9a vector	GCCCAGCCGGCGATGGCCATG GGCAAGAAACATCTGGGCGAC AGACTGG	GTGGTGGTGGTGTCCTGATCC CACCTCCTGGTTGTAGCGGCC AG	MG+ **K^{18}-V^{998}** + GSG+**H_{6x}**	C-t H_{6x}	BTMD (from K^{18} to V^{998}) hSKI1 in bacteria cells. As pCri9a-hSKI1 (18-429)-H6 construct for the cloning.
pCri4a-H6-TRX-TEV-hSKI1 (18-429)	K^{18}-S^{429} hSKI1	Human c-DNA pCri4a vector	GCCCAGCCGGCGATGGCCATG GGCAAGAAACATCTGGGCGAC AGACTGG	TTGCTCGAGTTAGCTCACTAAC AAGGTGACAGCAC	GAMG+ **K^{18}-S^{429}**	N-t H_{6x}+ TRX+TEV	Prodomain and serine protease domains of hSKI1 (from K^{18} to S^{429}) in bacteria cells. The gene was inserted by directional cloning between the NcoI and XhoI restriction sites.

pCri4a-H6-TRX-TEV--hSKI1 (18-473)	K^{18}-Q^{473} hSKI1	Human c-DNA pCri4a vector	GCCCAGCCGGCGATGGCCATGGGCAAGAAACATCTGGGCGACAGACTGG	TTGCTCGAGTTACTGATAGGCTCTGAGCAGATCG	GAMG+ **K^{18}-Q^{473}**	N-t H_{6x}+ TRX+TEV	Prodomain and subtilase domains of hSKI1 (from K^{18} to Q^{473}) in bacteria cells. As pCri4a-H6-TRX-TEV-hSKI1 (18-429) construct for the cloning.
pCri4a-H6-TRX-TEV--hSKI1 (18-998)	K^{18}- V^{998} hSKI1	Human c-DNA pCri4a vector	GCCCAGCCGGCGATGGCCATGGGCAAGAAACATCTGGGCGACAGACTGG	TTGCTCGAGTTACACCTCCTGGTTGTAGCGGCCAG	GAMG+ **K^{18}- V^{998}**	N-t H_{6x}+ TRX+TEV	BTMD (from K^{18} to V^{998}) hSKI1 in bacteria cells. As pCri4a-H6-TRX-TEV--hSKI1 (18-473) construct for the cloning.
pCri7a-hSKI1 (18-429)-H6	K^{18}-S^{429} hSKI1	Human c-DNA pCri7a vector	GCCCAGCCGGCGATGGCCATGGGCAAGAAACATCTGGGCGACAGACTGG	TTGCTCGAGGCTCACTAACAAGGTGACAGCAC	MG+ **K^{18}-S^{429}** +LE+**H_{6x}**	C-t H_{6x}	Prodomain and serine protease domains of hSKI1 (from K^{18} to S^{429}) in bacteria cells. The gene was inserted by directional cloning between the NcoI and XhoI restriction sites.
pCri7a-hSKI1 (18-473)-H6	K^{18}-Q^{473} hSKI1	Human c-DNA pCri7a vector	GCCCAGCCGGCGATGGCCATGGGCAAGAAACATCTGGGCGACAGACTGG	TTGCTCGAGCTGATAGGCTCTGAGCAGATCG	MG+ **K^{18}-Q^{473} +** LE+**H_{6x}**	C-t H_{6x}	Prodomain and subtilase domains of hSKI1 (from K^{18} to Q^{473}) in bacteria cells. As pCri7a-hSKI1 (18-429)-H6 construct for the cloning.
pCri7a-hSKI1 (18-998)-H6	K^{18}- V^{998} hSKI1	Human c-DNA pCri7a vector	GCCCAGCCGGCGATGGCCATGGGCAAGAAACATCTGGGCGACAGACTGG	TTGCTCGAGCACCTCCTGGTTGTAGCGGCCAG	MG+ **K^{18}- V^{998} +** LE+**H_{6x}**	C-t H_{6x}	BTMD (from K^{18} to V^{998}) hSKI1 in bacteria cells. As pCri7a-hSKI1 (18-429)-H6 construct for the cloning.
pCri7a-hSKI1 (188-477)-H6	A^{188}-S^{477}_ hSKI1	Human c-DNA pCri7a vector	GATGGCCATGGGCGCCATCCCGCGCCAGGTTGCCCAG	GTGGTGCTCGAGGCTGTTGAGGATCTGATAGGC	MG+ **A^{188}-S^{477} +** LE+**H_{6x}**	C-t H_{6x}	Long subtilase domain of hSKI1 (from A^{188} to S^{477}) in bacteria cells. The gene was inserted by directional cloning between the NcoI and XhoI restriction sites.
pCri7a-hSKI1 (210-477)-H6	A^{210}-S^{477}_ hSKI1	Human c-DNA pCri7a vector	GATGGCCATGGGCGCTAATGTAAGAGTTGCTG	GTGGTGCTCGAGGCTGTTGAGGATCTGATAGGC	MG+ **A^{210}-S^{477} +** LE+**H_{6x}**	C-t H_{6x}	Short subtilase domain of hSKI1 (from A^{210} to S^{477}) in bacteria cells. The gene was inserted by directional cloning between the NcoI and XhoI restriction sites.

* Restriction-site sequences and overhangs for restriction-free cloning are underlined.

** Peptide sequence of the expressed protein after fusion-tag removal. Amino acids derived from the construct are in bold. See also Fig. 9

*** Fused tags at the carboxy-terminus (C-t) or the amino-terminus (N-t). TEV, tobacco-etch virus peptidase; TRX, thioredoxin tag; BTMD, before transmembrane domain.

Table 10. Constructs, primers and plasmids of hSKI1 for expression in insect cells							
Plasmid name	**Protein**	**Parental DNA**	**Forward -primer ***	**Reverse-primer ***	**Protein sequence ****	**Tags *****	**Comments**
pMT-hSKI1 (1-429)-H8	M^{1}-S^{429} hSKI1	Human c-DNA pMT/V5-His vector	GTCCAGTGTGGTGGAATTCGGCATGGGAAAGCTTGTCAACATCTGGCT	GTGATGGTGATGATGGTGGTGACCGCTCACTAACAAGGTGACAGCAC	**M^{1}-S^{429}** +G+**H_{8x}**	C-t H_{8x}	Own signal peptide, prodomain and serine protease domains of hSKI1 (from M^{1} to S^{429}) in insect cells. The gene was inserted by restriction free cloning and in frame with the 8xHis tag into the pMT/BiP/V5-His vector for inducible expression in stably transfected cell lines.
pMT-hSKI1 (1-473)-H8	M^{1}-Q^{473} hSKI1	Human c-DNA pMT/V5-His vector	GTCCAGTGTGGTGGAATTCGGCATGGGAAAGCTTGTCAACATCTGGCT	GTGATGGTGATGATGGTGGTGACCCTGATAGGCTCTGAGCAGATCG	**M^{1}-Q^{473}** +G+**H_{8x}**	C-t H_{8x}	Own signal peptide, prodomain and subtilase domains of hSKI1 (from M^{1} to Q^{473}) in insect cells. As pMT-hSKI1 (1-429)-H8 construct for the cloning.
pMT-hSKI1 (1-998)-H8	M^{1}- V^{998} hSKI1	Human c-DNA pMT/V5-His vector	GTCCAGTGTGGTGGAATTCGGCATGGGAAAGCTTGTCAACATCTGGCT	GTGATGGTGATGATGGTGGTGACCCTGATAGGCTCTGAGCAGATCG	**M^{1}- V^{998}** +G+**H_{8x}**	C-t H_{8x}	BTMD (from M^{1} to V^{998}) construct of hSKI1 in insect cells with its own signal peptide. As pMT-hSKI1 (1-429)-H8 construct for the cloning.
pMT/GFP-hSKI1 (1-998) -H8	M^{1}- V^{998} hSKI1 +GFP	Human c-DNA pMT/GFP-His vector	CTCGATTTAAGAAGGAGATATACCATGGGAAAGCTTGTCAACATCTGG	GTTCTTCTCCTTTGCTAGCCATACCCACCTCCTGGTTGTAGCGGCCAG	**M^{1}- V^{998}** +G+GFP+**H_{8x}**	C-t GFP+ H_{8x}	BTMD (from M^{1} to V^{998}) construct of hSKI1 in insect cells with its own signal peptide. The gene was inserted by restriction free cloning and in frame with the GFP and 8xHis tag into the pMT/GFP-His vector for inducible expression in stably transfected cell lines. The GFP allows reporting protein expression.
pMT/BiP-hSKI1 (18-429)-H8	BiP SP + K^{18}-S^{429} hSKI1	Human c-DNA pMT/BiP/V5-His vector	GTTGGCCTCTCGCTCGGGAGAAAGAAACATCTGGGCGACAGACT	GTGATGGTGATGATGGTGGTGACCGCTCACTAACAAGGTGACAGCAC	R+ **K^{18}-S^{429}** +G+**H_{8x}**	C-t H_{8x}	Prodomain and serine protease domains of hSKI1 (from K^{18} to S^{429}) in insect cells. The gene was inserted by restriction free cloning (between the BiP signal peptide and the 8xHis tag) into the pMT/BiP/V5-His vector for inducible and secreted protein expression in stably transfected cell lines.
pMT/BiP-hSKI1 (18-473)-H8	BiP SP+ K^{18}-Q^{473} hSKI1	Human c-DNA pMT/BiP/V5-His vector	GTTGGCCTCTCGCTCGGGAGAAAGAAACATCTGGGCGACAGACT	GTGATGGTGATGATGGTGGTGACCCTGATAGGCTCTGAGCAGATCG	R+ **K^{18}-Q^{473}** +G+**H_{8x}**	C-t H_{8x}	Prodomain and subtilase domains of hSKI1 (from K^{18} to Q^{473}) in insect cells. As pMT/BiP-hSKI1 (18-429)-H8 construct for the cloning.

pMT/BiP - hSKI1 (18-998)-H8	BiP SP+ K^{18}- V^{998} hSKI1	Human c-DNA pMT/BiP/V5-His vector	GTTGGCCTCTCGCTCGGGAGAAGAAACATCTGGGCGACAGACT	GTGATGGTGATGATGGTGGTGACCCTGATAGGCTCTGAGCAGATCG	R+ **K^{18}- V^{998}** +G+**H_{8x}**	C-t H_{8x}	BTMD (from K^{18} to V^{998}) hSKI1 in insect cells. As pMT/BiP-hSKI1 (18-429)-H8 construct for the cloning.
pMT/GNBP3-hSKI1 (18-998)-H8	GNBP3 SP + K^{18}- V^{998} hSKI1	Human c-DNA pMT/ GNBP3-His vector	TTCCTGCTCCTTGGCGTTCAGGGGAAGAAACATCTGGGCGACAGACT	GTGATGGTGATGATGGTGATGACCCACCTCCTGGTTGTAGCGGCCAG	**K^{18}- V^{998}** +G+**H_{8x}**	C-t H_{8x}	BTMD (from K^{18} to V^{998}) hSKI1 in insect cells. The gene was inserted by restriction free cloning (between the GNBP3 signal peptide and the 8xHis tag) into the pMT/GNBP3-His vector for inducible and secreted protein expression in stably transfected cell lines.
pIE-hSKI1 (18-477)-H6	AKH+ K^{18}-S^{477} hSKI1	Human c-DNA pIEx vector	CAATAGGCCTCAAAGAAACATCTGGGCGACAGACTG	CAATACCGGTGCTGTTGAGGATCTGATAGGCTC	S+ **K^{18}- S^{477}** + **TG**+H_{6x}	C-t H_{6x}	Prodomain and subtilase domains of hSKI1 (from K^{18} to S^{477}) in insect cells. The gene was inserted by directional cloning (between *Stu*I and *Age*I) into the pIEx vector (transient expression) in frame with the AKH signal peptide sequence and the 6xHis tag.
pIE-hSKI1 (18-998)-H6	AKH SP + K^{18}- V^{998} hSKI1	Human c-DNA pIEx vector	CAATAGGCCTCAAAGAAACATCTGGGCGACAGACTG	CAATACCGGTCACCTCCTGGTTGTAGCGGCCAGG	S+ **K^{18}- V^{998}** + **TG**+H_{6x}	C-t H_{6x}	BTMD (from K^{18} to V^{998}) hSKI1 in insect cells. The gene was inserted by directional cloning (between *Stu*I and *Age*I) into the pIEx vector (transient expression) in frame with the AKH signal peptide sequence and the 6xHis tag.
pHroE2-hSKI1 (18-477)-H6	AKH SP + K^{18}-S^{477} hSKI1	pIE-hSKI1 (18-477)-H6 pHroE2 vector	GCGCCGGCCGATGTACAAGCTCACAGTCTTCC	GCGCGGTACCTTAATGGTGATGGTGATGATG	S+ **K^{18}- S^{477}** + **TG**+H_{6x}	C-t H_{6x}	Prodomain and subtilase domains of hSKI1 (from K^{18} to S^{477}) in insect cells. The gene was inserted by directional cloning (between *Eag*I and *Kpn*I) into the pHroE2 vector (Excellgene) for transient expression.
pHroE2-hSKI1 (18-998)-H6	AKH SP + K^{18}- V^{998} hSKI1	pIE-hSKI1 (18-998)-H6 pHroE2 vector	GCGCCGGCCGATGTACAAGCTCACAGTCTTCC	GCGCGGTACCTTAATGGTGATGGTGATGATG	S+ **K^{18}- V^{998}** + **TG**+H_{6x}	C-t H_{6x}	BTMD (from K^{18} to V^{998}) hSKI1 in insect cells. The gene was inserted by directional cloning (between *Eag*I and *Kpn*I) into the pHroE2 vector (Excellgene) for transient expression.

* Restriction-site sequences and overhangs for restriction-free cloning are underlined.

** Peptide sequence of the expressed protein after fusion-tag removal. Amino acids derived from the construct are in bold. See also Fig. 9

*** Fused tags at the carboxy-terminus (C-t) or the amino-terminus (N-t). BTMD, before transmembrane domain; GNBP3, Gram-negative bacteria-binding protein 3 (*Drosophila melanogaster*).

Table 11. Constructs, primers and plasmids of hSKI1 for expression in mammalian cells

Plasmid name	Protein	Parental DNA	Forward -primer *	Reverse-primer *	Protein sequence **	Tags ***	Comments
pXLG40-hSKI1 (1-477)-H6	M^{1}-S^{477} hSKI1	Human c-DNA pXLG40 vector	CAATACTAGTATGAAGCTT GTCAACATC	CAATTCTAGATTAATGGTGA TGGTGATGGTGGCTGTTGA GGATCTGAT	**M^{1}-S^{477} +H_{6x}**	C-t H_{6x}	Own signal peptide, prodomain and subtilase domains of hSKI1 (from M^{1} to S^{477}) in mammalian cells. The gene was inserted by directional cloning (between *Spe*I and *Xba*I) into the pXLG40 vector (Excellgene) for transient expression.
pXLG40-hSKI1 (1-998)-H6	M^{1}-V^{998} hSKI1	Human c-DNA pXLG40 vector	CAATACTAGTATGAAGCTT GTCAACATC	CAATTCTAGATTAATGGTGA TGGTGATGGTGCACCTCCT GGTTGTAGC	**M^{1}-V^{998} +H_{6x}**	C-t H_{6x}	BTMD (from M^{1} to V^{998}) construct of hSKI1 in mammalian cells with its own signal peptide. As pXLG40-hSKI1 (1-477)-H6 construct for the cloning.
pS6-hSKI1 (18-947)-H6	Ig κ leader+ K^{18}-W^{947} hSKI1	Human c-DNA pCMV-Sport6 vector	CTGGGTTCCAGGTTCCACT GGTGACAAGAAACATCTG GGCGACAGACTG	CGCCTAATGGTGATGGTGA TGGTGCCAAAGGTTACTGG GCGCCGTCTC	**K^{18}-W^{947} +H_{6x}**	C-t H_{6x}	BSS (from K^{18} to W^{947}) (before shedding site) construct of hSKI1 in mammalian cells. The gene was inserted by restriction-free cloning into the pCMV-Sport 6 vector in frame with the Ig κ leader sequence. Construct for transient protein expression.
pS6-hSKI1 (18-998)-H6	Ig κ leader+ K^{18}-V^{998} hSKI1	Human c-DNA pCMV-Sport6 vector	CTGGGTTCCAGGTTCCACT GGTGACAAGAAACATCTG GGCGACAGACTG	CGCCTAATGGTGATGGTGA TGGTGCACCTCCTGGTTGT AGCGGCCAGGC	**K^{18}-V^{998} +H_{6x}**	C-t H_{6x}	BTMD (from K^{18} to V^{998}) construct of hSKI1 in mammalian cells. The gene was inserted by restriction-free cloning into the pCMV-Sport 6 vector in frame with the Ig κ leader sequence. Construct for transient protein expression.

* Restriction-site sequences and overhangs for restriction-free cloning are underlined.

** Peptide sequence of the expressed protein after fusion-tag removal. Amino acids derived from the construct are in bold. See also Fig.**9**

*** Fused tags at the carboxy-terminus (C-t) or the amino-terminus (N-t). BSS, before shedding site; BTMD, before transmembrane domain.

3.2.2.3 Protein expression in bacteria

Different fragments of hSKI1 were cloned into pCri vectors, aiming for soluble protein expression (Fig. 46 and Table 9). In all these constructs, an 8x or 6x His-tag was linked to the protein sequence to assist in protein purification. In the case of pCri4a, a thioredoxin (TRX) tag and a TEV (tobacco-etch virus peptidase) cleavage site were fused to the protein to assist in protein folding and purification, respectively. In the case of pCri9a, a periplasmic signal peptide allowed for protein purification from periplasmic space. And for pSP9, the outer membrane protein A (OmpA) SP for extracellular secretion.

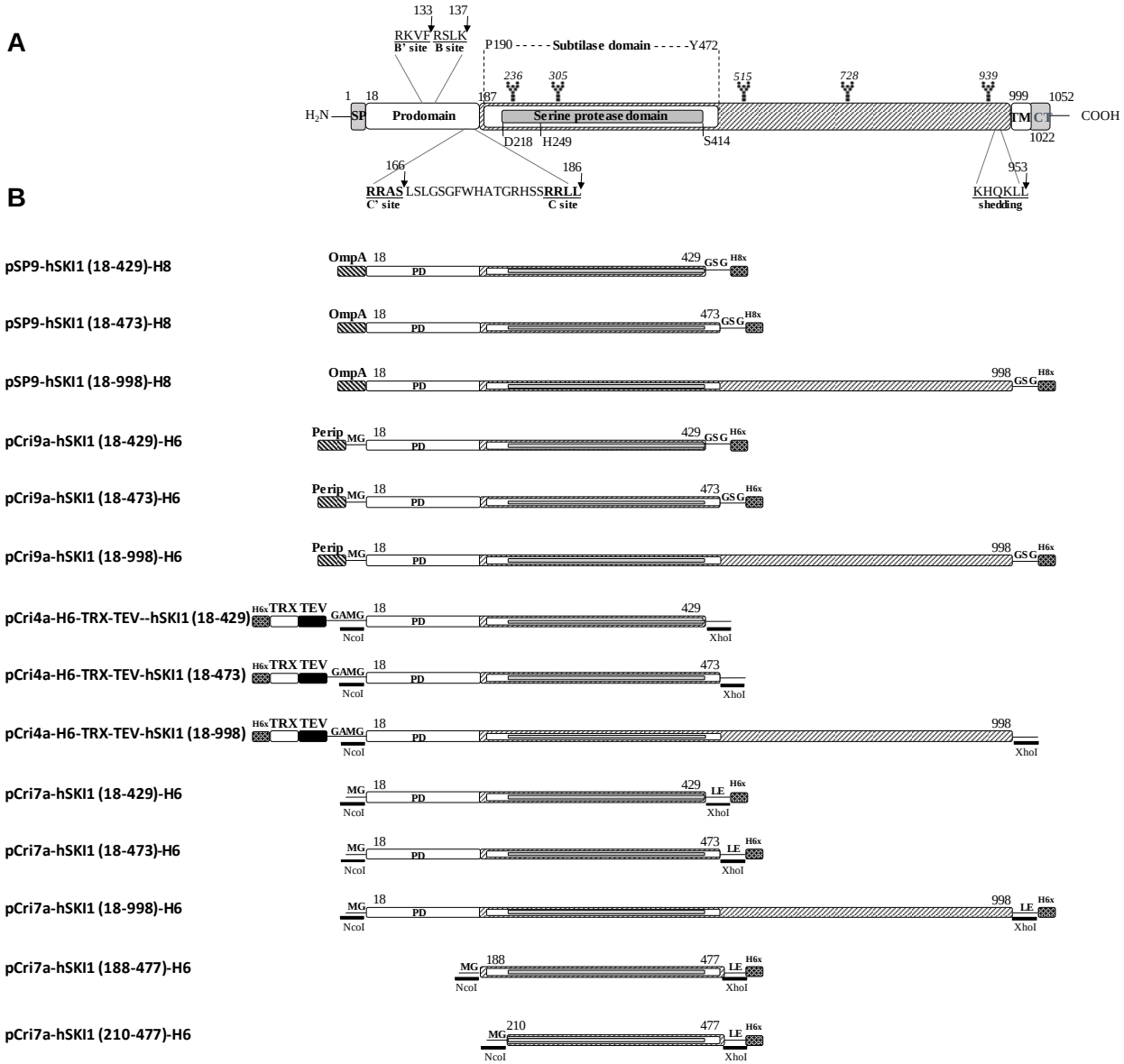


Figure 46. Overview of hSKI1 protein and fusion proteins produced for expression in bacteria. The residue numbers correspond to UP, entry Q14703. **(A)** Functional domains of zymogen SKI1 are the signal peptide (SP), prodomain (PD), transmembrane (TM) and cytosolic tail (CT). N-linked glycosylation sites are highlighted with a sugar chain. Cleavages and shedding motifs are shown in magnification and are underlined. The exact cleavage is also indicated with an arrow. **(B)** Human SKI1 fusion proteins produced with SP9 and pCri vectors. Signal peptides (OmpA and periplasmatic), TRX fusion protein, TEV site, His-tags and restriction sites are graphically represented. The length of each construct is also shown.

The successfully expressed constructs of hSKI1 were:

- pCri9a-hSKI1 (18-429)-H6 (Fig. 47)
- pCri7a-hSKI1 (18-429)-H6 (Fig. 48).
- pCri7a-hSKI1 (18-473)-H6 (Fig. 49).
- pCri7a-hSKI1 (18-998)-H6 (Fig. 50).
- pCri4a-hSKI1 (18-429)-H6 (Fig. 51).
- pCri4a-hSKI1 (18-473)-H6 (Fig. 51).
- pCri7a-hSKI1 (188-477)-H6 (Fig. 52).
- pCri7a-hSKI1 (210-477)-H6 (Fig. 52).

Expression from pCri9a construct (periplasmic space and refolding): In the case of pCri9a-hSKI1-H6 constructs, three fragments of hSKI1 were cloned (residue 18 to 429; residue 18 to 473 and residue 18 to 998) (Fig. 47A) and the only construct with detectable expression was the pCri9a-hSKI1 (18-429)-H6 (Fig. 47C). Afterwards, the protein identity was confirmed by peptide mass fingerprinting (result is not shown here).

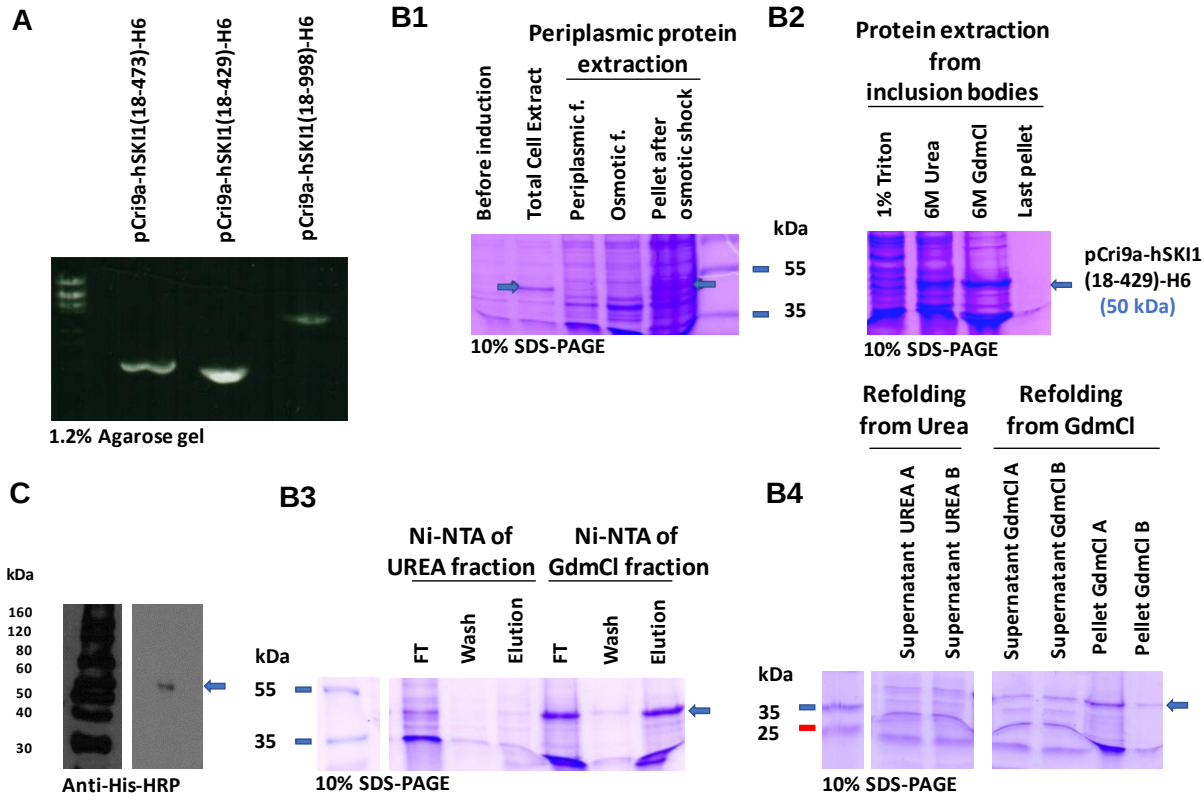


Figure 47. Periplasmic expression of hSKI1 (18-429) protein in BL21(DE3) cells with 1mM IPTG, O/N and at 20°C.. (A) Agarose gel electrophoresis analysis of different hSKI-1 fragments inserted into pCri9a. **(B1)** Protein expression induction of pCri9a-hSKI1 (18-429)-H6 construct in *E.coli* BL21(DE3) cells (before induction and after = total cell extract) and periplasmic protein extraction (periplasmic and osmotic fractions and final pellet). **(B2)** Protein extraction from inclusion bodies (sequential washes with EDTA, urea and GdmCl; protein purification protocol from inclusion bodies A). **(B3)** Affinity purification of the solubilized proteins in urea and GdmCl treatments (flow through (FT), wash and elution). **(B4)** Refolded proteins using dialysis membranes and two different buffers (refolding protocol A, second step, buffers B.A and B.B)). **(C)** Western blot analysis of the GdmCl elution using the monoclonal His-HRP Conjugated Antibody. Arrows indicate purified hSKI1 proteins.

After having analysed the resulting fractions from the periplasmic protein extraction protocol (Fig. 47B1), we concluded that most of the protein expressed from pCri9a construct and targeting a periplasmic secretion signal, was not in the periplasmic space, but remained in the insoluble fraction. The protein in the insoluble fraction was solubilized with 6M guanidinium chloride (GdmCl) after previous washes with Triton-EDTA and 6M urea (Fig. 47B2) and purified by Ni-NTA (Fig. 47B3) (inclusion bodies purification protocol A). Finally, we attempted the dialysis mediated refolding of the eluted proteins, using two different refolding buffers (Fig. 47B4) (refolding protocol A).

Using this experimental approach, most of the protein remained insoluble and precipitated into the dialysis membrane. No soluble protein was obtained to proceed with crystallization trials.

Expression from pCri7a construct in different strains: Regarding the pCri7a-hSKI1-H6 constructs, they were expressed in different bacterial strains, including BL21(DE3), Rosetta (which enhances the expression of eukaryotic proteins that contain codons rarely used in *E. coli*) and Origami 2 (with thioredoxin reductase and glutathione reductase genes, for disulphide bond formation in the cytoplasm), that allow for expression after IPTG induction. Despite that we found similar levels of expression of hSKI1 proteins in all three strains, the Rosetta strain was finally chosen because it seemed to get the cleanest insoluble fraction (Fig. 48A). Following that, different IPTG concentrations (1, 0.4 or 0.1 mM) and expression conditions (6h at 37°C or O/N at 20°C or 18°C) were also assayed. For all the cases, the expression of hSKI1 protein was still in inclusion bodies and refolding protocol (A) was unsuccessful.

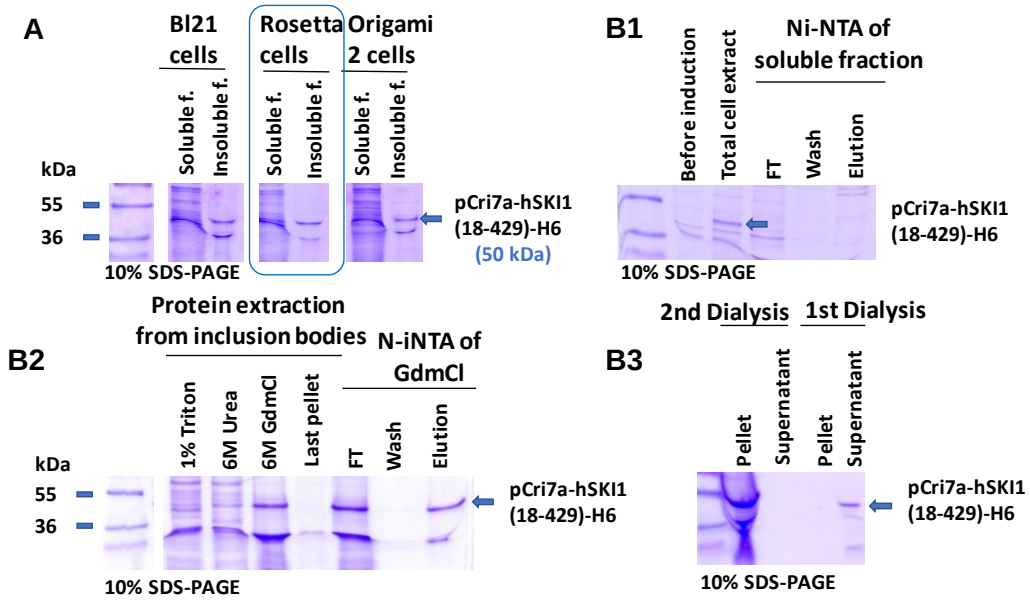


Figure 48. Overexpression of hSKI1 (18-429) protein in *E. coli* Rosetta cells with 1mM IPTG, O/N and at 20°C. (A) Several bacterial strains were tested at same expression conditions. **(B1)** Affinity purification of the soluble fraction. **(B2)** Protein extraction from inclusion bodies and Ni-NTA of the GdmCl fraction. **(B3)** Refolded protein using a two-step refolding dialysis. Arrows and square indicate purified hSKI1 proteins and the selected bacterial strain for expression, respectively.

Multi-step refolding assay for pCri7a constructs:

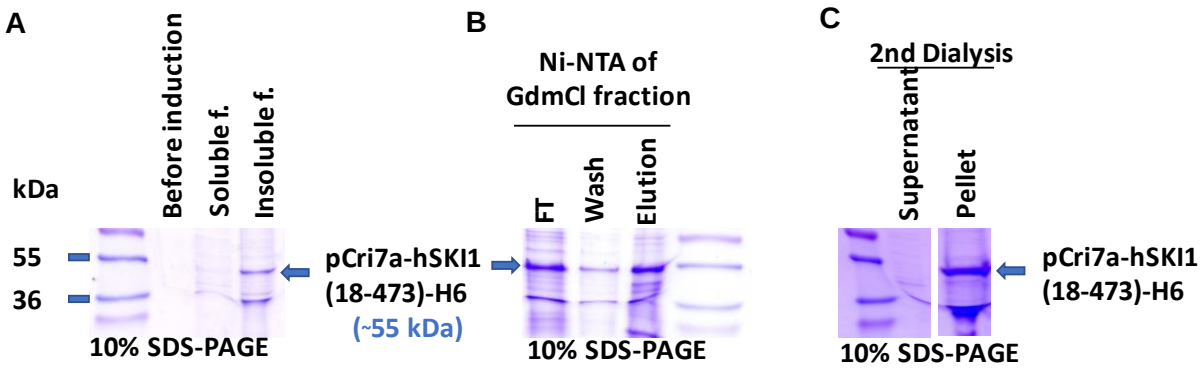


Figure 49. Overexpression of hSKI1 (18-473) protein in *E. coli* Rosetta cells with 1mM IPTG, O/N and at 20°C. (A) Protein expression induction (before induction and after = total cell extract, showing the two fractions separately, soluble and insoluble fractions). **(B)** Solubilization of the insoluble fraction by GdmCl treatment, followed by affinity purification. **(C)** Protein refolding by two-step dialysis protocol. Arrows indicate purified hSKI1 proteins

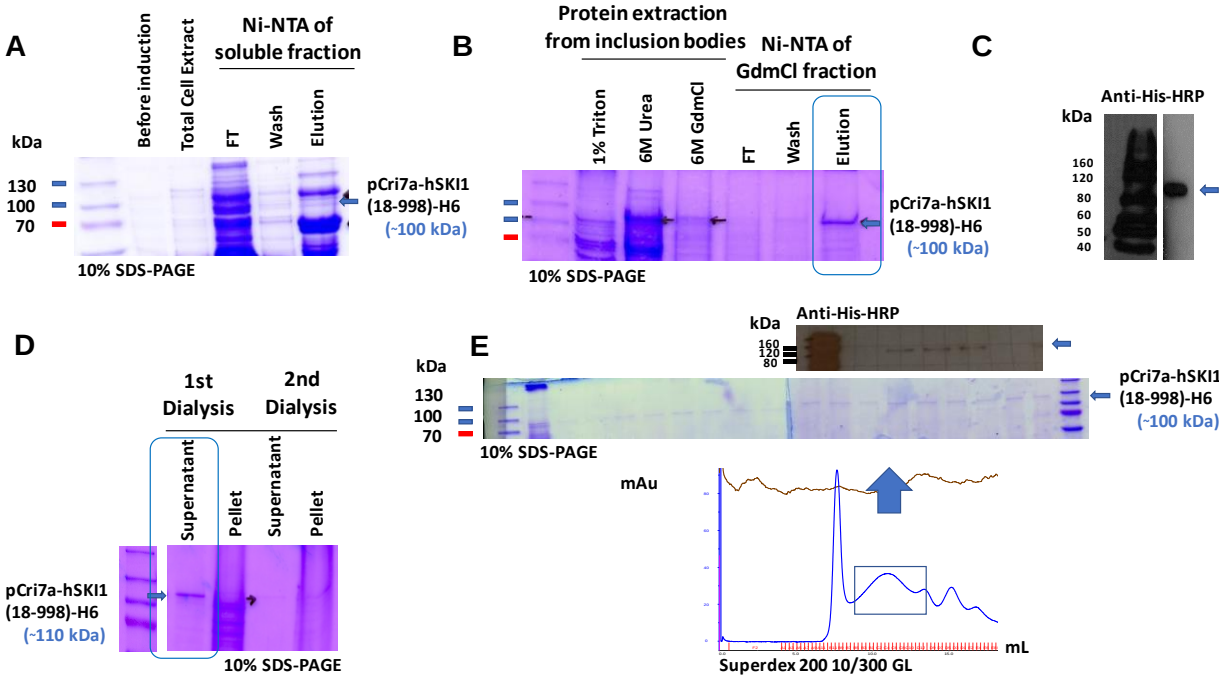


Figure 50. Overexpression of -hSKI1 (18-998) protein in *E. coli* Rosetta cells with 1mM IPTG, O/N and at 20°C. (A) Protein expression induction (before induction and after = total cell extract) and affinity purification of soluble fraction. **(B)** Solubilization of the insoluble fraction by triton, urea and GdmCl treatments, followed by affinity purification of the last treatment. **(C)** Validation of the GdmCl elution band by Western blot with an antibody against the His-tag**. (D)** Refolding by two-step dialysis protocol and selection of the first supernatant. **(E)** Size exclusion chromatography (Superdex 200 10/300) of the soluble protein from the first refolding step (L-Arg 0.5 M, but no GdmCl, present). Protein analysed and identified by 10% SDS-PAGE with Coomassie staining and Western blot against the His-tag. Arrows and square indicate purified hSKI1 proteins and selected samples, respectively.

In a successive refolding assay with two step dialysis (removing in the first step the GdmCl, and in a second the L-arginine), we observed that protein precipitation occurred mainly after removing the

L-arginine from the buffer (Fig. 48B3). Same behaviour was followed by hSKI1 (18-473) protein (Fig. 49) and hSKI1 (18-998) protein (Fig. 50). Therefore, it was decided to not remove L-arginine during the successive polishing steps. Thus, we performed a size exclusion chromatography, using the BTMD protein (hSKI1 (18-998))with L-arginine and glycerol in the running buffer (50mM Tris-HCl pH 8.0, 0.5 M L-arginine 5% glycerol) (Fig. 50). Best fractions of this final step of purification were pooled, concentrated up to 1.73 mg/mL and used for crystallization trials, despite that reaching supersaturation for initiating crystallization is difficult in the presence of these additives and a low protein concentration. As expected, protein crystals were never obtained. In this sense, considering protein behaviour during purification, it was certainly possible that the concentrated fractions were in fact partially unfolded SKI-1 soluble aggregates.

Expression with other tags and minimal constructs Regarding the results obtained up to this point, we tried the pCri4a vector with a thioredoxin tag and a TEV cleavage site to assist in protein folding and purification. However, hSKI1 proteins kept being expressed insoluble and solubilization failed to provide soluble folded protein (Fig. 51).

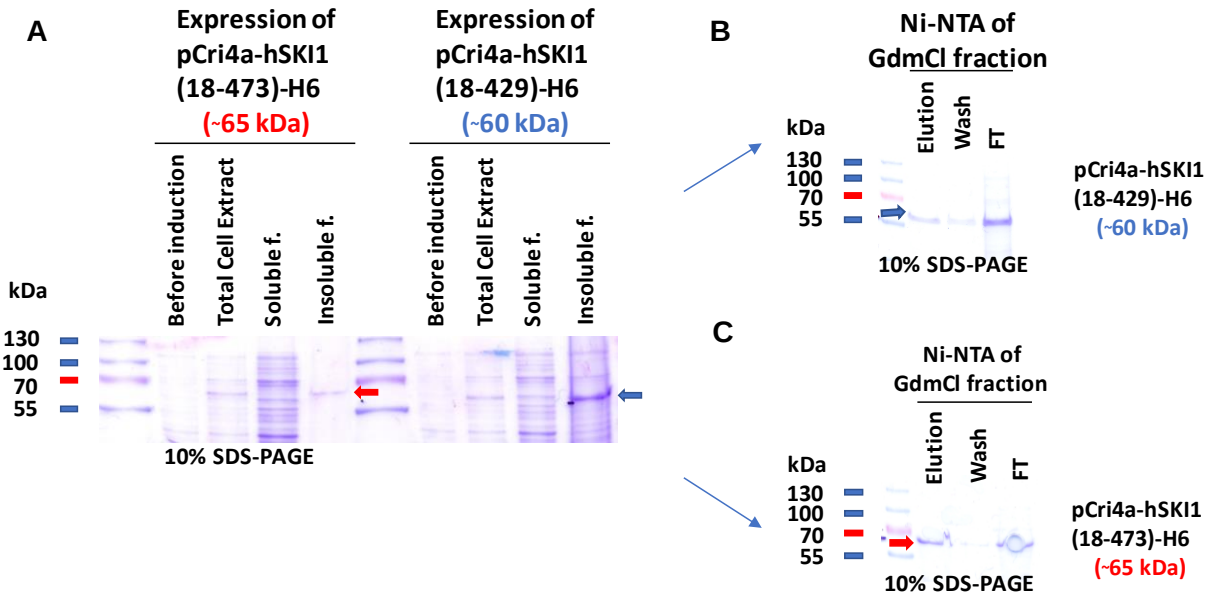


Figure 51 . Overexpression of hSKI1 (18-429) protein (in blue) and hSKI1 (18-473 protein(in red) in *E. coli* Origami 2 cells with 1mM IPTG, O/N and at 20°C. (A) Protein expression and purification (before induction and after = total cell extract, showing the two fractions separately, soluble and insoluble fractions). **(B)** For hSKI1 (18-429) protein: solubilization of the insoluble fraction by GdmCl treatment, followed by its affinity purification. **(C)** The same as B but for hSKI1 (18-473) protein. Arrows indicate purified hSKI1 proteins.

Expression and refolding of the minimal part of the protein (catalytic domain (residues from 188 to 477, and residues from 210 to 477) into pCri7a): Both constructs were expressed in inclusion bodies, but their purifications were slightly different (see the protocol B for protein purification from inclusion bodies) and protein refoldings was performed in column with a Sephadex G25 Pd10 (GE Healthcare) (refolding protocol C). In any case, the yields of the obtained refolded proteins were too little to continue with crystallization trials (Fig. 52).

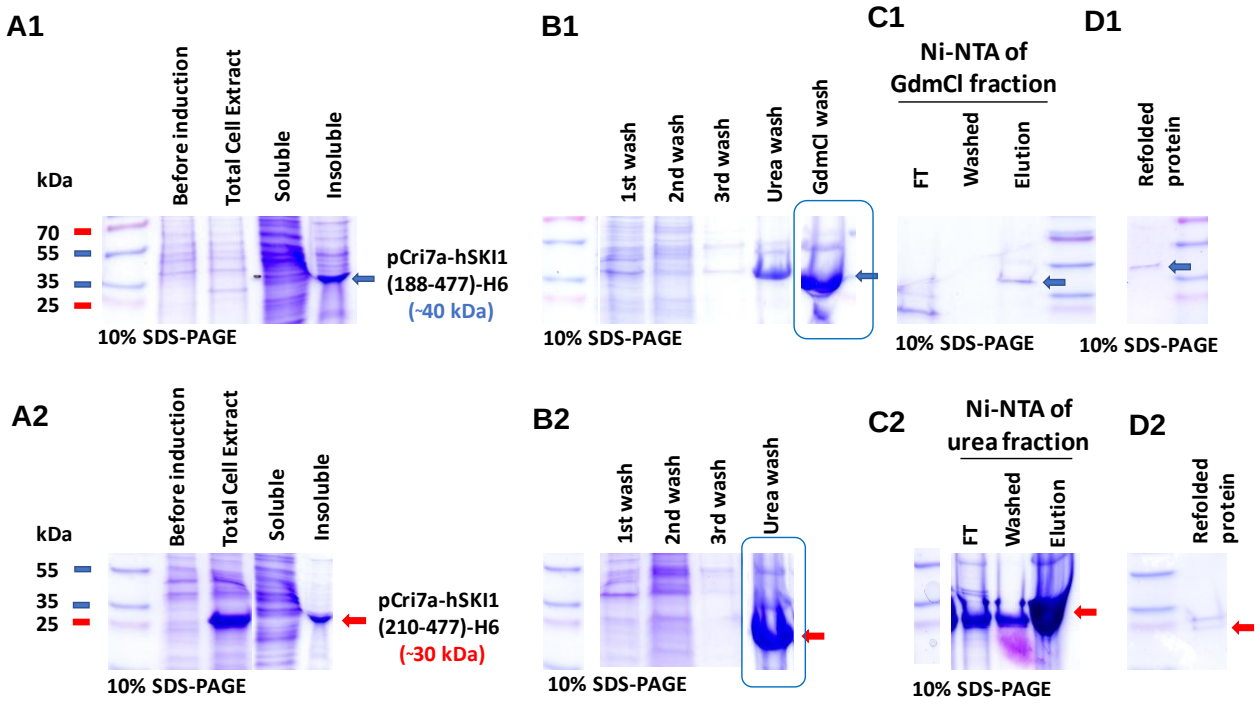


Figure 52. Overexpression of hSKI1 (188-477) protein (in blue) and -hSKI1 (210-477) protein (in red) in BL21(DE3) cells with 0.4mM IPTG, O/N and at 20°C. (A1, A2) Protein expressions induction by IPTG (before induction and after = total cell extract, showing separately, soluble and insoluble fractions). **(B1, B2)** Protein extractions of the insoluble fraction; **(C1, C2)** affinity purifications of the solubilized proteins and **(D1, D2)** refolded proteins (elutions) using a PD-10 desalting column. Arrows and square indicate purified hSKI1 proteins and selected samples, respectively.

In conclusion, despite of the diversity of followed strategies to get soluble protein, from all the successfully expressed constructs, proteins were expressed in inclusion bodies (insoluble protein), correctly solubilized but ultimately reluctant to complete refolding with different approaches. The final recovered proteins after this long protocol were not enough to proceed with the crystallization trials, therefore we proceed with the establishment of the insect protein expression systems.

3.2.2.4 Protein expression in insect cells

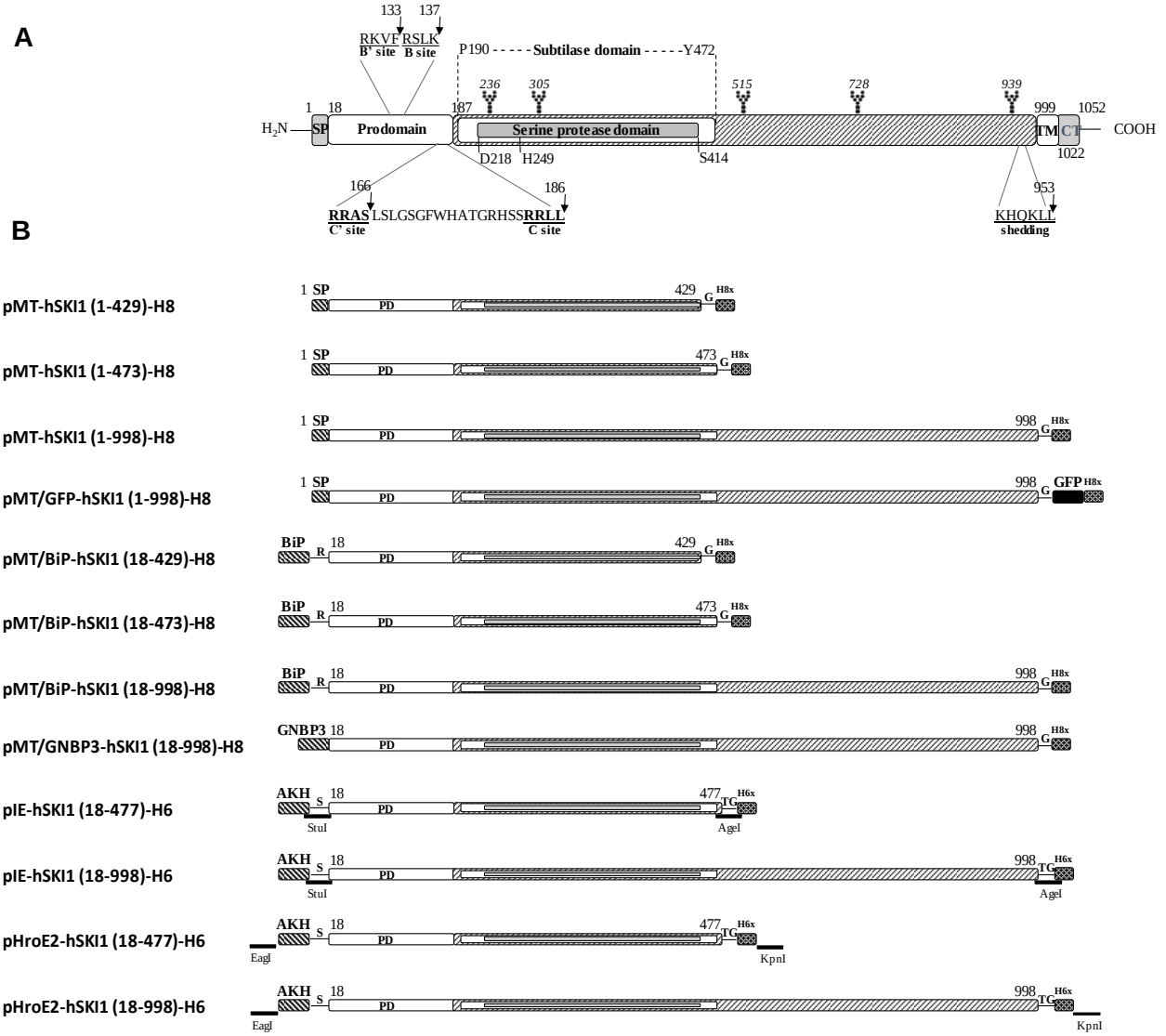


Figure 53. Overview of hSKI1 protein and fusion proteins produced for expression in insect cells. The residue numbers correspond to UP, entry Q14703. **(A)** Functional domains of zymogen SKI1 are the signal peptide (SP), prodomain (PD), transmembrane (TM) and cytosolic tail (CT). N-linked glycosylation sites are highlighted with a sugar chain. Cleavages and shedding motifs are shown in magnification and are underlined. The exact cleavage is also indicated with an arrow. **(B)** Human SKI1 proteins produced with its own signal peptide or signal peptides from vectors (BiP, GNBP3 and AKH signal peptides), in fusion with GFP protein, His-tags and restriction sites are graphically represented. The length of each construct is also shown.

Selected fragments of hSKI1 were also inserted into insect vectors for protein expression in *Drosophila* insect cells (Fig. 53 and Table 10). Here, we followed two strategies:

- Stably transfected cell lines: pMT/V5 and pMT/BiP/V5 vectors cotransfected with the selection vector (pCoPuro vector).
- Transient protein expression: pIEx and pHroE2 vectors.

Independently of the approach, an 8x or 6x His tag was linked to the protein sequence to facilitate protein purification. In the pMT/GFP-hSKI1 (1-998)-H8 construct, a GFP was also fused after the hSKI1 protein to report its expression by fluorescence and Western blot by using an antibody against the GFP.

About the first protein expression strategy, stably transfected cell lines allow for the selection of the successfully transfected cells and their amplification by antibiotic selection (puromycin resistance in the case of the pCoPuro vector). This strategy is very useful to increase the protein expression, when it is very poor in transient expression. Moreover, it is usually accompanied by cotransfection with an inducible vector, such as pMT and pMT/BiP vectors (with a *Drosophila* metallothionein (MT) promoter) (Bunch et al., 1988)., to control the protein expression by adding a metal inductor ($CuSO_4$ in this case).

We first generated six stably transfected cell lines a part from the control ones (GNBP3 and GFP proteins for extracellular and intracellular expressions, respectively), where three constructs had the original signal peptide from hSKI1 (pMT-hSKI1 (1-429)-H8, pMT-hSKI1 (1-473)-H8 and pMT-hSKI1 (1-998)-H8; but in the other three, the original signal peptide was replaced for an extracellular one (the *Drosophila* BiP (Kirkpatrick et al., 1995))(pMT/BiP-hSKI1 (18-429)-H8; pMT/BiP-hSKI1 (18-473)-H8 and pMT/BiP-hSKI1 (18-998)-H8. After their cotransfection with the puromycin selection vector and three weeks of selection with puromycin, from the eight cell lines (six for hSKI1 and two for the controls), the only detected expression signals came from the controls and the truncated full-length of hSKI1 with the extracellular signal peptide (construct pMT/BiP-hSKI1 (18-998)-H8 (Fig. 54). Subsequently, we set up the purification steps with the stably transfected control proteins (*Drosophila* Gram-negative bacteria-binding protein 3 or GNBP3 and Green fluorescent protein or GFP). However, the level of expression of the only expressed hSKI1 construct was so low than this construct was modified, and the signal peptide was replaced with the GNBP3 one (pMT/GNBP3-hSKI1 (18-998)-H8). In this case, the protein obtained from the stably transfected cell line was better expressed because of the GNBP3 signal peptide. The hSKI1 protein was detectable in the extracellular media and then purified by immobilized metal affinity chromatography (IMAC). Two bands were detected by 10% SDS-PAGE and

Western blot against the His-tag (~ 120 and 100 kDa) (Fig. 55). and both were identified by peptide mass fingerprinting and only the 100 kDa band by Edman degradation (recognised ALDS sequence).

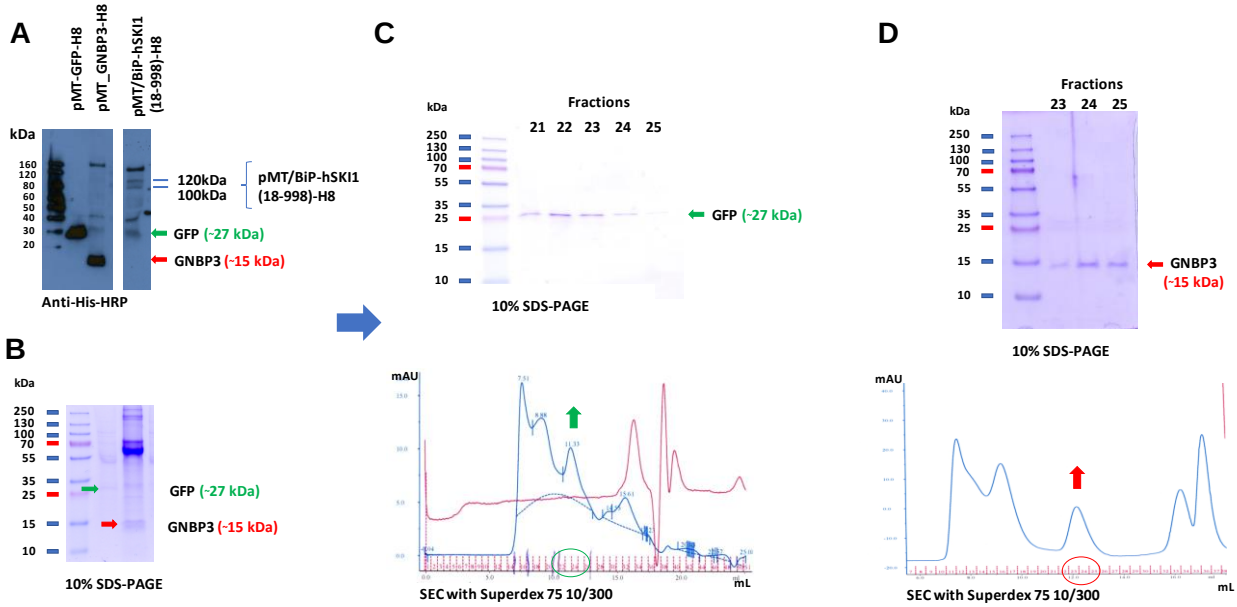


Figure 54. Overexpression of protein controls (*Drosophila* Gram-negative bacteria-binding protein 3 (GNBP3) and Green fluorescent protein (GFP)) of stably transfected cell lines and set up of the purification steps. (A) Western blot analysis of the control proteins (GNBP3 and GFP) and hSKI1 (18-998) protein against the His-tag. **(B)** 10% SDS-PAGE analysis of control proteins. **(C)** and **(D)** Size exclusion chromatographies with a Superdex 75 10/300 column for GFP **(C)** and GNBP3 **(D)** proteins. Proteins were analysed by 10% SDS-PAGE. Arrows and circles indicate purified proteins and analysed fractions, respectively.

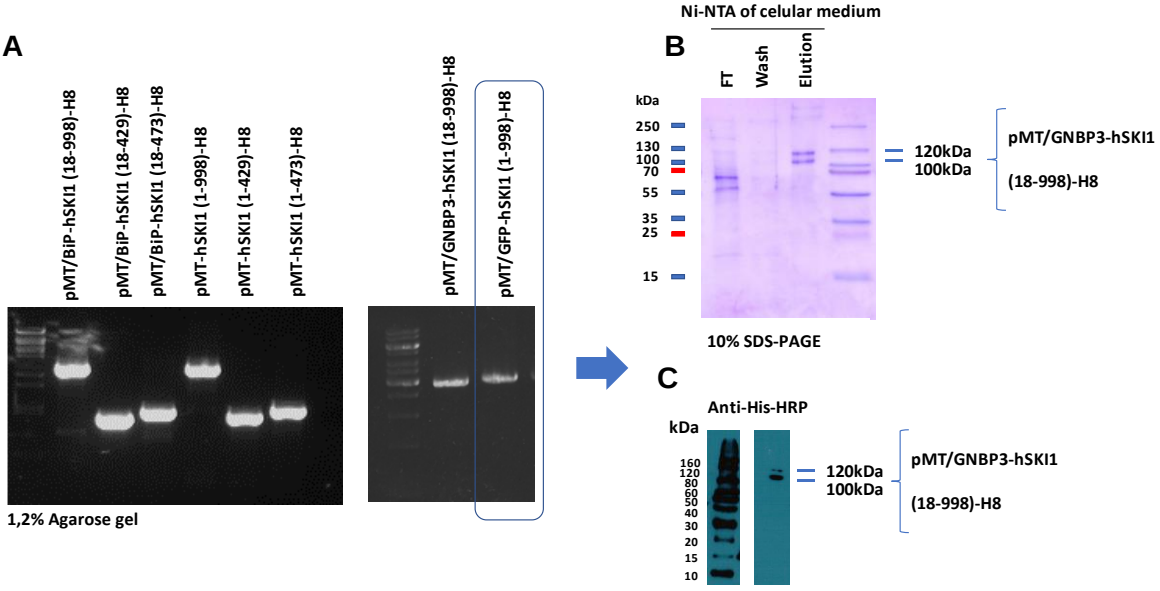


Figure 55. Overexpression of hSKI1 (18-998) protein in S2 cells. Analysis of the cloning by agarose gel electrophoresis **(A),** the protein expression by 10% SDS-PAGE **(B)** and Western blot against the His tag **(C).** Arrows and square indicate purified proteins and extracellularly expressed construct in stably transfected cell lines for hSKI1.

Two series of optimization trials were performed in order to increase the final protein yield and protein stability, related to transfection conditions and protein purification steps. About transfection conditions: the ideal induction with $CuSO_4$ was at cell density of ~$4x10^6$ cells/mL, the $CuSO_4$ concentration was from 500 to 1000 µM and from three harvesting days post-transfection (Fig. 56). About protein purification trials: several immobilized metal chelating resins were tested (nickel and cobalt (TALON)) for the affinity chromatography step, with the TALON giving the best recovery; and because protein precipitation and degradation into the column in the SEC step were observed, different detergents at several concentrations were also assayed The best behaviour of hSKI1 in SEC was observed with running buffers containing 0,5% Nonidet P40 (Igepal) or 0,5% octyl-β-glucoside (βOG)(Fig.58).

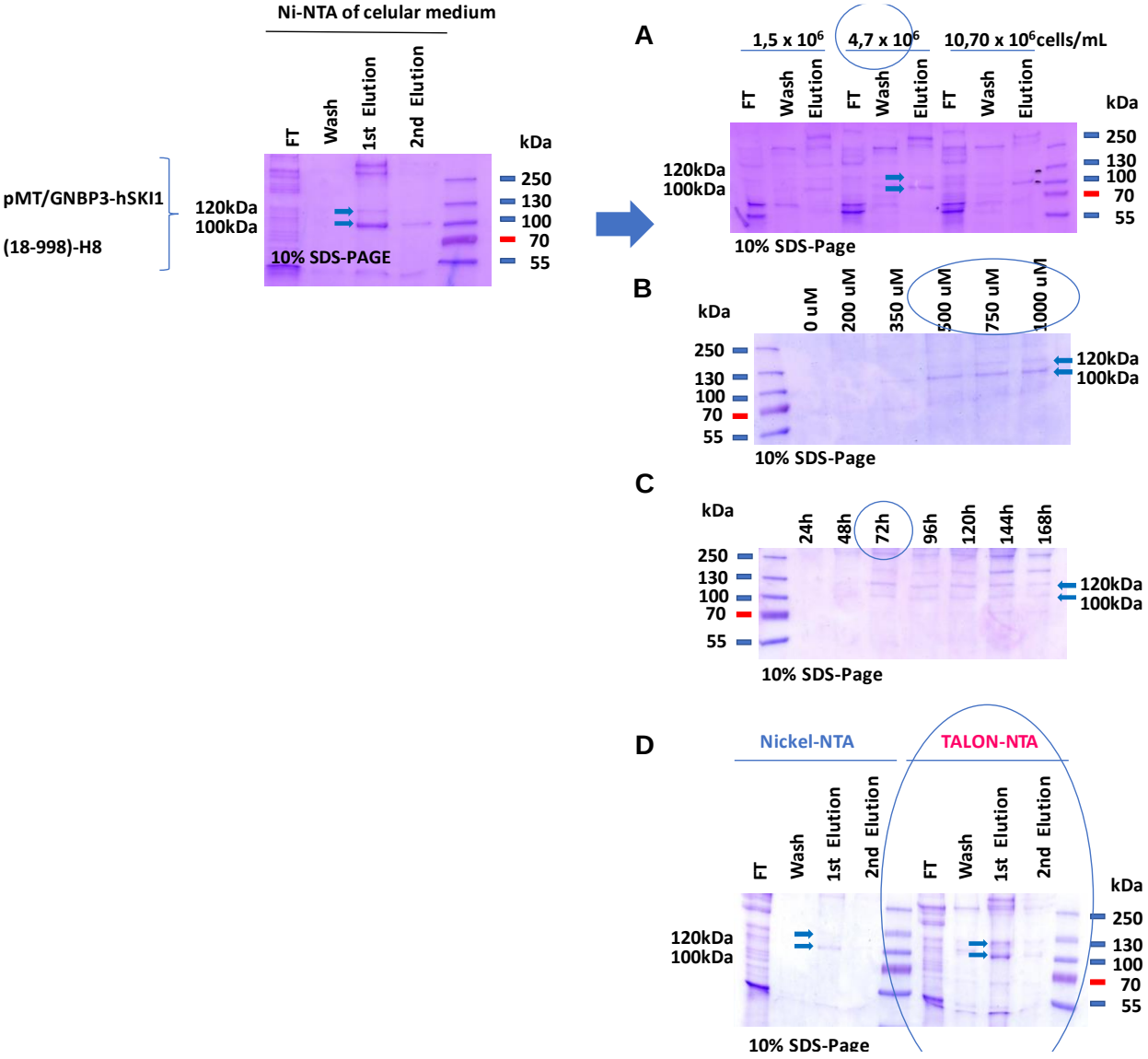


Figure 56. Optimization of transfection conditions and first purification step of hSKI1 (18-998) protein expressed in S2 cells. Establishment of optimal transfection conditions: **(A)** cell density for induction**, (B)** CuSO4 concentration for protein induction**, (C)** harvesting days to collect the protein and **(D)** metal affinity resin for protein purification**.** Arrows and circles indicate purified proteins and best conditions, respectively.

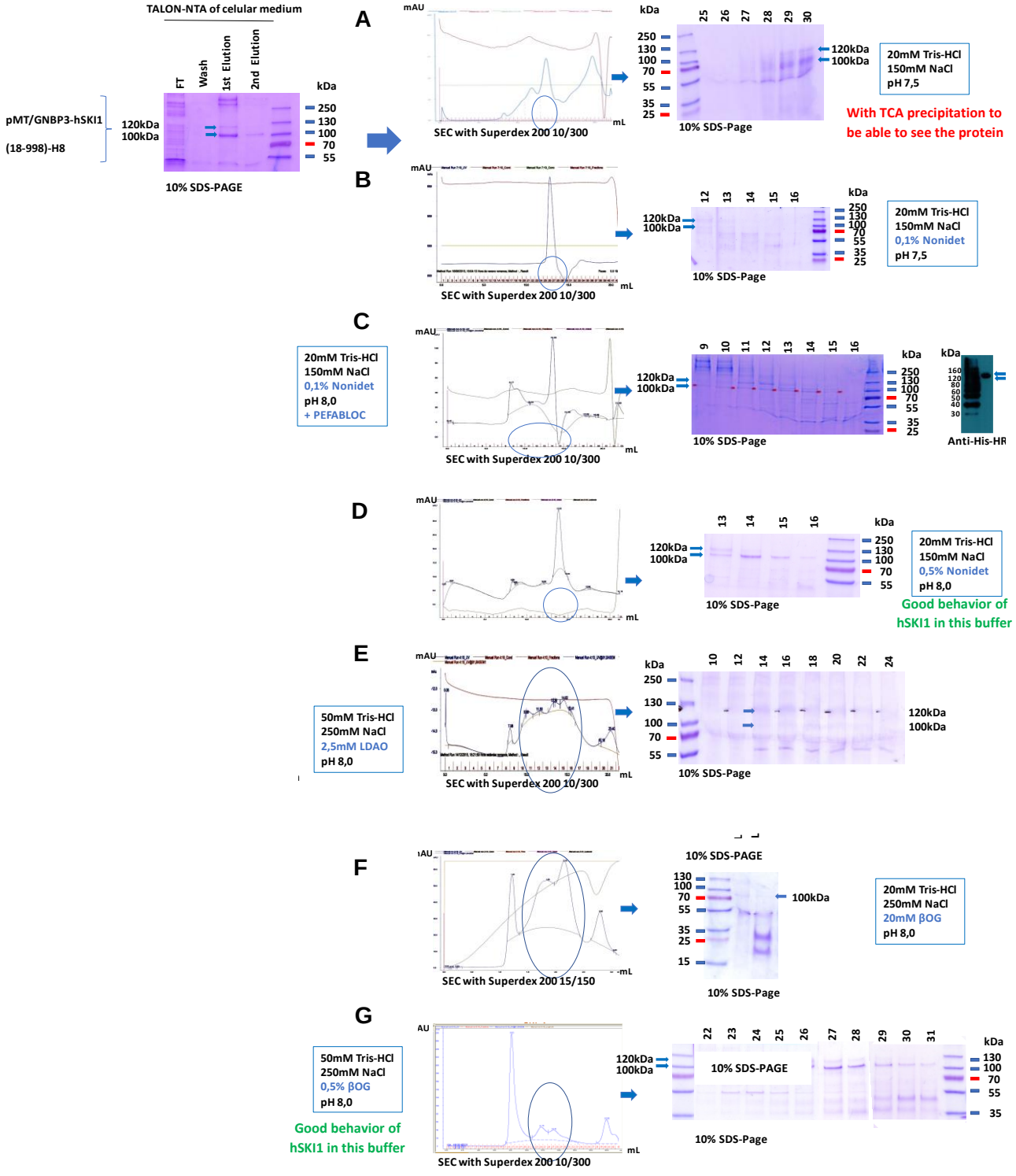


Figure 57. Optimization of size exclusion chromatography step of -hSKI1 (18-998) protein expressed in S2 cells. Trials with detergents were assayed to avoid protein precipitation into the column and denaturalization of the protein: **(A)** buffer without detergents, **(B)** 0,1% Nonidet-P40, **(C)** 0,1% Nonidet-P40 and Pefabloc, **(D)** 0,5% Nonidet-P40, **(E)** 2,5mM LDAO, **(F)** 20mM βOG and **(G)** 0,5% βOG. Results are shown by SEC and SDS-PAGE analysis of the corresponding peaks and fractions where hSHKI1 elutes. Arrows and circles indicate purified proteins and analysed fractions, respectively.

Another construct with a fused GFP was engineered (pMT/GFP-hSKI1 (1-998)-H8), but the protein expression was undetectable against His and GFP-tags.

A second strategy was performed with two different vectors for transient expression, pIEx and pHroE2 and two hSKI1 fragments (catalytic domain, residues 18 to 477; and truncated full-length, residues 18 to 998). In both cases, the AKH (Adipokinetic hormone) signal peptide was used for extracellular protein expression and in both cases, expression was detected by SDS-PAGE and Western blot against His tag.

3.2.2.5 Protein expression in mammalian cells

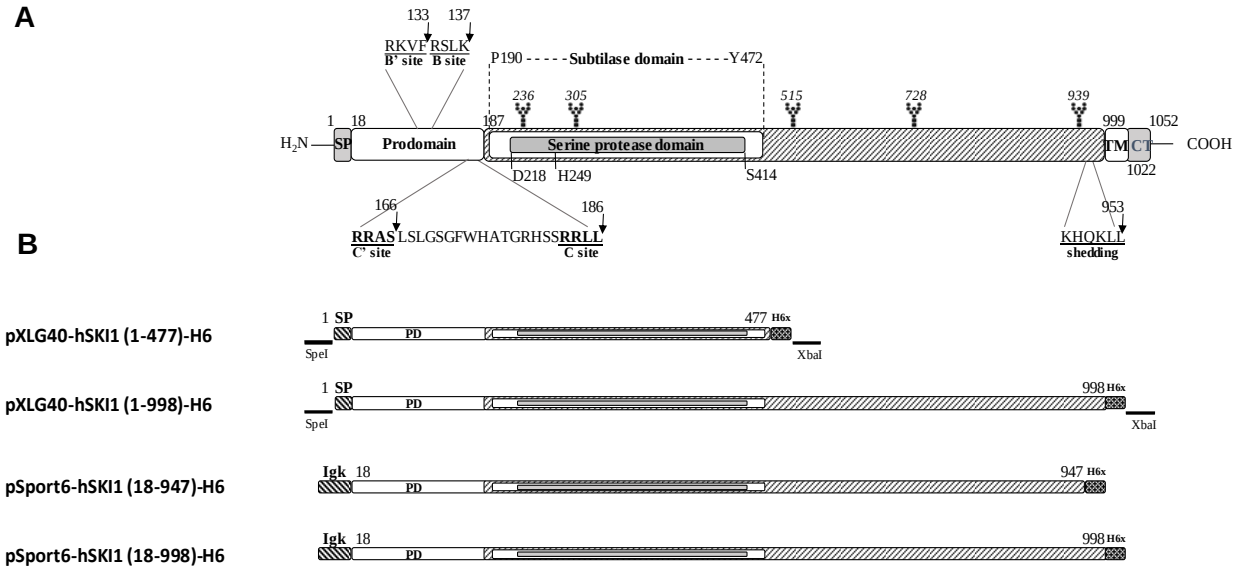


Figure 58. Overview of hSKI1 protein and fusion proteins produced for expression in mammalian cells. The residue numbers correspond to UP, entry Q14703. **(A)** Functional domains of zymogen SKI1 are the signal peptide (SP), prodomain (PD), transmembrane (TM) and cytosolic tail (CT). N-linked glycosylation sites are highlighted with a sugar chain. Cleavages and shedding motifs are shown in magnification and are underlined. The exact cleavage is also indicated with an arrow. **(B)** Human SKI1 fusion proteins produced with its own signal peptide and with the mouse Ig κ-chain ladder sequence (secreted signal peptide), His-tags and restriction sites are graphically represented. The length of each construct is also shown.

A last effort was performed to improve the protein yield of hSKI1 obtained with insect cells (1.12 mg/L cell culture after the TALON affinity purification; expressed in S2 cells, but with many contaminants) and its behavior in protein purification. Four constructs were generated to be expressed in mammalian cells (CHO and Expi293F cells) (Fig. 58 and Table 11):

- Two constructs with the endogenous signal peptide of hSKI1 with two different C-terminal ends (see Fig. 9):
  - pXLG40-hSKI1 (1-477)-H6: Ending at the residue 477 for the expression of the catalytic domain of hSKI1.
  - pXLG40-hSKI1 (1-998)-H6: Ending at the residue 998 for the expression and secretion of the truncated (before the transmembrane domain) full-length hSKI1.
- And two constructs with a signal peptide (Ig κ) for secretion and two different C-terminal ends:
  - pSport6-hSKI1 (18-998)-H6: Ending at the residue 998 for the expression of the full-length hSKI1, truncated before the transmembrane domain (BTMD construct).
  - pSport6-hSKI1 (18-947)-H6: Ending at the residue 947 for the expression of the full-length hSKI1, truncated before the transmembrane domain and as well, before the shedding site (BSS construct).

Two transient mammalian vectors were used (pXLG40 and pSport6), originally with very similar features, but later we modified the pSport6 to introduce the Ig κ signal peptide for secreted expression.

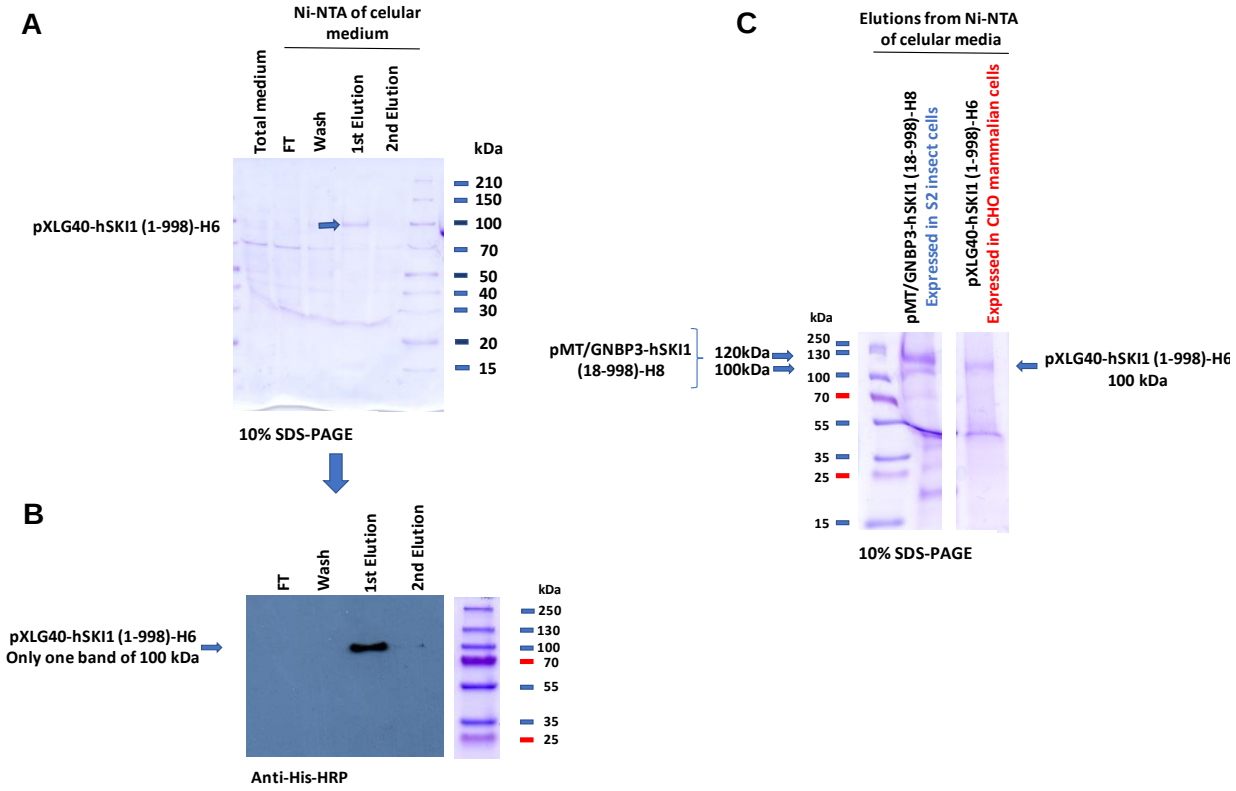


Figure 59. Overexpression of hSKI1 (1-998) protein in Expi293F™ cells. (A) Analysis of the protein expression after Ni-NTA purification by 10% SDS-PAGE**. (B)** Validation of the protein expression by Western blot against the His tag. **(C)** Comparation of hSKI1 full-length proteins expressed in insect and mammalian expression systems. Arrows indicate purified proteins.

With mammalian cells, the three truncated full-length constructs of hSKI1 were expressed and detected by SDS-PAGE and Western blot analysis in extracellular medium, independently of the used signal peptide. Purified proteins retained less contaminants, with better protein yield (1 mg/L cell culture after the Ni-NTA) and with a unique ~100 kDa band in the case of the construct with the endogenous signal peptide of hSKI1 (pXLG40-hSKI1 (1-998)-H6). All these data indicated that this protein was properly processed and it probably corresponded to the soluble form of hSKI1, ending in the residue 998 (because without the transmembrane part, the shedding is not possible) (Elagoz et al., 2001) (Fig. 59).

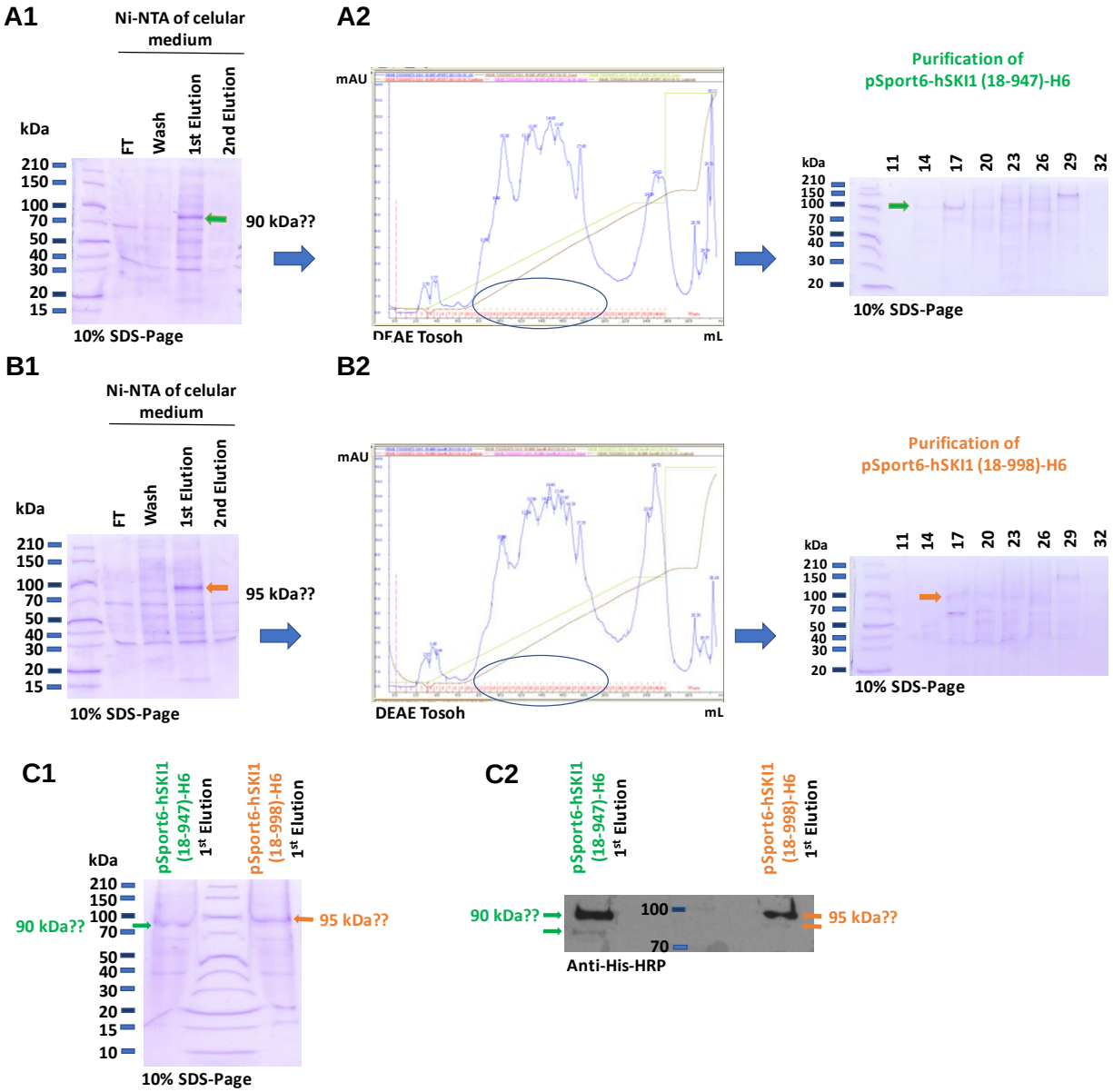


Figure 60. Overexpression of hSKI1 (18-947) and-hSKI1 (18-998) proteins in Expi293F™ cells. (A1) Analysis of the protein expression of pSport6-hSKI1 (18-947)-H6 construct after Ni-NTA purification by 10% SDS-PAGE**. (A2).** Purification by anion exchange chromatography and its analysis by SDS-PAGE. **(B1) and (B2)** The same for pSport6-hSKI1 (18-998)-H6 construct. **(C1)** and **(C2)** comparation of first elutions of these constructs after the Ni-NTA and analysis by 10% SDS-PAGE and Western blot against the His tag. Arrows and circles indicate purified proteins and analysed fractions, respectively.

In the case of the full-length hSKI1 proteins expressed from residue 18 and extracellular signal peptide (pSport6-hSKI1 (18-947)-H6 and pSport6-hSKI1 (18-998)-H6 constructs), the protein yields were 1.8 and 1.65 mg/ L cell culture after the Ni-NTA, respectively. However, these protein yields were not accurate because of the high number of contaminants eluted together with hSKI1. Even when we performed an anion exchange chromatography, the full-length proteins were not properly separated from their contaminants, and as a result, almost all the protein was lost in the process. They had a MW around 100 kDa, but another lower band could be detected by Western blot against the His-tag (Fig. 60). We therefore concluded that all these contaminants could be forming aggregates with hSKI1. These two proteins were being forced to be expressed extracellularly, when it is well known (da Palma et al., 2014), that the hSKI1 processing takes places in organelles (first RE and then, Golgi) and that all the species of hSKI1 travel together, activating each other and also, carrying the prodomain pieces (see Fig. 9). If this very complex process is moved to the extracellular medium, it will probably become even more complicated without the proper membrane signals for processing, activation and secretion of the full-active secreted form (residues 187 to 953) of hSKI1.

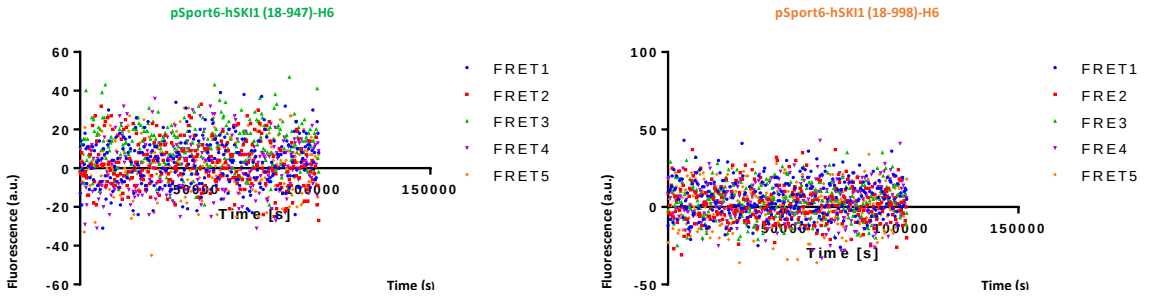


Figure 61. Kinetics of the proteolysis of fluorogenic substrates (FRETS 1, 2, 3, 4 and 5) for hSKI1(18-947) and hSKI1(18-998) purified proteins. (A) For pSport6-hSKI1 (18-947)-H6 purified protein and **(B)** for pSport6-hSKI1 (18-998)-H6 purified protein.

Apart from analysing the expression of these proteins (pSport6-hSKI1 (18-947)-H6 and pSport6-hSKI1 (18-998)-H6 constructs), we also tested their proteolytic activity on five fluorogenic peptide substrates of sequence: Abz-Lys-Asp-Glu-Ser-Tyr-Arg-K(dnp) (FRET1), Abz-Thr-Val-Leu-Glu-Arg-Ser-K(dnp) (FRET2), Abz-Asp-Tyr-Val-Ala-Ser-Glu-K(dnp) (FRET3), Abz-Tyr-Gly-Lys-Arg-Val-Phe-K(dnp) (FRET4) and Abz-Val-Lys-Phe-Tyr-Asp-Ile-K(dnp) (FRET5) (Goulas et al., 2011) (where Abz is ortho-aminobenzoic acid and K(Dnp) is Ne-2,4-dinitrophenyl-lysine with free carboxyl group) (Cotrin et al., 2004). These reactions were performed per triplicate as described in section 2.7.1. No proteolytic activity against these protein substrates was detected for both proteins (Fig. 61). This result could be explained because these substrates did not have the proper recognition sequence for hSKI1 or because, these proteins were simply not active (they were not the full active protein of hSKI1, which is secreted after being processed and cleaved the full prodomain (up to residue 186) and shedded (at residue 953), resulting a protein of ~98 kDa)

3.2.3 CONCLUSIONS AND FUTURE PERSPECTIVES

When we initiated this project, we aspired to achieve two main goals:

- The expression of this complex protein, which is first synthesized as a zymogen that is progressively processed in different versions of immature proteins in the organelles until getting the mature and full-active protein, which is secreted. This protein is a transmembrane protein, very glycosylated and naturally placed into organelles (ER and Golgi) and only a part is secreted under normal conditions (Pasquato et al., 2014).
- To establish a good purification protocol to obtain enough pure protein to be able to characterize it biochemically and biophysically, and finally, structurally by the obtention of protein crystals and atomic diffraction.

To accomplish these objectives, we generated and expressed different fragments of hSKI1 (catalytic domain constructs of different lengths and the BTMD) in three protein expression systems (bacteria, insect and mammalian cells). In the case of bacteria, several constructs were successfully expressed, but they were unfortunately expressed intracellularly in inclusion bodies in all the cases. Bacteria did not have the necessary machinery to fold this complicated protein. That happened independently of the signal peptide, tag, bacterial strain or culture conditions used for their expression.

Despite of being expressed in inclusion bodies, we struggled to purify and fold the insoluble proteins with several protocols for extraction and renaturation. Although we recovered high amounts of protein, these proteins were not properly folded, so 0.5 M L-arginine was necessary to keep these soluble aggregates in solution. We set up crystallization plates with this additive in the buffer, but we never obtained protein crystals.

We decided to increase the complexity of the protein expression system and test the expression in *Drosophila* S2 cells. In this scenario, we generated numerous constructs (catalytic domain constructs of different lengths and the BTMD) for transient and inducible expression (in selected stably transfected cell lines), with the endogenous signal peptide of hSKI1 and also replaced by a secretion signal. hSKI1 was not detectable in transient transfection for any fragment of hSKI1. Nevertheless, a positive expression in media (so, soluble protein) was detected from one of these stably transfected cell lines, whose positive clones had been selected and amplified. That was for the construct with the BiP secretion signal and the BTMD fragment (pMT/BiP-hSKI1 (18-998)-H8 construct). This meant that the expression was so low that it needed to be amplified to be detected and maybe the luminal part of hSKI1 was also required for folding and protein stability (because none of the catalytic fragments were expressed with the same strategy).

Moreover, we replaced the successful signal peptide for the GNBP3 one, and optimized its expression and purification conditions. This construct (pMT/GNBP3-hSKI1 (18-998)-H8) was secreted to the extracellular medium with two main bands of approximately 120 and 100 kDa. The protein identity of these two bands was confirmed by peptide mass fingerprinting and Western blot against the His-tag. However, the N-terminal Edman degradation could only recognize the 100 kDa band (because the amino terminal part of the 120 kDa band was very degraded) and the processing of this protein was being detected before the B'/B cleavages. This data suggested that both proteins could correspond to pro-hSKI1 (immature forms and totally inactive).

The protein yield of this construct (pMT/GNBP3-hSKI1 (18-998)-H8) was 1.12 mg/L cell culture after the TALON, but considering the many contaminants present, we estimated the real protein yield was around 0.4 mg/L cell culture (in agreement with the already published protein yields of hSKI1 in insect cells) (Bodvard et al., 2006). We tried to get rid of these contaminants by using different affinity resins and chromatography columns (for size and anion exchange chromatographies), but without yielding good results. In addition, the eluted proteins from IMAC needed a detergent to be stable and not precipitate into the chromatography columns. This also indicated that these proteins were not well folded (and maybe answering why their prodomains had not been properly cleaved as the Edman degradation analysis suggested).

Therefore, we moved to another eukaryotic expression system to skip all these inconveniences. For the expression in mammalian cells, we generated four constructs: one for the catalytic domain fragment and three for the BTMD expression. Regarding the BTMD constructs (truncated before the transmembrane domain), two were ending in residue 998 (but with different signal peptides: the endogenous signal for intracellular expression and a secretion signal) and one ending in residue 947, which is before the shedding motif. Fortunately, this expression system was successful in obtaining soluble BTMD protein secreted to cellular medium. The catalytic domain fragment was once again not expressed, validating our theory about the luminal domain importance for folding and protein stability.

When analysing the BTMD expressions in supernatants, we mainly detected differences in protein yield and purity. The only expressed construct with the endogenous signal peptide (pXLG40-hSKI1 (1-998)-H6) behaved with only one band about 100 kDa and with a 1 mg/L cell culture of very pure protein, which is the highest yield that has been published to date (0.4-0.8 mg/L cell culture by BEVS in Sf9 insect cells) (Bodvard et al., 2006) . This protein must be the secreted hSKI1, as it was naturally being secreted to the medium, but ending in the residue 998 instead of 953 (because we could detect its C-terminal His-tag by Western blot). This was expected, as it has been published by Elagoz et al. in 2001, that for the shedding to happen, the transmembrane domain is needed. It was also published

by da Palma et al. in 2014 that a proper cleavage of the prodomain is crucial for trafficking and for obtaining this mature protein with full activity (which is the final step of this complex and progressive processing of the pro-hSKI1). In conclusion, the only way to obtain a proper folded hSKI1 protein secreted to the medium would be by expressing the full-length or the truncated form before the transmembrane domain with its endogenous signal peptide.

The other BTMD constructs (ending at 947 and 998 positions) with the secretion signal were also expressed into the medium with only a single main band (around 90 and 95 kDa, respectively) and with protein yields of 1.8 and 1.65 mg/ L cell culture after the Ni-NTA, respectively. However, these protein yields were not accurate because of the high number of contaminants eluted together with hSKI1. Even performing anion exchange chromatography, the BTMD proteins were not properly separated from these contaminants (but they were not precipitating into the column, like in the case of insect cells, so they seemed to be well folded). We then concluded that all these contaminants could be prodomain pieces of hSKI1 forming heterodimers with immature hSKI1 proteins, as it was previously proposed by Pasquato et al. in 2014 and Kunz et al. in 2015. In addition to this, the presence of chaperone proteins coeluting during hSKI1 purifications has been previously described (Espenshade et al., 1999), which may form transient complexes with it (like they usually do with many other ER proteins).

To summarize, the BTMD hSKI1 proteins expressed from residue 18, with the extracellular signal peptide (pSport6-hSKI1 (18-947)-H6 and pSport6-hSKI1 (18-998)-H6 constructs) and truncated before the transmembrane domain, were forced to be expressed extracellularly, when it is well known, that the hSKI1 processing takes places in organelles (first, RE and then, Golgi) and this is the reason why all the species of hSKI1 travelled together. They activate each other and also, carry the prodomain pieces (see Fig. 10) (da Palma et al., 2014). If this very complex process is moved to the extracellular medium, it will probably become even more complicated without the proper membrane signals for the processing, activation and secretion of the full-active secreted form (from residues 187 to 953) of hSKI1.

All this would explain why the protein is pure when it is naturally secreted to the medium (using the secreted form of hSKI1, whereas all the immature forms of hSKI1 with their prodomain pieces and chaperone proteins would stay at ER and Golgi) and impure and with aggregates when the secretion is forced. We would only need to perform a N-terminal sequencing of the expressed protein starting in the natural Met^{1} (pXLG40-hSKI1 (1-998)-H6) to validate this hypothesis and analyse by peptide mass fingerprinting several of the “contaminants” bands in the pSport6 expressions starting in 18.

Also, the separation of these PD pieces and their analysis can be performed: by reversed phase high performance liquid chromatography (RP-HPLC), as seen before (Da Palma et al., 2014 and Touré at al., 2000). PD fragments may be linked by hydrophobic interactions (also seen in other subtilisins) (Gallagher et al., 1995; Ohta et al., 1991).

Regarding the kinetic assays (proteolytic activity assays), a kinetic experiment should be performed with pure secreted protein from the construct starting in Met^{1} (pXLG40-hSKI1 (1-998)-H6) at 25°C in 25 mM MES, 25mM Tris/HCl and 2,5 mM CaCl2, pH 7.4 and with the two fluorogenic substrates already published: Abz-RSLK234YAESDY(NO2) A and Ac-RRLL-pNA (Bodvard et al., 2006 and Touré et al., 2000).

Other interesting experiments that should be performed, are the generation and expression in mammalian cells of:

- The full-length of hSKI1 (pSport6-hSKI1 (1-1052)-H6 construct): Because this is the only construct which is going to be able to produce the natural secreted hSKI1 form, starting in 187 and properly cleaved in 953, and 100% active (Elagoz et al., 2002).

- The “engineered shed hSKI1” (pSport6-hSKI1 (1-953)-H6 construct): To be compared in activity and folding with the previous full-length construct. The best construct should be assayed for crystallization of the “secreted hSKI1” (from 187 to 953 residues).

- The activity mutant of the hSKI1 full-length (pSport6-hSKI1 (1-1052)-H6 construct) with H242A mutation: This mutation was previously described to fully inactivate hSKI1 and get only one band of around 128 kDa in cellular pellet (da Palma et al., 2014; da Palma et al., 2015; Elagoz et al., 2001). That could be useful to crystallize the pro-hSKI1 with all the prodomain sequence.

- The activity mutant of the BTMD (pSport6-hSKI1 (18-998)-H6 construct with H242A mutation: To purify the pro-hSKI1 in only one band from the supernatant, if it is secreted.

And finally, if the protein yield is not enough for crystallization trials (which are highly protein demanding), perform stably transfected lines in mammalian cells with the more successful constructs or recombinant Vaccinia virus infection (Seidah et al., 1999).

In recent studies (da Palma et al., 2015), it was wisely observed that *Drosophila melanogaster* S1P lacks the BC region (between B’/B and C’/C cleavages) of human SKI1’s prodomain. A *Drosophila*-human chimera construct was designed, allowing the expression and secretion of a mature and a partially active enzyme in mammalian CHO-K1 and SRD12B cells (residues: 1 to 22 (human SP); 19 – 141 (*D. melanogaster* prodomain) and 187 – 1052 (human SKI1)). Indeed, this strategy should be also

tried and mimicked with the aim of obtaining crystals of hSKI1, in insect or mammalian expression systems.

Finally, the expression and crystallization of the hSKI1 prodomain have never been assayed, and the expression of the 1-186 and 18-186 hSKI1 constructs in BL21 (DE3) and Rosetta cells could be worth it. Moreover, if some of these protein expressions is successful, we could use it like an external “foldase” and try to refold some of the hSKI1 proteins expressed in bacteria in inclusion bodies, like for example pCri7a-hSKI1 (18-998)-H6 and pCri7a-hSKI1 (188-477)-H6, using it in the refolding buffer. This strategy was used by Ohta et his collaborators in 1991 and in this case, they achieved the refolding and activation of the subtilisin E with a synthetic propeptide.

Figure reproduced from “pixpa.com” (Austin Siadak)

RESULTS

3. 3 Matrix metalloproteinases (MMP) outside vertebrates

Marino-Puertas L., Goulas T and Gomis-Rüth FX. (2017). Matrix metalloproteinases outside vertebrates. *Biochimica Biophysica Acta Molecular Cell Research, 1864*(11 Pt A):2026-2035.

3.3.1 OBJECTIVES

The matrix metalloproteinase (MMP) family belongs to the metzincin clan of zinc-dependent metallopeptidases. Due to their enormous implications in physiology and disease, MMPs have mainly been studied in vertebrates. They are engaged in extracellular protein processing and degradation, and present extensive paralogy, with 23 forms in humans. One characteristic of MMPs is a ~165-residue catalytic domain (CD), which has been structurally studied for 14 MMPs from human, mouse, rat, pig and the oral-microbiome bacterium *Tannerella forsythia*. These studies revealed close overall coincidence and characteristic structural features, which distinguish MMPs from other metzincins and give rise to a sequence pattern for their identification. Here, we are going to review the literature available on MMPs outside vertebrates and perform database searches for potential MMP CDs in invertebrates, plants, fungi, viruses, protists, archaea and bacteria.

3.3.2 RESULTS AND DISCUSSION

3.3.2.1 Matrix metalloproteinases in invertebrates

Matrix metalloproteinases (MMPs) have been extensively studied in vertebrates, particularly in mammals, due to their widespread implications in human health and disease (Fanjul-Fernández et al., 2010; López-Otín et al., 2009; Mittal et al., 2016; Page-McCaw et al., 2007). However, they are also widely present in invertebrates, where they participate in tissue turnover processes such as dendritic remodelling, regeneration, tracheal growth, axon guidance, histolysis and matrix degradation during hatching. They also participate in cancer processes (Page-McCaw, 2008). Studied organisms belong to the phyla Arthropoda, including insects (dipterans, lepidopterans, hymenopterans and coleopterans) and crustaceans; Echinoderma, including sea urchins and sea cucumbers; Mollusca, including mussels and snails; Nematoda, Annelida and Planaria; invertebrate Chordata, including sea squirt and lancelet; and Cnidaria, including hydra and jellyfish (see Table 12). In addition, we found MMP-like sequences in the genomes of organisms as primitive as starlet sea anemone (*Nematostella vectensis*; UniProt database (www.uniprot.org) access code [UP] A7RJ22), which is also from the phylum Cnidaria, and sea gooseberry (*Pleurobrachia pileus*; GenBank database (www.ncbi.nlm.nih.gov/genbank) access code [GB] FQ005948) from the phylum Ctenophora, i.e. MMPs are present across the subkingdom Eumetazoa. In contrast, MMPs are apparently absent from the genomes of the sponge *Amphimedon queenslandi*ca (phylum Porifera; (Pisani et al., 2015) and of *Trichoplax adherens* (phylum Placozoa; (Srivastava et al., 2008), which belong to the basalmost clades of metazoans.

Eumetazoans are animals comprised of tissues organized into germ layers, with a gastrula stage during embryogenesis. Primitive animals of this lineage up to the emergence of the clade Bilateria likely had few MMP copies within their genomes, as currently found in several Drosophila species (two copies) and other insects such as red flour beetle, silkworm or malaria mosquito (three copies) () (Fanjul-Fernández et al., 2010; Ishimwe et al., 2015; Knorr et al., 2009; Llano et al., 2000; Llano et al., 2002; Page-McCaw, 2008). In contrast, higher animals show gene polyplication, which likely arose during early vertebrate evolution and led to the substantial paralogy that is currently found, e.g., in humans (24 genes), mice (23 genes) and zebrafish (26 genes). However, development of this complexity was not linear: while seven genes are found in the sea squirt *Ciona intestinalis* and nine in the mosquito *Aedes aegypti*, the sea urchin *Strongylocentrotus purpuratus* has 26 (Angerer et al., 2006; Fanjul-Fernández et al., 2010; Huxley-Jones et al., 2007; Kantor et al., 2017; Massova et al., 1998; Page-McCaw, 2008; Quesada et al., 2010; Small and Crawford, 2016)

The complex evolution of MMPs in Eumetazoa is also reflected by the presence of several ancillary domains in the full-length enzymes (see Section 1.4.1 in Introduction and Fig. 1 in (Fanjul-Fernández et al., 2010)). It can be speculated that primitive MMPs consisted of standalone catalytic domains (CDs) or possibly prodomain (PD)+ CD tandems, which underwent duplication, gene fusion and exon shuffling to result in multi-domain architectures (Fanjul-Fernández et al., 2010; Massova et al., 1998; Murphy et al., 1991; Pal and Guda , 2006). However, in some instances evolution progressed in the opposite direction, i.e. multidomain enzymes underwent truncation to yield proteins of fewer domains, even in mammals. This is the case for matrilysins, which span only a signal peptide, a PD with a cysteine-switch motif and a CD (Fanjul-Fernández et al., 2010; Massova and Mobashery, 1998; Puente et al., 2003). This minimal architecture is also predominant across invertebrates (Fanjul-Fernández et al., 2010), although many MMP sequences in insects—e.g. in Drosophila (Page-McCaw, 2008)—further comprise a furin recognition sequence, an HD and a GPI anchor.

Table 12 — MMPs reported from invertebrates

Phylum *Annelida*	
Class *Polychaeta*	
Boneworm (*Osedax japonicus*)	(Miyamoto et al., 2017)
Phylum *Arthropoda*	
Class *Branchiopoda*	
Water flea (*Daphnia pulex*)	(Spanier et al., 2010)
Class *Insecta*	
Yellow-fever mosquito (*Aedes aegypti*)	(Kantor et al., 2017)
Malaria mosquito (*Anopheles gambiae*)	Goulielmaki et al., 2014)
Honeybee (*Apis mellifera*)	Ueno et al., 2009)
Silkworm (*Bombyx mori*)	(Guan et al., 2009)
Fruitfly (*Drosophila melanogaster*)	(Llano et al., 2000; Llano et al., 2002; Page-McCaw, 2008)
Greater wax moth (*Galleria mellonella*)	(Altincicek and Vilcinskas, 2008)
Tobacco hornworm (*Manduca sexta*)	(Vishnuvardhan et al., 2013)
Red flour beetle (*Tribolium castaneum*)	(Knorr et al., 2009)
Cabbage looper (*Trichoplusia ni*)	(Means and Passarelli, 2010)
Phylum *Chordata*	
Class *Ascidiacea*	
Sea squirt (*Ciona intestinalis*)	(Huxley-Jones et al., 2007)
Class *Leptocardii*	
Lancelet (*Branchiostoma japonicum*)	(Zhang et al., 2012)
Phylum *Cnidaria*	

Class *Hydrozoa*	
Common hydra (*Hydra vulgaris* / *Hydra magnipapillata*)	(Leontovich et al., 2000; Shimizu et al., 2002)
Class *Scyphozoa*	
Nomura's jellyfish (*Nemopilema nomurai*)	(Kang et al., 2014)
Phylum *Echinodermata*	
Class *Echinoidea*	
Asian sea urchin (*Hemicentrotus pulcherrimus*)	(Nomura et al. 1997)
Mediterranean purple sea urchin (*Paracentrotus lividus*)	(Ghiglione et al., 1994)
Pacific purple sea urchin (*Strongylocentrotus purpuratus*)	(Angerer et al., 2006)
Class *Holothuroidea*	
Japanese sea cucumber (*Apostichopus japonicus*)	(Sun et al., 2011)
Rock sea cucumber (*Holothuria glaberrima*)	(Quiñones et al., 2002)
Phylum *Mollusca*	
Class *Bivalvia*	
American oyster (*Crassostrea virginica*)	(Nikapitiya et al., 2014)
Mediterranean mussel (*Mytilus galloprovincialis*)	(Mannello et al., 2001)
Class *Gasteropoda*	
Giant African snail (*Achatina fulica*)	(Indra et al., 2005)
Bloodfluke planorb (*Biomphalaria glabrata*)	(Yoshino et al., 2014)
Many-colored abalone (*Haliotis diversicolor*)	(Wang et al., 2008)
Red abalone (*Haliotis rufescens*)	(Chovar-Vera et al., 2015)
Korean common dogwhelk (*Thais clavigera*)	(Rhee et al., 2012)
Phylum *Nematoda*	
Class *Secernentea*	
Rat lungworm (*Angiostrongylus cantonensis*)	(Sun et al., 2012)
Roundworm (*Caenorhabditis elegans*)	(Altincicek et al., 2010; Coates et al., 2000; Wada et al., 1998)
Yellow potato cyst nematode (*Globodera rostochiensis*)	(Kovaleva et al., 2004)
Parasitic nematode (*Gnathostoma spinigerum*)	(Uparanukraw et al., 2001)
Soybean cyst nematode (*Heterodera glycines*)	(Kovaleva et al., 2004)
Phylum *Platyhelminthes*	
Class *Rhabditophora*	
Freshwater planarian flatworm (*Dugesia japonica*)	(Isolani et al., 2013)
Freshwater planarian flatworm (*Schmidtea mediterranea*)	(Isolani et al., 2013)

Taxonomy according to the Catalogue of Life (http://www.catalogueoflife.org/col; (Roskov et al., 2017)).

3.3.2.2 Matrix metalloproteinases in plants and algae

In addition to Eumetazoa, MMPs have been reported from higher plants (phylum Tracheophyta), where they are generally present in lower copy numbers than in animals. Enzymes were described

from soybean, mouse-ear cress, cucumber, barrel medic, tobacco, loblolly pine and tomato (see Table 13). In addition, sequences were referenced from sugar beet, rice, corn, sugar cane and barley (Table 13). Moreover, metalloproteinase (MP) activity attributable to an MMP was also described for the jack bean *Canavalia ensiformis*, though validation is still pending (Gonçalves et al., 2016). Finally, current sequence similarity searches identified several hundreds of potential hits in higher plants, Table 13 provides a curated selection of them.

Plant MMPs localize to the plasma membrane or the extracellular space and have been found to be involved in remodeling of the extracellular matrix during plant growth and development processes, such as germination, programmed cell death and senescence, as well as in biotic and abiotic stress responses (Clark et al., 2013; Flinn, 2008; Marino and Funk, 2012; Zimmermann et al., 2016) . Sequence analyses revealed that plant MMPs have a homogenous domain architecture and mostly comprise a signal peptide, a PD with a cysteine-switch motif that occasionally diverges from the consensus (see Section 1.4.1 in Introduction and (Marino and Funk, 2012), and a CD. Within the latter, some sequences have the general base/acid glutamate mutated to glutamine (Marino and Funk, 2012), a change that in mammalian MMPs leads to ablation or strong reduction of proteolytic activity (Rowsell et al., 2002). Uniquely, plant MMPs encompass a specific consensus sequence, D-L-E-S/T, two residues upstream of the zinc-binding motif (Clark et al., 2013; Li et al., 2015; Maidment et al., 1999). In addition, the loop that connects strand βV with helix αB is up to 10 residues longer in plant MMPs than in mammalian counterparts. In the absence of experimental structures, these two combined features point to a specific structural element of plant MMPs, putatively an extra cation-binding site. Downstream of the CD, plant MMPs only contain ~40-residue C-terminal GPI anchors or transmembrane segments for localization to the plasma membrane (Li et al., 2015).

Table 13 — MMPs referenced from plants and selected sequences within genomes

Phylum *Tracheophyta*	
Class *Liliopsida*	
Barley (*Hordeum vulgare*): F2DC11, F2D2W1, F2DGF5, F2D593	(Li et al., 2015)
Rice (*Oryza sativa*): A2X9G3, A2YB49, A2ZA53, A2ZA54, A2ZA55, A2ZA56, A2ZA53, A2ZA53	(Li et al., 2015)
Sugar cane (*Saccharum hybrid cultivar*): A0A059Q041	(Ramos and Selistre-de-Araujo, 2001)
Corn (*Zea mays*): B6U4D1, B4FZU1	(Li et al., 2015)
Eelgrass (*Zostera marina*): A0A0K9P9J4*, A0A0K9PG73*, A0A0K9PXQ4*, A0A0K9PK10*, A0A0K9PLB4*, A0A0K9PY38*, A0A0K9PBG9*, A0A0K9P805*	
Class *Magnoliopsida*	
Lyre-leaved rock cress (*Arabidopsis lyrata*): D7LBU9*	
Mouse-ear cress (*Arabidopsis thaliana*): At1/2/3/4/5-MMP (O23507, O04529, Q5XF51, Q8GWW6, Q9ZUJ5)	(Flinn, 2008; Golldack et al.,

	2002; Liu and Graham, 1998; Maidment et al., 1999; Marino et al., 2014; van der Hoorn et al., 2011)
Sugar beet (*Beta vulgaris*): A0A0J8F9H7, A0A0J8B8R0, GB XP_010692448, GB XP_010695895	(Broccanello et al., 2015)
Cucumber (*Cucumis sativus*): Cs1-MMP (Q9LEL9)	(Delorme et al., 2000)
Soybean (*Glycine max*): Gm1/2-MMP, Gm-Slti114-MMP (C6TNN5, Q93Z89, B6CAM2) and C6TNN5	(Cho et al., 2009; Graham et al., 1991; Li et al., 2015; Liu et al., 2001; McGeehan et al., 1992; Pak et al., 1997; Ragster and Chrispeels, 1979)
Tree cotton (*Gossypium arboreum*): A0A0B0ME77*	
Barrel medic (*Medicago truncatula*): Mt1-MMP (Q9ZR44)	(Combier et al., 2007)
Tobacco (*Nicotiana tabacum*): Nt1-MMP (C3PTL6)	(Mandal et al., 2010; Schiermeyer et al., 2009)
Western balsam poplar (*Populus trichocarpa*): B9I3X8*	
Castor bean (*Ricinus communis*): B9RUG7*	
Tomato (*Solanum lycopersicum*): Sl1/2/3/4/5-MMP (I7JCM3, I7KJ40, K4BWG3, K4CNL4, A0A0G3ZAU2, K4CYZ6)	(Li et al., 2015)
Cocoa (*Theobroma cacao*): A0A061DVV2*, S1S448*	
Class *Pinopsida*	
Loblolly pine (*Pinus taeda*): Pta1-MMP (B7TVN4)	(Ratnaparkhe et al., 2009)

All sequence codes are from UniProt (UP; www.uniprot.org) or GenBank (GB; www.ncbi.nlm.nih.gov/genbank).
Taxonomy according to the Catalogue of Life (http://www.catalogueoflife.org/col; (Roskov et al., 2017)).
Nomenclature of validated plant MMPs according to (Clark et al., 2013).
* Sequences annotated in UniProt as "matrix metalloproteinases" and manually curated.
Searches completed on 10 January 2017.

The unicellular green alga *Chlamydomonas reinhardtii* encodes two MPs dubbed gametolysins (also known as gamete lytic enzymes, MMP1 and MMP2, and autolysins), which are engaged in cell-wall

turnover and have been recurrently associated with the MMP family (Claes, 1971; Clark et al., 2013; Kinoshita et al., 1992; Kubo et al., 2001; Marino and Funk, 2012; Matsuda, 2013). Similarly, to MMPs, these ~635-residue enzymes comprise a signal peptide and a putative PD with a cysteine-switch-like motif. These enzymes are similar to four enzymes from the multicellular green alga *Volvox carteri* (VMP1-VMP4), in which the first zinc-binding histidine is replaced with glutamine (Hallmann et al., 2001; Heitzer and Hallmann, 2002; Shimizu et al., 2002). Current sequence similarity searches identified new potential paralogues of these MPs in C. reinhardtii and V. carteri, as well as in the green algae *Gonium pectorale* and *Chlorella variabilis*. However, in the absence of three dimensional structures, the primary sequences of these MPs deviate from the extended sequence pattern of MMPs (see Section 1.4.1 in Introduction). Consequently, they belong to a separate metzincin family, which was tentatively dubbed gametolysins in the past (Cerdà-Costa and Gomis-Rüth, 2014; Gomis-Rüth, 2003; Kubo et al., 2001). This is consistent with their adscription to a family separate from MMPs within the MEROPS peptidase database (M11 vs. M10; see merops. sanger.ac.uk; (Rawlings et al., 2016)). Of note, searches for true MMPs within eukaryotic algae revealed six potential relatives, two in phytoplankton *Emiliania huxleyi* (UP R1DV14 and R1E2E0/R1EHE7) and one each in G. pectoral (UP A0A150GPR5), *Symbiodinium minutum* (KEGG Gene symbB.v1.2.029330.t2; see www.genome.jp), *Aureococcus anophagefferens* (UP F0Y382) and *V. carteri* (UP D8UD00). This restricted presence of MMPs within algae is consistent with recent genome-wide analyses of the secretomes of nine brown algae belonging to the phyllum Phaeophyceae, which revealed no potential MMP ortholog (Terauchi et al., 2017). Furthermore, no other sequences were presently found in other Protista/Chromista.

3.3.2.3 Matrix metalloproteinases in fungi

Although some MPs were described from fungi (McHenry et al., 1996; Pavlukova et al., 1998; Terashita et al., 1985), none has yet been confirmed to be an MMP. We thus performed a database search, which revealed several potential sequences in the phylum Ascomycota but not in other phyla (Table 14). These hypothetical proteins span between 255 and 659 residues, some have their catalytic glutamate replaced with glutamine (see Sections 1.4.1 in Introduction and 3.3.2.2 in Results), and most comprise N-terminal extensions (NHS) with a potential cysteine-switch-like motif. Those lacking this motif possess potential NTSs that are significantly shorter than in standard MMPs, so they may correspond to incomplete sequences or truncated variants. Several sequences also contain large N- and/or C-terminal extensions that could correspond to additional domains. Hit fungal species contain only one sequence each, with the exception of *Arthrobotrys oligospora* and *Dactylellina haptotyla,* with five sequences each, and all but two species are fungal pathogens of higher plants, nematodes,

insects and animals—including humans—or endophytic fungi. Accordingly, their lifestyle entails very intimate contact with organisms that contain MMPs.

Table 14 — Fungal MMP sequences

Phylum *Ascomycota*		
Class *Dothideomycetes*		
Bipolaris oryzae	388 residues	GB XP_0077684757 / UP W6ZMI0
Bipolaris victoriae	388 residues	GB XP_014556326 / UP W7EIS0
Bipolaris zeicola	388 residues	GB XP_007709037 / UP W6YG21
Cochliobulus heterostrophus (Bipolaris maydis)	388 residues	GB XP_014075121 / UP N4X7I4, UP M2V381 **
Cochliobulus sativus (Bipolaris sorokiniana)	320 residues	GB XP_007696738 / UP M2TCL8
Paraphaeosphaeria sporulosa	411 residues	GB XP_018039150 / UP A0A177CMM7
Phaeosphaeria nodorum (Parastagonospora nodorum) (Septoria nodorum)	326 residues	GB XP_001794978 / UP Q0UUK1
Pyrenochaeta sp.	565 residues	UP A0A178DJW7
Stagonospora sp.	393 residues	UP A0A178BDB5
Stemphylium lycopersici	563 residues	UP A0A0L1HKC2
Class *Eurotiomycetes*		
Aspergillus calidoustus	290 residues	UP A0A0U5G9D2
Aspergillus flavus	274 residues	GB XP_002379978 / UP B8NI13, UP A0A0D9MRC7 **
Aspergillus lentulus	311 residues	UP A0A0S7DKQ5
Aspergillus udagawae	282 residues	UP A0A0K8L198
Endocarpon pusillum	306 residues	GB XP_007803399 / UP U1HKZ8
Exophiala aquamarina	282 residues	GB XP_013264742 / UP A0A072PQX4
Neosartorya fischeri (Aspergillus fischerianus)	616 residues	GB XP_001261124 / UP A1DIM0
Class *Leotiomycetes*		
Pseudogymnoascus sp.	255 residues	UP A0A094DQI1, UP A0A094ITE6 *
Class *Orbiliomycetes*		
Arthrobotrys oligospora	609 residues	GB XP_011127847 / UP G1XU64
Arthrobotrys oligospora	599 residues	GB XP_011127846 / UP G1XU63
Arthrobotrys oligospora	290 residues	GB XP_011126672/ UP G1XQB0
Arthrobotrys oligospora	315 residues	GB XP_011126207 / UP G1XNZ5
Arthrobotrys oligospora	232 residues	GB XP_011120912 / UP G1X8V0
Drechslerella stenobrocha	210 residues	UP W7IFU4
Class *Sordariomycetes*		
Dactylellina haptotyla	659 residues	GB XP_011116834 / UP S7ZXB5
Dactylellina haptotyla	308 residues	GB XP_011114406 / UP S8A464
Dactylellina haptotyla	292 residues	GB XP_011112095 / UP S8AFF5
Dactylellina haptotyla	268 residues	GB XP_011113422 / UP S8BTL0
Dactylellina haptotyla	246 residues	GB XP_011107496 / UP S8AU63
Metarhizium anisopliae	619 residues	UP A0A0B4G9D2
Metarhizium anisopliae	607 residues	UP A0A0D9NZH8
Metarhizium brunneum	619 residues	GB XP_014544223 / UP A0A0B4FXN2

Metarhizium robertsii	619 residues	GB XP_007825075 / UP E9F9D7, A0A014N9M3 **
Pochonia chlamydosporia	636 residues	GB XP_018148437 / UP A0A179G3B6
Purpureocillium lilacinum	627/629 residues	GB XP_018174930 / UP A0A179GT13, A0A179GDA3 *

All sequence codes are from UniProt (UP; www.uniprot.org) or GenBank (GB; www.ncbi.nlm.nih.gov/genbank).
Sequence searches were performed with structurally validated MMP CDs within UniProt (www.uniprot.org) or the National Center for Biotechnology Information (blast.ncbi.nlm.nih.gov/Blast.cgi) using standard parameters.
Taxonomy according to the Catalogue of Life (http://www.catalogueoflife.org/col; (Roskov et al., 2017)).
* These sequences have only minimal differences.
** These entries are identical.
Searches completed on 19 January 2017.

3.3.2.4 Matrix metalloproteinases in viruses

In 2000, the functional characterization of an MMP from *Xestia cnigrum granulovirus* was reported (Ko et al., 2000). In the absence of structural information, analysis of its amino acid sequence reveals an MMP CD. The upstream N-terminal segment lacks the canonical cysteine-switch motif, but shows sequence stretch C42-G-G-G-N-H-R-R-T-K-R52 immediately before the predicted mature N-terminus, which includes a cysteine-glycine pair reminiscent of those motifs, as well as a recognition sequence characteristic of furin-activatable MMPs (see (Fanjul-Fernández et al., 2010) and Section 1.4.1 in Introduction). This notwithstanding, the full-length protein was active, thus suggesting it was in an at least partially competent state without hypothetical activation (Ko et al., 2000). Threading calculations (see the legend to Fig. 62A) with the protein segment downstream of the CD of Xestia MMP indicated that it most likely contained an HD (Fig. 62A), preceded by an intermediate “threonine-rich region”. More recently, *Cydia pomonella granulovirus* was also shown to express a functional MMP, the only other viral family member studied to date (Ishimwe et al., 2015). Like the Xestia ortholog, it contained a CD preceded by a potential furin-recognition sequence. However, it lacked any cysteine in the N-terminal fragment, so its function and/or activation mechanism might diverge from Xestia MMP. In contrast to previous hypotheses (Ishimwe et al., 2015), but in agreement with the prediction for Xestia MMP, threading calculations suggested that a threonine-rich region and an HD are present in the C-terminal part of the protein (Fig. 62B).

Granuloviruses belong to the genus Betabaculovirus within the family Baculoviridae and all species sequenced to date within this genus contain putative MMP homologs (see Table 15 and (Ishimwe et al., 2015)). In contrast, MMPs are absent from the other *Baculoviridae geni*, viz. Alphabaculovirus, Deltabaculovirus and Gammabaculovirus. In addition to Baculoviridae, we also found MMP sequences

in the families Iridoviridae, Nudiviridae, Poxviridae (subfamily Entomopoxvirinae) and Ascoviridae Table 15) but not in any other viruses or viroids. Collectively, all these virus families are double-stranded DNA viruses with no RNA stage, which infect invertebrates (arthropods, lepidopters, hymenopters, diptera and decapods), i.e. organisms that contain MMPs. The potential viral MMPs vary in length, some contain a signal peptide, a potential cysteine-switch-like motif, a putative furin-cleavage site, a predicted HD and a threonine-rich region as in Xestia and Cydia MMPs, but others do not. Moreover, and consistent with the host specificity of the harboring viruses, sequences cluster closely with insect MMPs (Ishimwe et al., 2015), which possess a similar domain architecture (see Section 3.3.2.1 in Results and (Page-McCaw, 2008)). In insects, housekeeping MMPs participate in physiological remodeling of the basal lamina, which lines the midgut to prevent systemic infections (Passarelli, 2011). In turn, as part of the infective process, ingested viral particles reach the midgut of target insects and breach the basal lamina, a task that might be carried out by viral MMPs (Passarelli, 2011).

Table 15 — Viral MMP sequences

Family *Ascoviridae*		
Genus *Ascovirus*		
Heliothis virescens ascovirus 3e	474 residues	GB YP_001110872 / UP A4KX75
Heliothis virescens ascovirus 3f	439 residues	GB AJP08985 / UP A0A171PVB2
Heliothis virescens ascovirus 3g	439 residues	GB AFV50272 / UP K4NY21
Spodoptera frugiperda ascovirus 1a	386 residues	GB YP_762369 / UP Q0E587
Trichoplusia ni ascovirus 2c	501 residues	GB YP_803380 / UP Q06VC5
Genus *Toursvirus*		
Diadromus pulchellus ascovirus 4a	223 residues	GB YP_009220674 /UP F2NYV8
Family *Baculoviridae*		
Genus *Betabaculovirus*		
Adoxophyes orana granulovirus	395 residues	GB NP_872491 / UP Q7T9X8
Cryptophlebia leucotreta granulovirus	486 residues	GB NP_891890 / UP Q7T5P6
Cydia pomonella granulovirus #	546 /545 residues	GB AIU36692 / UP A0A097P0M6 // GB NP_148830 / UP Q91F09 *
Helicoverpa armigera granulovirus	596 residues	GB YP_001649020 / UP A9YMN0
Phthorimaea operculella granulovirus	469 residues	GB NP_663206 / UP Q8JS18
Plodia interpunctella granulovirus	620 residues	GB YP_009330170 / UP
Plutella xylostella granulovirus	402 / 403 residues	GB NP_068254 / UP Q9DVZ7 // GB ANY57554 / UP A0A1B2CSG4 *
Pseudalatia unipuncta granulovirus	593 residues	GB YP_003422377 / UP B6S6Q7
Trichoplusia ni granulovirus LBIV-12	593 residues	GB AOW41373 / UP A0A1D8QL47
Xestia c-nigrum granulovirus #	469 residues	GB NP_059188 / UP Q9PZ03
Unclassified *Betabaculoviridae*		

Agrotis segetum granulovirus	481 residues	GB YP_006303 / UP Q6QXG0
Choristoneura occidentalis granulovirus	488 residues	GB YP_654454 / UP Q1A4R3
Clostera anachoreta granulovirus	412 residues	GB YP_004376233 / UP F4ZKQ2
Clostera anastomosis granulovirus	412 residues	GB YP_008719974 / UP U5KB82
Clostera anastomosis granulovirus	486 residues	GB AKS25377 / UP A0A0K0WSE3
Cnaphalocrocis medinalis granulovirus	441 residues	GB ALN41975 / UP A0A0X9FQ45 // GB YP_009229958 / UP A0A109WW48 *
Diatraea saccharalis granulovirus	370 residues	GB YP_009182238 / UP A0A0R7EZ65
Epinotia aporema granulovirus	359 residues	GB YP_006908552 / UP K4ERV0
Erinnyis ello granulovirus	462 residues	GB YP_009091878 / UP A0A097DAI6
Mocis sp. granulovirus	580 residues	GB YP_009249873 / UP A0A162GWM9
Pieris rapae granulovirus	432 residues	GB YP_003429361 / UP D2J4K4
Pieris rapae granulovirus	444 residues	GB ADO85463 / UP E7BN22
Spodoptera frugiperda granulovirus	539 residues	GB YP_009121819 / UP A0A0C5B309
Spodoptera litura granulovirus	464 residues	GB YP_001256988 / UP A5IZN9
Family *Iridoviridae*		
Genus *Iridovirus*		
Invertebrate iridescent virus 6 (Chilo iridescent virus)	264 residues	GB NP_149628 / UP O55761
Invertebrate iridovirus 22	352 residues	GB YP_008357315 / UP S6DDP6
Invertebrate iridescent virus 22	365 residues	GB YP_009010779 / UP W8W1A0
Invertebrate iridovirus 25	362 residues	GB YP_009010550 / UP W8W2D9
Invertebrate iridescent virus 30	367 residues	GB YP_009010310 / UP W8W249
Wiseana iridescent virus (Insect iridescent virus type 9)	346 residues	GB YP_004732919 / UP G0T5G2
Genus *Chloriridovirus*		
Aedes taeniorhyncus iridescent virus	363 residues **	GB ABF82125
Invertebrate iridescent virus 3 (IIV-3) (Mosquito iridescent virus)	363 residues **	GB YP_654667 / UP Q196W5
Unclassified *Iridoviridae*		
Anopheles minimus irodovirus	366 residues	GB YP_009021109 / UP W8QE20
Family *Nudiviridae*		
Genus *Betanudivirus*		
Helicoverpa zea nudivirus 2 (HzNV-2)	789 residues	GB YP_004956816 / UP G9I094
Heliothis zea nudivirus	792 residues	GB AAN04364 / UP Q8JKP2
Family *Poxviridae* (Subfamily *Entomopoxvirinae*)		
Genus *Alphaentomopoxvirus*		
Anomala cuprea entomopoxvirus	268 residues	GB YP_009001641 / UP W6JIV8
Genus *Betaentomopoxvirus*		
Amsacta moorei entomopoxvirus (AmEPV)	252 residues	GB AAG02776, GB NP_064852 / UP Q9EMX9 **
Mythimna separata entomopoxvirus 'L'	411 residues	GB YP_008003705 / UP R4ZFL1
Unclassified dsDNA viruses, no RNA stage		

Apis mellifera filamentous virus	830 residues	GB AKY03287, GB YP_009165969 / UP A0A0K1Y874 **
Apis mellifera filamentous virus	1354 residues	GB AKY03074, GB YP_009165756 / UP A0A0K1Y866 **

All sequence codes are from UniProt (UP; www.uniprot.org) or GenBank (GB; www.ncbi.nlm.nih.gov/genbank).
Sequence searches were performed with MMP CDs within UniProt (www.uniprot.org) or the National Center for Biotechnology Information (blast.ncbi.nlm.nih.gov/Blast.cgi) using standard parameters.
Taxonomy according to the International Committee on Taxonomy of Viruses (ictvonline.org/virusTaxonomy.asp), NCBI (www.ncbi.nlm.nih.gov/taxonomy) or the Catalogue of Life (http://www.catalogueoflife.org/col; Nagase et al., 2006).
[#] *Xestia* and *Cydia* MMPs are the only viral enzymes studied.
* These entries have only minimal differences.
** These entries are identical.
Searches completed on 10 January, 2017.

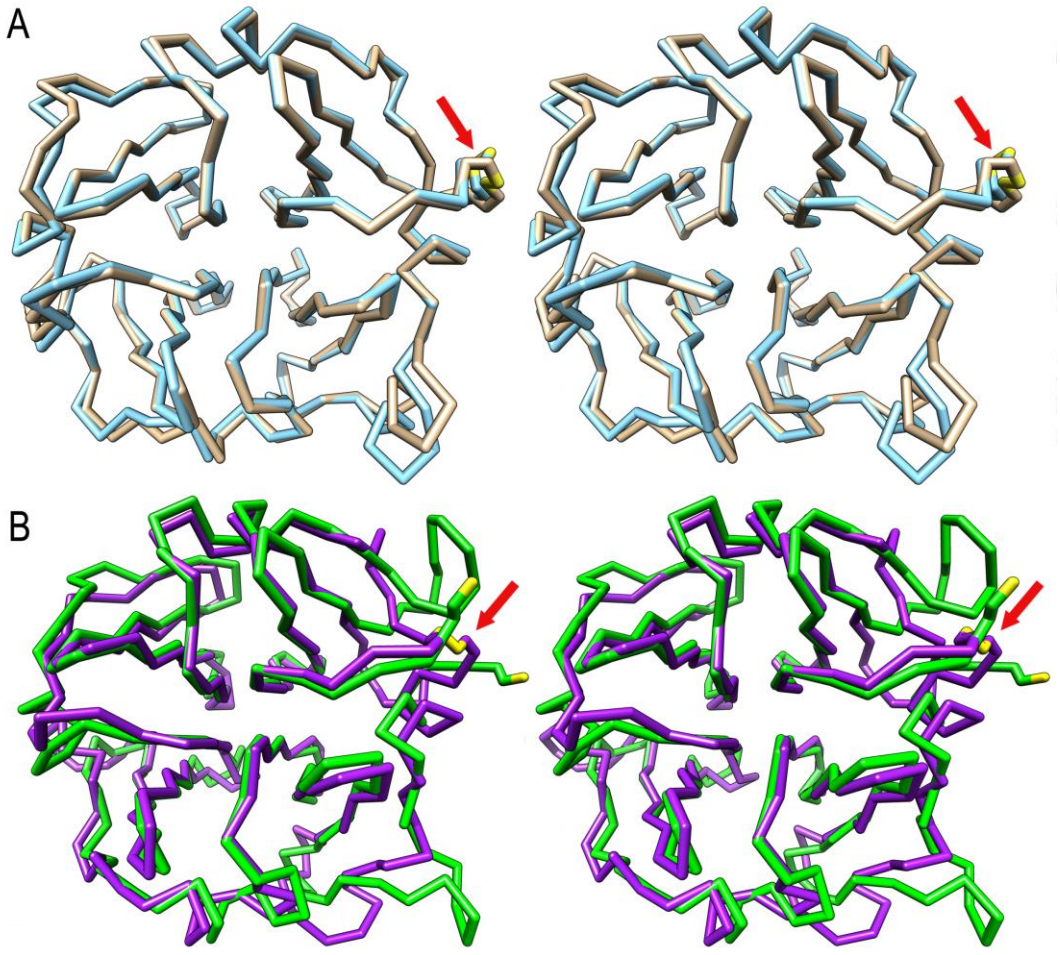


Figure 62. Predicted structure of baculoviral MMP C-terminal domains. (A) Superposition in cross-eye stereo of the two top-ranked homology models (in tan and cyan respectively) of *Xestia cnigrum* granulovirus MMP segment C280-C469 (UP Q9PZ03) by threading calculations with the LOMETS meta-server using programs cdPPAS (Z-score 16.9; (Wu and Zhang, 2007)) and SP3 (Z-score 36.6; (Zhou and Zhou, 2005)), respectively. Full-length modelling was automatically carried out with program MODELLER (Eswar et al., 2006). The corresponding Cα-traces and the critical disulphide bond linking the terminal β-leaflets

(C280-C469; red arrow), which is usually found in fourfold β-propeller structures such as MMP hemopexin domains (Gomis-Rüth, 2004), are depicted. In both cases, the automatically selected template structure was that of the hemopexin domain of human MMP-14 (PDB 3C7X; (Tochowicz et al., 2011). **(B)** Same as (A) but for *Cydia pomonella* granulovirus MMP (UP Q91F09) segment C359-C545. The two top predictions (in purple and green, respectively) were obtained with the program HHSEARCH (Z-scores 29.6 and 35.9; (Söding, 2005)). Selected templates were the hemopexin domains of human MMP-2 (PDB 1GEN; Libson et al., 1995) and MMP-1 (PDB 1SU3; Jozic et al., 2005). In the second automatic homology model, the two cysteines are not linked but close to each other in space.

3.3.2.5 Matrix metalloproteinases in archaea and bacteria

The only prokaryotic MMP characterized at the functional and structural level to date is karilysin from *Tannerella forsythia*. This is a human oral-microbiome bacterium from the phylum Bacteroidetes that is engaged in odontopathogenic infections (Bryzek et al., 2014; Cerdà-Costa et al., 2011; Guevara et al., 2013; Jusko et al., 2012; Karim et al., 2010; Koziel et al., 2010; López-Pelegrín et al., 2015; Potempa et al., 2013; Siddiqi et al., 2016; Skottrup et al., 2012). Karilysin comprises a CD flanked upstream by a signal peptide for secretion and a 14-residue PD, which does not proceed over a cysteine-switch mechanism but an "aspartate-switch" mechanism, as found in the otherwise unrelated astacin family within the metzincin clan (Cerdà-Costa et al., 2011; Gomis-Rüth et al., 2012; López-Pelegrín et al., 2015; Stöcker and Gomis-Rüth, 2013). The CD is followed by two domains of unknown structure and function, collectively spanning 275 residues, which comprise the C-terminal residues K-L-I-K-K. A C-terminal domain similar to karilysin is found in other unrelated peptidases within T. forsythia, Bacteroides sp. and Prevotella sp., which have been collectively termed KLIKK proteases (Ksiazek et al., 2015).

The CD of karilysin was characterized at the structural level (see Fig. 11A), which revealed that it fulfilled all the structural criteria of MMP CDs of mammals (Cerdà-Costa et al., 2011) (Fig. 11B). These studies also suggested that karilysin is evolutionary closer to forms from mosquitoes that are insect vectors of malaria (*Anopheles gambiae*), dengue fever, Chikungunya and yellow fever (*Aedes aegypti*), and West Nile virus and Zika virus infections (*Culex quinquefasciatus*) than to bacterial counterparts (Cerdà-Costa et al., 2011). The lifecycle of these mosquitoes entails feeding on human blood and they are mostly found in poor countries, which have the highest incidence of odontopathogenic bacterial infections (Axelsson et al., 2002).

Further to karilysin in bacteria, a potential MMP ortholog in Bacillus anthracis, MmpZ, was shown to participate in the extracellular degradation of anthrax toxin components and anthrolysin O at the onset of the stationary growth phase of the bacterium (Pomerantsev et al., 2011). However, detailed inspection of its protein sequence (UP Q81NM7) reveals that it deviates from the extended sequence pattern of MMP CTSs. Therefore, it cannot be assigned unambiguously to MMPs until its three-

dimensional structure has been resolved. Moreover, there have been other reports postulating the existence of bacterial MPs, which were hailed as ancestral forms of MMPs (Small and Crawford, 2016). In particular, Bacteroides fragilis toxins alias fragilysins were thought to accomplish this role (Massova et al., 1998; Moncrief et al., 1995; Obiso et al., 1997; Sears, 2009). However, when the structures of profragilysin-3 and the closely-related metalloproteinase II were reported, it became obvious that fragilysins, which are present in enterotoxigenic B. fragilis strains but not commensal ones, represent a metzincin family on their own, which is closer to adamalysin/ADAMs than MMPs, if at all (Cerdà-Costa and Gomis-Rüth, 2014; Goulas et al., 2011; Goulas et al., 2013; Shiryaev et al., 2014).

To complete the picture of MMP distribution in prokaryotes, we conducted sequence similarity searches and identified several hundred potential MMP orthologs across archaeal and bacterial genomes, some of them with several copies. A representative selection of manually curated sequences

Table 16 — Selected archaeal MMP sequences

Phylum *Euryarchaeota*		
Class *Halobacteria*		
Haladaptatus sp.	344 residues	GB WP_066144928 / UP A0A166SWC5
Halomicrobium mukohataei *(Haloarcula mukohataei)*	392 residues	UP C7NZX4
Class *Methanomicrobia*		
Methanocalculus sp.	177 residues	UP A0A101H2S1 *
Methanococcoides burtonii	231 residues	GB WP_011499923 / UP Q12UU6
Methanococcoides methylutens	233 residues	UP A0A099T1M7
Methanoculleus sp.	194 residues	UP A0A0Q1AIF2
Methanohalobium evestigatum	240 residues	GB WP_013195212 / UP D7EBD0
Methanohalophilus mahii	231 residues	GB WP_013038296 / UP D5E866
Methanolobus tindarius	241 residues	GB WP_023845349 / UP W9DPH7
Methanolobus psychrophilus	235 residues	GB WP_015053499 / UP K4MC96
Methanomicrobiales archaeon	177 residues	UP A0A117MHP9 *
Methanomethylovorans hollandica	242 residues	GB WP_015323622 / UP L0KWK3
Methanoregula formicica	179 residues	GB WP_015286272 / UP L0HJR6
Methanosalsum zhilinae *(Methanohalophilus zhilinae)*	227 residues	GB WP_013899163 / UP F7XME3
Methanosarcina acetivorans	211 residues	UP Q8TIZ9
Methanosarcina barkerii	217 residues	UP A0A0E3R7C9
Methanosarcina flavescens	217 residues	GB WP_054299817 / UP A0A0P6VKR0
Methanosarcina horonobensis	210 residues	UP A0A0E3SB37
Methanosarcina lacustris	215 residues	GB WP_048124635 / UP A0A0E3RZU1
Methanosarcina mazei *(Methanosarcina frisia)*	220 residues	GB KKG06863 / UP A0A0F8EZR8
Methanosarcina siciliae	211 residues	UP A0A0E3PQS9
Methanosarcina vacuolata	217 residues	UP A0A0E3LI77
Phylum *Thaumarchaeota*		
Class — / Order *Cenarchaeales*		

Cenarchaeum symbiosum	345 residues	UP A0RWX1
Class — / Order *Nitrosopumilales*		
Candidatus Nitrosoarchaeum koreensis	253 residues	UP F9CWY2
Candidatus Nitrosoarchaeum limnia	283 residues	GB WP_010190351 / UP S2E6F9
Candidatus Nitrosopumilus adriaticus	296 residues	UP A0A0D5C2T1
Candidatus Nitrosopumilus koreensis	238 residues	UP K0B4I4
Candidatus Nitrosopumilus salaria	269 residues	GB WP_008298808 / UP I3D381
Class — / Order *Nitrososphaeria*		
Candidatus Nitrososphaera evergladensis	251 residues	UP A0A075MRP9
Unclassified *Thaumarcheota*		
Candidatus Nitrosotalea devanaterra	423 residues	UP A0A128A209
Thaumarchaeota archaeon	540 residues	GB WP_048194249 / UP V6AQP1
Candidatus Nitrosopelagicus brevis	505 residues	UP A0A0A7V069
Unclassified *Archaea*		
Candidatus Pacearchaeota archaeon	301 residues	UP A0A1F6ZFA3

All sequence codes are from UniProt (UP; www.uniprot.org) or GenBank (GB; www.ncbi.nlm.nih.gov/genbank).
Sequence similarity searches were performed with karilysin CD (UP D0EM77) within UniProt using standard parameters. Only one representative from each species or genus (when sp.) was chosen. All sequences were manually curated.
Taxonomy according to the Catalogue of Life (http://www.catalogueoflife.org/col; Nagase et al., 2006) or, when absent, UniProt (www.uniprot.org).
* Identical sequences.
Searches completed on 23 January, 2017.

is provided in Tables 16 and 17. Inspection of the archaeal sequences reveals that they cluster to phyla Euryarchaeota and Thaumarchaeota, which populate the human digestive tract together with Crenarchaeota (Gaci et al., 2014). The healthy gut microbiome is also dominated in humans by the bacterial phyla Firmicutes and Bacteroidetes, with Actinobacteria, Proteobacteria and Verrucomicrobia present in smaller proportions (Lynch and Pedersen, 2016). Upstream in the gastrointestinal tract, six major bacterial phyla populate the oral microbiome: Firmicutes, Bacteroidetes, Proteobacteria, Actinobacteria, Spirochaetes, and Fusobacteria (Dewhirst et al., 2010). Bacterial species that potentially contain MMPs also belong to these phyla, as does karilysin, with the notable exception of a few bacterial sequences from the phyla Planctomycetes, an aquatic phylum present in brackish and fresh marine waters, and Cyanobacteria, which also inhabit waters and moist soils (Table 17). Accordingly, the distribution of archaeal and bacterial sequences is patchy and they are almost exclusively found in species of phyla highly represented in human microbiomes.

Table 17 — Selected bacterial MMP sequences

Phylum *Acidobacteria*		
Class *Solibacteres*		
Solibacter usitatus	442 residues	GB WP_011688765 / UP Q02908
Phylum *Actinobacteria*		
Class *Acidimicrobiia*		
Acidithrix ferrooxidans	358 residues	UP A0A0D8HKH4
Class *Actinobacteria*		
Pseudarthrobacter phenanthrenivorans (Arthrobacter phenanthrenivorans)	687 residues	GB WP_013599449 / UP F0M7V4
Pseudarthrobacter siccitolerans	686 residues	UP A0A024GZ92
Streptomyces lincolnensis	610 residues	GB WP_067438548 / UP A0A1B1MG04
Streptomyces scabiei	359 residues	UP A0A086GV39
Class - / Order *Actinomycetales*		
Alloactinosynnema sp.	556 residues	GB WP_054055428 / UP A0A0H5CVK8
Arthrobacter nitrophenolicus	634 residues	GB WP_009356397 / UP L8TT82
Knoellia aerolata	271 residues	GB WP_052113283 / UP A0A0A0JTC1
Pseudonocardia dioxanivorans	500 residues	GB WP_013678450 / UP F4CV22
Phylum *Bacteroides*		
Class *Bacteroidia*		
Parabacteroides distasonis	446 residues	UP A0A174VNZ8
*Tannerella forsythia (Bacteroides forsythus) **	472 residues	UP D0EM77
Class *Cytophagia*		
Cyclobacterium marinum (Flectobacillus marinus)	482 residues	GB WP_014022474 / UP G0J0B7
Phylum *Cyanobacteria*		
Class *Cyanophyceae*		
Microcystis aeruginosa	268 residues	GB WP_052734169 / UP A0A0F6U407
Synechococcus sp.	439 residues	UP B4WJ70
Class - / Order *Synechococcales*		
Acaryochloris marina	580 residues	GB WP_012164578 / UP B0CCT5
Phormidesmis priestleyi Ana	516 residues	UP A0A0P7ZK60
Phylum *Firmicutes*		
Class *Bacilli*		
Bacillus cereus	258 residues	GB WP_000775479 / UP R8XLT0
Bacillus mycoides	253 residues	UP A0A090ZAN4
Halobacillus halophilus (Sporosarcina halophila)	390 residues	GB WP_014642607 / UP I0JKN6
Lactobacillus acidifarinae	228 residues	GB WP_057801390 / UP A0A0R1LQY6
Lactobacillus acidophilus	222 residues	GB WP_003548046, GB YP_194247 / UP Q5FJA8

Lactobacillus apodemi	193 residues	GB WP_035459587 / UP A0A0R1TU85
Lactobacillus brevis	223 residues	GB WP_011666964 / UP Q03TZ3
Lactobacillus plantarum	266 residues	GB WP_003643352 / UP M4KPC6
Lactobacillus nodensis	240 residues	UP A0A0R1KCP7
Lactobacillus senmaizukei	238 residues	UP A0A0R2DS50
Lactobacillus sunkii	233 residues	GB WP_057825587 / UP A0A0R1KWI9
Leuconostoc kimchii	226 residues	GB WP_013103335 / UP D5T391
Paenibacillus terrae	280 residues	GB WP_044649112 / UP A0A0D7WTR0
Streptococcus suis	231 residues	GB WP_044681537 / UP A0A0Z8JY51
Phylum *Nitrospirae*		
Class *Nitrospira*		
Candidatus Magnetoovum chiemensis	426 residues	UP A0A0F2J1S3
Phylum *Plantomycetes*		
Class *Plantomycetia*		
Candidatus Brocadia fulgida	418 residues	UP A0A0M2UXB6
Candidatus Jettenia caeni	420 residues	GB WP_007220894 / UP I3IJ81
Candidatus Scalindua brodae	382 residues	UP A0A0B0EQJ8
Class *Plantomycetacia*		
Isosphaera pallida	435 residues	GB WP_013565370 / UP E8QY87
Pirellula staleyi (Pirella staleyi)	281 residues	GB WP_012911897 / UP D2R985
Planctomyces sp.	499 residues	GB WP_068412701 / UP A0A142YDV0
Singulisphaera acidiphila	538 residues	UP L0DHV3
Phylum *Proteobacteria*		
Class *Alphaproteobacteria*		
Aurantimonas sp.	321 residues	GB WP_055845852 / UP A0A0Q6EYY6
Aureimonas sp.	484 residues	GB WP_056689492 / UP A0A0Q6EFI8
Bradyrhizobium elkanii	454 residues	UP A0A1E3EYM1
Bradyrhizobium jicamae	318 residues	GB WP_057837259 / UP A0A0R3LGN7
Bradyrhizobium lablabi	324 residues	GB WP_057857189 / UP A0A0R3N9Q1
Hyphomicrobium nitrativorans	291 residues	GB WP_023788197 / UP V5SJS8
Labrenzia alexandrii (Stappia alexandrii)	378 residues	UP B9QV57
Methylobacterium sp.	470 residues	UP B0UC88
Puniceibacterium sp.	271 residues	GB WP_053078859 / UP A0A0J5QD38
Rhizobium sp.	404 residues	GB WP_062553422 / UP A0A0Q7YFP6
Rhodomicrobium udaipurense	248 residues	GB WP_037237908 / UP A0A037UYG6
Rhodospirillaceae bacterium	357 residues	UP A0A0F2S3S3
Sphingomonas changbaiensis	556 residues	UP A0A0E9MSC9
Sulfitobacter geojensis	501 residues	GB WP_064223228 / UP A0A196QSA6
Class *Betaproteobacteria*		
Burkholderia gladioli (Phytomonas marginata)	463 residues	UP A0A0M2QDA5

Burkholderia vietnamiensis	1065 residues	GB WP_059890570 / UP A0A132D6D9
Paraburkholderia glathei	1101 residues	UP A0A0J1D520
Rubrivivax gelatinosus	633 residues	UP I0HQ73
Class *Deltaproteobacteria*		
Chondromyces apiculatus	463 residues	UP A0A017SZ72
Chondromyces crocatus	433 residues	UP A0A0K1EA09
Labilithrix luteola	314 residues	UP A0A0K1Q883
Sandaracinus amylolyticus	588 residues	GB WP_053234042 / UP A0A0F6W3W9
Sorangium cellulosum (Polyangium cellulosum)	247 residues	UP A0A150T1A1
Stigmatella aurantiaca	477 residues	GB WP_002612142 / UP Q098Y1
Vulgatibacter incomptus	602 residues	GB WP_050725892 / UP A0A0K1PDK5
Class *Gammaproteobacteria*		
Acinetobacter sp.	205 residues	GB WP_005215616 / UP N9NAE2
Methylophaga aminisulfidivorans	237 residues	GB WP_007146704 / UP F5T331
Methylothermaceae bacteria	432 residues	UP A0A148N5W8
Pseudoalteromonas luteoviolacea	1466 residues	GB WP_065788537 / UP A0A1C0TT69
Candidatus Tenderia electrophaga	261 residues	UP A0A0S2TI48
Thiolapillus brandeum	387 residues	UP W0TNH9
Unclassified *Bacteria*		
Dadabacterium bacterium	302 residues	UP A0A0T5ZUN1
Latescibacteria bacterium	337 residues	UP A0A0S7XD53
Parcubacteria bacterium	310 residues	UP A0A0L0LDW4

All sequence codes are from UniProt (UP; www.uniprot.org) or GenBank (GB; www.ncbi.nlm.nih.gov/genbank).

Sequence similarity searches were performed with karilysin CD (UP D0EM77) within UniProt using standard parameters. Only one representative from each species or genus (when sp.) was chosen. All sequences were manually curated.

Taxonomy according to the Catalogue of Life (http://www.catalogueoflife.org/col; [1]) or, when absent, UniProt (www.uniprot.org).

* *T. forsythia* karilysin is the only bacterial MMP studied.

Searches completed on 27 January 2017.

3.3.3 CONCLUSIONS AND FUTURE PERSPECTIVES

MMPs are arguably the best studied MPs at the molecular, functional, physiological and structural levels, but most reports are restricted to humans and a few other animals. However, a comprehensive review of the literature and current sequence similarity searches revealed that validated and potential MMP CDs are widespread. In general, most proteins possess common ancestors in Eubacteria or Archaea, so their presence within the latter indicates that inheritance follows the Darwinian tree-based pathway or vertical descent model (Apic et al., 2001; Pal and Guda, 2006). This is the case for ~60% of human protein domains, which have their origins in these kingdoms and eukaryotic nodes before the metazoan era (Pal and Guda, 2006). However, some proteins originate at nodes that appeared later in evolution (Aravind and Subramanian, 1999), as reported for the large, multidomain pan-peptidase inhibitors of the α2-macroglobulin (α2M) family (Budd et al., 2004). These> 1000-residue proteins are widely distributed across metazoans but missing in all non-metazoan eukaryotic lineages. Unexpectedly, homologous proteins were found in several bacterial proteomes (Garcia-Ferrer et al., 2015; Goulas et al., 2017), but their distribution was patchy and incompatible with vertical descent from a common ancestral eubacterium. As most of these bacterial species encoding α2Ms exploited higher eukaryotes as hosts, either as pathogenic invaders or commensal colonizers, it was proposed that they were acquired by eukaryotic-toprokaryotic horizontal gene transfer (Budd et al., 2004) , similarly to some metabolic enzymes (Doolittle et al., 1990; Martin and Cerff, 1986).

MMPs are likewise widespread, perhaps across all kingdoms of life, where they are possibly involved in extracellular processing of proteins. However, they only show a homogenous gene distribution that is probably consistent with a vertical descent model within animals of the subkingdom Eumetazoa, as they are absent from more primitive metazoans. Within plants, they have only been found in higher plants. Here, the domain architecture is reminiscent of invertebrate MMPs, which suggests that plant and invertebrate MMPs are more closely related to each other than to vertebrate MMPs. This, in turn, hints that they could be modern representatives of an ancient MMP ancestor, common to the three groups (Fanjul-Fernández et al., 2010; Massova et al., 1998). In fungi, protists, viruses, bacteria and archaea, the presence of MMP sequences is reduced and patchy, which violates the vertical descent model. Generally, distribution is restricted to phyla with a lifecycle entailing intimate, direct or indirect, pathogenic or commensal, interaction with members of the subkingdom Eumetazoa. This suggests that MMPs in those kingdoms, like bacterial α2Ms, are xenologs coopted several times during evolution from eumetazoan hosts by independent horizontal gene transfer events (see e.g. Fig. 63), which would include uptake of mRNA by competence or abiotic mechanisms (Kotnik and Weaver, 2016), followed by subsequent spreading and polyplication within phyla.

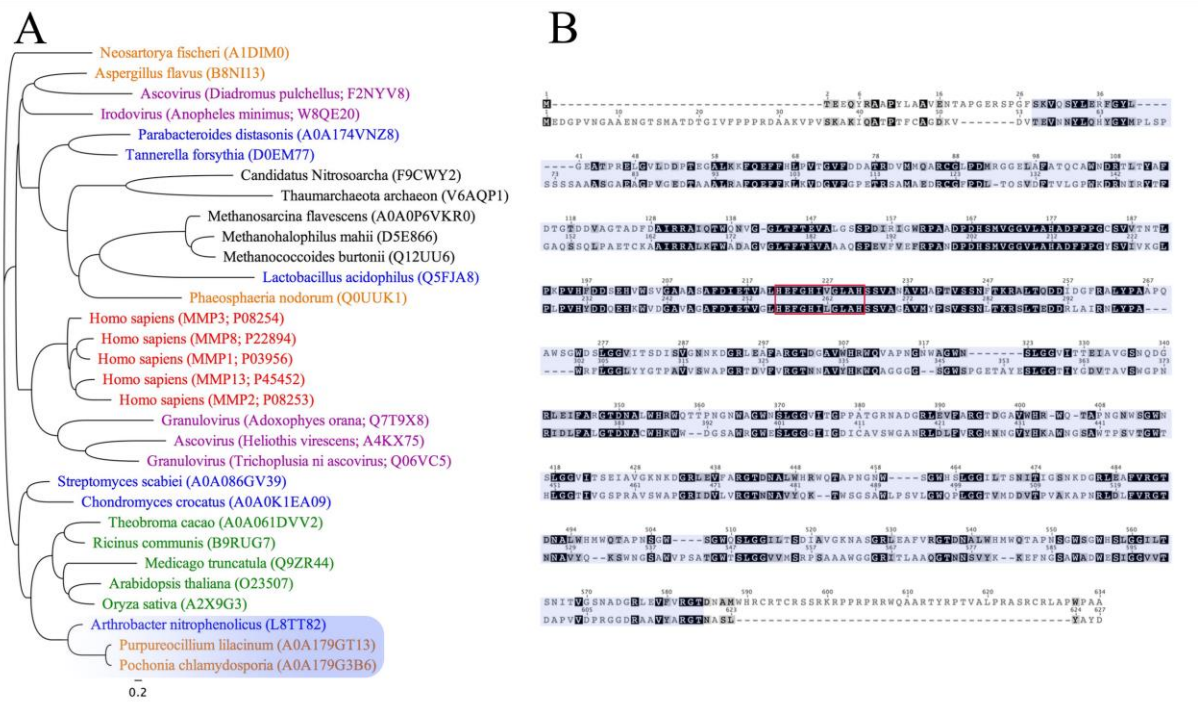


Figure 63: Phylogenetic studies. (A) Rooted phylogenetic tree reflecting evolutionary distances among selected protein sequences (in parenthesis the UniProt codes) of organisms from different kingdoms of life. Names in red are mammals, in green plants, in orange fungus, in lilac viruses, in blue bacteria and in black archaea. In blue background are highlighted proteins of different kingdoms (bacterial and fungal) that display high sequence similarity (see in B). This could be an indication of an eukaryotic-to-prokaryotic horizontal gene transfer event. The bar represents 0.2 substitutions per site. **(B)** Alignment of protein sequences from *Arthrobacter nitrophenolicus* (UniProt code: L8TT82) and *Purpureocillium lilacinum* (UniProt code: A0A179GT13) displaying more than 42% sequence identity in the blue backgrounded area. In red rectangle the zinc-binding motif (H-E-X-X-H-X-X-G/N-X-X-H/D) found in MMPs.

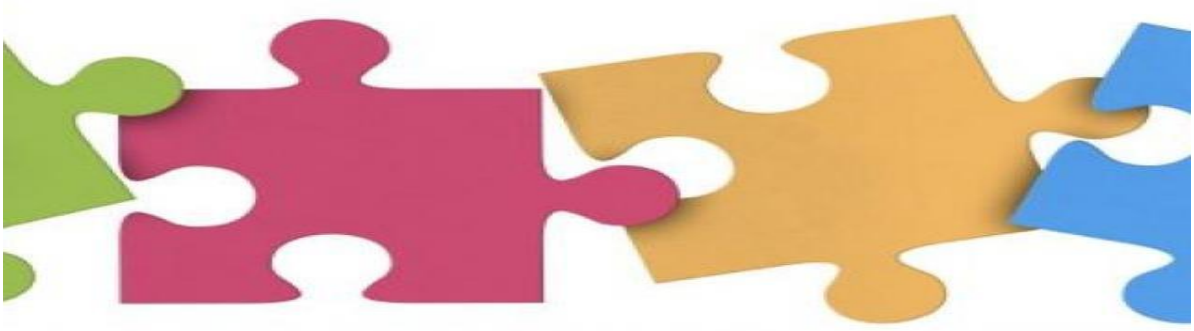


Figure reproduced from “Owips.com”

3. 4 Functional and structural studies on human α2-macroglobulin (hα2M) and interaction studies with recombinant G-related α2-macroglobulin binding protein (GRAB) and latent transforming growth factor-β2 (pro-TGF β2)

Marino-Puertas, L., Del Amo-Maestro, L., Taulés, M., Gomis-Rüth, F. X., & Goulas, T. (2019). Recombinant production of human α2-macroglobulin variants and interaction studies with recombinant G-related α2-macroglobulin binding protein and latent transforming growth factor-β2. *Scientific reports, 9*(1), 9186.

Del Amo-Maestro, L., Marino-Puertas, L., Goulas, T. & Gomis-Rüth, F. X. (2019). Recombinant production, purification, crystallization, and structure analysis of human transforming growth factor β2 in a new conformation. *Scientific reports, 9*(1), 8660. doi: 10.1038/s41598-019-44943-4. First co-author.

3.4.1 OBJECTIVES

The protein human α2-macroglobulin (hα2M) regulates key physiological processes that control the distribution and activity of many proteins, including peptidases, cytokines, hormones and other physiological effectors. Mediated by peptidase cleavage, they transit between native, intact forms and activated, induced forms. hα2Ms have been studied over decades using wild-type material from primary sources, which was limited by sample heterogeneity and contaminants.

To shed light on the molecular features of the intricate working mechanism of hα2M, we were interested in working on:

- The development of eukaryotic expression systems of hα2M variants and purification of the wild-type hα2M protein from plasma blood.
- The biochemical characterization of the recombinant hα2M variants.
- The determination and improving of the crystal structure of wild-type hα2M (native and induced) from blood and crystallization of the recombinant hα2M variants (alone or in protein complexes).
- The expression of the hα2M ligands (GRAB, pro-TGFβ2 and LRP1 receptor) in bacteria (*Escherichia coli*) and eukaryotic cells as *Drosophila* Schneider 2 or human Expi293F cells.
- The biochemical, biophysical and structural characterizations of the interactions between wild-type hα2M (native and induced forms) with GRAB, pro-TGFβ2 and LRP1.
- And finally, map down these interactions using the recombinant hα2M proteins forming complexes with GRAB, pro-TGFβ2 or LRP1.

3.4.2 RESULTS AND DISCUSSION

3.4.2.1 Analysis of hα2M sequence and generation of hα2M constructs

Human α2-macroglobulin gene (hα2M) encodes a very glycosylated protein of 1474-residues with an estimated molecular weight of 180 kDa per monomer. hα2M is generally found in blood as a 720 kDa homotetramer.

According to Uniprot (entry P01023), each 180 kDa monomer starts with a 23 residues-signal peptide for secretion and 1451 residues chain, which is formed by eleven domains. It is also stabilized by thirteen disulphide bonds (eleven intra-chain disulphide bonds and two inter-chain bonds) and eight glycosylations per monomer (Fig 64A).

hα2M was cloned from human c-DNA and the sequence was validated by sequencing with its internal primers (see Table 18).

Multiple constructs spanning different hα2M fragments were then generated and inserted into different expression vectors for testing protein expression, biochemical and biophysical characterizations and finally, crystallization (further details about these plasmids can be found in the "List of expression vectors" and in Table 19).

The used protein expression systems were:

- The insect expression system with transient protein expression, using an AKH signal peptide for extracellular expression into a pIEx vector.
- The mammalian expression system with transient protein expression, using an Ig κ ladder for secreted expression into a pCMV-Sport6 vector.

The generated constructs for hα2M were (Fig 64B) (see Table 19):

- The full-length hα2M (S^{24}-$N1^{473}$ and S^{27}-S^{1468} for insect and mammalian expressions, respectively) to characterize the nature of the thioester bond and study its interaction with GRAB and pro-TGFβ2.
- The N-terminal half of hα2M (S^{24}-E^{908} and S^{27}-A^{788} for insect and mammalian expressions, respectively) to study its role in the possible interactions: hα2M-GRAB and hα2M-pro-TGFβ2.

- The C-terminal half of hα2M (R^{787}-N^{1473} and F^{789}-S^{1468} for insect and mammalian expressions, respectively) to characterize the nature of the thioester bond and study its interaction with GRAB, pro-TGFβ2 and LRP1 receptor.
- The C-terminal half of hα2M without the RBD domain (F^{789}-E^{1337} for insect expressions) to remove a high flexible domain of the protein and facilitate the crystallization of this half hα2M fragment.
- The RBD domain of hα2M (P^{1340}-N^{1473} and P^{1340}-S^{1468} for insect and mammalian expressions, respectively) to assay its role in the interaction of hα2M with the LRP1 receptor.
- N/Q mutants in residues 55, 70, 247, 396, 410, 869, 991 and 1424 to test the biological importance of the glycosylations in hα2M and facilitate the crystallization of the hα2M fragments (see Table 20).

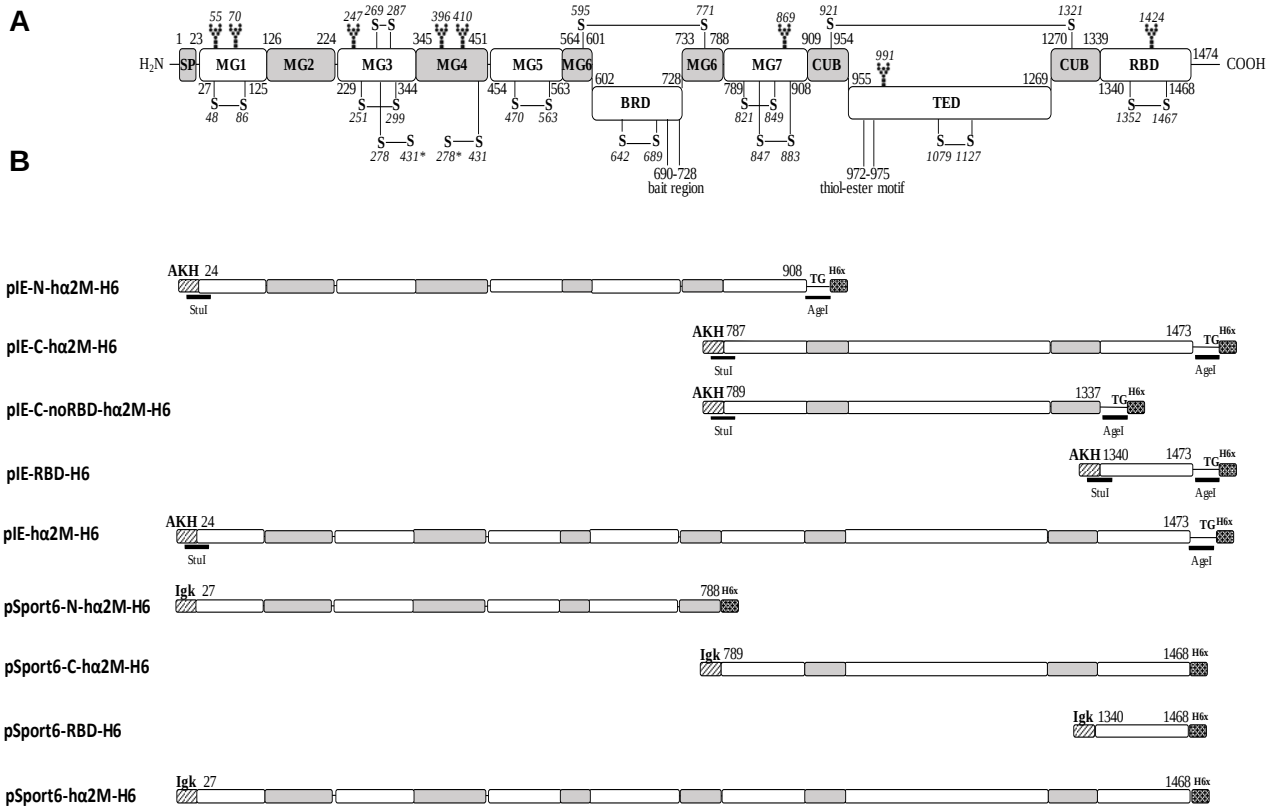


Figure 64. Overview of the hα2M and fusion proteins produced for expression in insect and mammalian cells. Scheme depicting the domain structure of hα2M **(A)** and the constructs studied **(B).** The residue numbers correspond to UP P01023. **(A)** Functional regions and domains are the signal peptide (SP); macroglobulin domains 1-to-7 (MG1-MG7); the bait-region domain (BRD); the CUB domain; the thioester domain (TED); and the receptor-binding domain (RBD). Disulphide bonds are shown in black and linked cysteines are labelled. An interchain disulphide is pinpointed with an asterisk and N-linked glycosylation sites are highlighted with a sugar chain. **(B)** hα2M fusion proteins produced with plasmids pIEx and pCMV-Sport6. The AKH signal peptide sequence, the mouse Ig κ-chain ladder sequence, His6×-tags and restriction sites are graphically represented.

Table 18. Internal sequencing primers of hα2M and sequencing primers of the main expression vectors			
Primer name	**Parental DNA**	**Primer sequence (from 5' to 3')**	**Comments**
α2M_seq_fwd 300	hα2M c-DNA	ATGTTCCTCACTGTCCAA	Forward primer to sequence the full-length and N-terminal fragments of hα2M DNA. Primer annealing in the MG1 domain of hα2M protein.
α2M _rev 400	hα2M c-DNA	TGTCTGGACAAAGACCAG	Reverse primer to sequence the N-terminal end of the full-length and N-terminal fragments of hα2M DNA. Primer annealing in the MG1 domain of hα2M protein.
α2M _fwd 875	hα2M c-DNA	ACAGCTAAACAGCCATGG	Forward primer to sequence the full-length and N-terminal fragments of hα2M DNA. Primer annealing in the MG3 domain of hα2M protein.
α2M _fwd 1255	hα2M c-DNA	CTTACTGTTAGGGTCAATTAC	Forward primer to sequence the full-length and N-terminal fragments of hα2M DNA. Primer annealing in the MG4 domain of hα2M protein.
α2M _seq_fwd 1630	hα2M c-DNA	GCTGTTTTACCTACCGGG	Forward primer to sequence the full-length and N-terminal fragments of hα2M DNA. Primer annealing in the MG5 domain of hα2M protein.
α2M _seq_fwd 2416	hα2M c-DNA	CCTTGAGTGTGAAGGCCT	Forward primer to sequence the full-length, N-terminal (cloned into pIEx vector), C-terminal and C-terminal without the RBD domain fragments of hα2M DNA. Primer annealing in the MG7 domain of hα2M protein.
α2M _fwd 3125	hα2M c-DNA	AGCCTTTGTTCTGAAGAC	Forward primer to sequence the full-length, C-terminal and C-terminal without the RBD domain fragments of hα2M DNA. Primer annealing in the TED domain of hα2M protein.
α2M _fwd 3501	hα2M c-DNA	CAAGTCACTTAATGAGGAAGC	Forward primer to sequence the full-length, C-terminal and C-terminal without the RBD domain fragments of hα2M DNA. Primer annealing in the TED domain of hα2M protein.
α2M _fwd 3875	hα2M c-DNA	GGACAACAACAACCGCC	Forward primer to sequence the full-length, C-terminal and C-terminal without the RBD domain fragments of hα2M DNA. Primer annealing in the CUB domain of hα2M protein.
α2M _fwd 4375	hα2M c-DNA	GCAATTGCTGAGTACAAT	Forward primer to sequence the full-length, C-terminal and RBD domain fragments of hα2M DNA. Primer annealing in the RBD domain of hα2M protein.
IE1_fwd	pIEx vector	TGGATATTGTTTCAGTTGCAAG	Forward primer to sequence the N-terminal end of the clonings performed and inserted into pIEx vector.
AS-Stag_rev	pIEx vector	GTCCATGTGCTGGCGTTC	Reverse primer to sequence the C-terminal end of the clonings performed and inserted into pIEx vector.
pIEx5_rev	pIEx vector	TTATTAGTGATGGTGATG	Reverse primer to sequence the C-terminal end of the clonings performed and inserted into pIEx vector.
CMV_fwd	pCMV-Sport 6	CGCAAATGGGCGGTAGGCGTG	Forward primer to sequence the N-terminal end of the clonings performed and inserted into pCMV-SPORT 6 vector.
M13_fwd	pCMV-Sport 6	TGTAAAACGACGGCCAGT	Reverse primer to sequence the C-terminal end of the clonings performed and inserted into pCMV-SPORT 6 vector.
Further information and feature maps of the vectors in the section "List of expression vectors".			

Table 19. Constructs, primers and plasmids of hα2M for expression in insect and mammalian cells							
Plasmid name	**Protein**	**Parental DNA**	**Forward -primer ***	**Reverse-primer ***	**Protein sequence ****	**Tags *****	**Comments**
pIE-hα_2M-H6	hα_2M	Human c-DNA pIEx vector	CATTAGGCCTCAGTCTCTG GAAAACCGCAGTATATG	CATTACCGGTAGGATTTCC AAGATCTTTG	S^{24}-N^{1473}+**PTG+H_{6x}**	C-t H_{6x}	Full-length hα_2M in S2 cells. The gene was inserted by directional cloning (between *Stu*I and *Age*I) into the pIEx vector in frame with the AKH signal peptide sequence. Partial digestion with Stu I (there are two restriction sites for StuI into the sequence).
pIE-N-hα_2M-H6	N-hα_2M	pIE-hα_2M-H6	GATCTTGGAAATCCTACCG GTCATCATCAC	CAATACCGGTTTCAGGTTC AACCAACAGAG	S^{24}-E^{908}+**TG+H_{6x}**	C-t H_{6x}	As above for the N-terminal half of hα_2M, but with only digestion with AgeI and then ligation into the already digested pIEx vector.
pIE-C-hα_2M-H6	C-hα_2M	pIE-hα_2M-H6	CCCGAGCCTTCCAGCCCTT CTTTGTG	CATTACCGGTAGGATTTCC AAGATCTTTG	R^{787}- N^{1473}+**PTG+H_{6x}**	C-t H_{6x}	As above for the C-terminal half of hα_2M, but with only digestion with AgeI and then ligation into the already digested pIEx vector.
pIE- C-hα_2M-no RBD-H6	C-hα_2M-No RBD	pIE-C-hα_2M-H6	GCTTTCGTCATCATCGCTG AGGCCTTCCAGCCCTTCTT TGTGGAGCTCA	ATGGTGATGGTGATGATG ACCGGTTTCCTTTTCTGGG AGAATATTGTA	F^{789}-E^{1337}+**TG+H_{6x}**	C-t H_{6x}	C-hα_2M without the RBD domain in S2 cells. The gene was inserted by restriction-free cloning into the pIEx vector in frame with the AKH signal peptide sequence.
pIE-RBD-H6	RBD	pIE-hα_2M-H6	CATTAGGCCTCACCCTTTG CTTTAGGAGTGC	CATTACCGGTAGGATTTCC AAGATCTTTG	P^{1340}-N^{1473}+**PTG+H_{6x}**	C-t H_{6x}	As above for the full-length hα_2M.
pS6-hα_2M-H6	hα_2M	pIE-hα_2M-H6 pCMV-Sport 6 vector	TGGGTTCCAGGTTCCACTG GTGACGGAAAACCGCAGT ATATGGTTCTG	CGCCTAATGGTGATGGTG ATGGTGGCTGCAAGGAGC ATTGTACTCAGC	S^{27}-S^{1468}+**H_{6x}**	C-t H_{6x}	Full-length hα_2M in Expi293F cells. The gene was inserted by restriction-free cloning into the pCMV-Sport 6 vector in frame with the Ig κ leader sequence.
pS6-N-hα_2M-H6	N-hα_2M	pIE-N-hα_2M-H6 pCMV-Sport 6 vector	TGGGTTCCAGGTTCCACTG GTGACGGAAAACCGCAGT ATATGGTTCTG	CGCCTAATGGTGATGGTG ATGGTGGGCTCGGAGAGA GGCAGTGGAAGA	S^{27}-A^{788}+**H_{6x}**	C-t H_{6x}	As above for the N-terminal half of hα_2M cloned into the pCMV-Sport 6 vector.
pS6-C-hα_2M-H6	C-hα_2M	pIE-C-hα_2M-H6 pCMV-Sport 6 vector	TGGGTTCCAGGTTCCACTG GTGACTTCCAGCCCTTCTT TGTGGAGCTC	CGCCTAATGGTGATGGTG ATGGTGGCTGCAAGGAGC ATTGTACTCAGC	F^{789}-S^{1468}+**H_{6x}**	C-t H_{6x}	As above for the C-terminal half of hα_2M cloned into the pCMV-Sport 6 vector.

pS6-RBD-H6	RBD	pS6-hα_2M-H6 pCMV-Sport 6 vector	TGGGTTCCAGGTTCCACTGGTGACCCCTTTGCTTTAGGAGTGCAGACT	CgCCTAATGGTGATGGTGATGGTGGCTGCAAGGAGCATTGTACTCAGC	P^{1340}-S^{1468}+$\mathbf{H_{6x}}$	C-t H_{6x}	As above for the RBD domain of hα_2M cloned into the pCMV-Sport 6 vector.

* Restriction-site sequences and overhangs for restriction-free cloning are underlined.

** Peptide sequence of the expressed protein after fusion-tag removal. Amino acids derived from the construct are in bold. See also Fig. 64

*** Fused tags at the carboxy-terminus (C-t).

Table 20. Mutations, parental DNA, proteins and primers to mutate the glycosylation sites of hα2M				
Mutation	**Forward -primer ***	**Reverse-primer ***	**Parental DNA**	**Protein ****
55 N/Q	CTGAGCTACCTG**CAG**GAG ACAGTGAC	GTCACTGTCTC**CTG**CAGGTAGC TCAG	pIE-hα2M-H6	
			pIE-N-hα2M-H6	
70 N/Q	TCTGTCAGGGGA**CAG**AGG AGCCTC	GAGGCTCCT**CTG**TCCCCTGACA GA	pIE-hα2M-H6	
			pIE-N-hα2M-H6	
247 N/Q	TTGGAAGAAGAGATG**CAG** GTATCAGTGT	ACACTGATAC**CTG**CATCTCTTC TTCCAA	pIE-hα2M-H6	
			pIE-N-hα2M-H6	
396 N/Q	AAACTATTACTCC**CAG**GCT ACCACGGAT	ATCCGTGGTAGC**CTG**GGAGTA ATAGTT T	pIE-hα2M-H6	
			pIE-N-hα2M-H6	
410 N/Q	CAGTTCTCTATC**CAG**ACCA CCAATGTT	AACATTGGTGGT**CTG**GATAGA GAACTG	pIE-hα2M-H6	
			pIE-N-hα2M-H6	
869 N/Q	TCATTAGGAAATGTG**CAG** TTCACTGTGAG	CTCACAGTGAA**CTG**CACATTTC CTAA TGA	pIE-hα2M-H6	
			pIE-N-hα2M-H6	
			pIE-C-hα2M/C- (no RBD)-H6	
991 N/Q	GTACTGGATTATCTA**CAG**G AAACACAGCAG	CTGCTGTGTTTC**CTG**TAGATAA TCCAGTAC	pIE-hα2M-H6	
			pIE-C-hα2M-H6	
			pIE-C-hα2M/C- (no RBD)-H6	
1424 N/Q	GATAAGGTGTCA**CAG**CAG ACACTGAGC	GCTCAGTGTCTG**CTG**TGACACC TTATC	pIE-hα2M-H6	
			pIE-C-hα2M-H6	
hα2M without glycosylations	N-α2M without glycosylations	C-α2M without glycosylations	C-α2M -no RBD without glycosylations	
Cloning by site directed mutagenesis with overlapping primers				
* Substituted sequence to mutate the residue is underlined and in bold ** Several mutations were performed sequentially to get a final protein totally deglycosylated				

Table 20. Deglycosylated mutants of hα2M fragments and full-length protein.

3.4.2.2 Expression of hα2M fusion-proteins

We designed several short constructs of hα2M in order to improve the current X-ray structure which is proved to be a complex process due to difficulties arising from sample heterogeneity owing to glycosylation but also from the large size and intrinsic flexibility of the molecule (Goulas et al., 2014b).

We were also interested in studying the biochemistry of the thioester bond and in the mapping of the studied protein interactions.

The final yield of pure recombinant hα2M proteins after all the purification and polishing steps are shown in the Fig. 65 and Table 21, respectively. Further details of the protein purifications are going to be explained bellow and in Material and Methods (M&M).

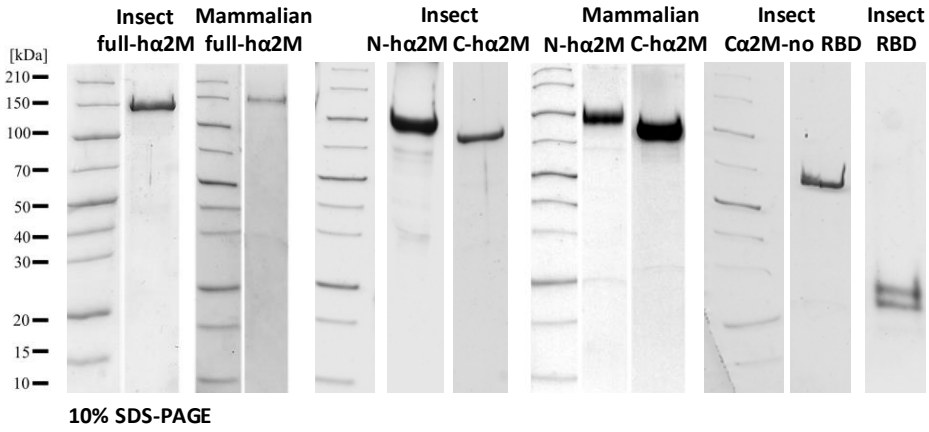


Figure 65. Protein production and purification of hα2M fusion-proteins. 10% SDS-PAGE analysis of recombinant hα2M proteins expressed in insect and mammalian cells.

Proteins	Expressed in Drosophila S2 cells	Expressed in human Expi293F cells
Full hα2M	~1.0 mg/L	~0.4 mg/L
N-hα2M	~1.6 mg/L	~3.2 mg/L
C-hα2M	~1.3 mg/L	~4.6 mg/L
C-hα2M-no RBD	~1.2 mg/L	
RBD	~0.9 mg/L	

Table 21. Protein yields per expression liter of pure recombinant hα2M proteins in the corresponding cell expression system (See Fig. 64 for the fusion proteins). Proteins expressed in *Drosophila* S2 cells, were cloned into a pIEx vector, and in Expi293F cells, were cloned into a CMV-Sport6 vector.

3.4.2.2.1 Overexpression in insect cells

All the protein expressions in insect cells, were performed in *Drosophila* S2 cells adapted to grow in suspension in Sf-900™ II SFM culture medium (Gibco). S2 cells were transfected with the corresponding pIEx construct at 15 × 10^6 cells/mL with linear 25 kDa polyethylenimine (PEI). Protein expressions were harvested after seven days and further collected for protein purification.

After all the purification and polishing steps, the protein yield for full-hα2M was ~1.0 mg of pure protein per liter of expression medium (Table 21) and the protein migrated as a tetramer of ~690 β

according to SEC (Fig. 66A). Expression of N-hα2M yielded ~1.6 mg per liter of insect cell culture (Table 21) and formed a dimer of ~170 kDa due to the presence of an intermolecular disulphide bond (C^{278}–C^{431}), which is also required for dimerization of the full-length protein (Fig. 64A). Consistently, the protein migrated as a monomer of ~85 kDa in the presence of reducing agents. Thus, several SEC conditions (extreme pH and reducing agents as 1,4-dithiothreitol, DTT, and tris[2-carboxyethyl]phosphine, TCEP) were assayed to break the disulphate bond and get only the 85 KDa monomer for the crystallization trials, however, the protein was found degraded (Fig. 66B). In turn, C-hα2M was monomeric (~75 kDa) (Fig. 66C) and yielded ~1.3 mg per liter of insect cell culture (Table 21). C--hα2M without the RBD domain (C-α2M no RBD) migrated as a monomer of ~70 kDa (Fig. 66D) and yielded ~1.2 mg per liter of insect cell culture (Table 21). However, this protein was very instable and very fast degraded. Expression of RBD yielded ~0.9 mg per liter of insect cell culture (Table 21) and was the half of this protein unfolded (Fig. 66E).

The corresponding deglycosylated hα2M fusion-proteins (N/Q in residues 55, 70, 247, 396, 410, 869, 991 and 1424) were also expressed in *Drosophila* insect cells to facilitate the crystallization of the hα2M fragments (see Table 10). The yield of these proteins was very low, and they were extremely unstable, meaning that the glycosylations were fundamental to keep soluble and folded the full-length hα2M protein and fragments (Fig. 67).

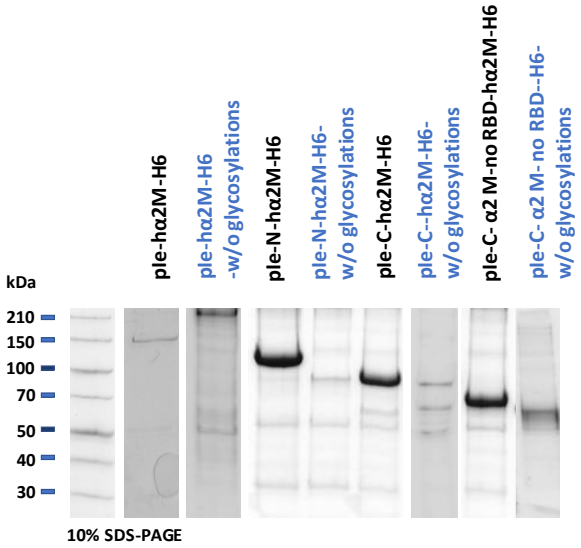
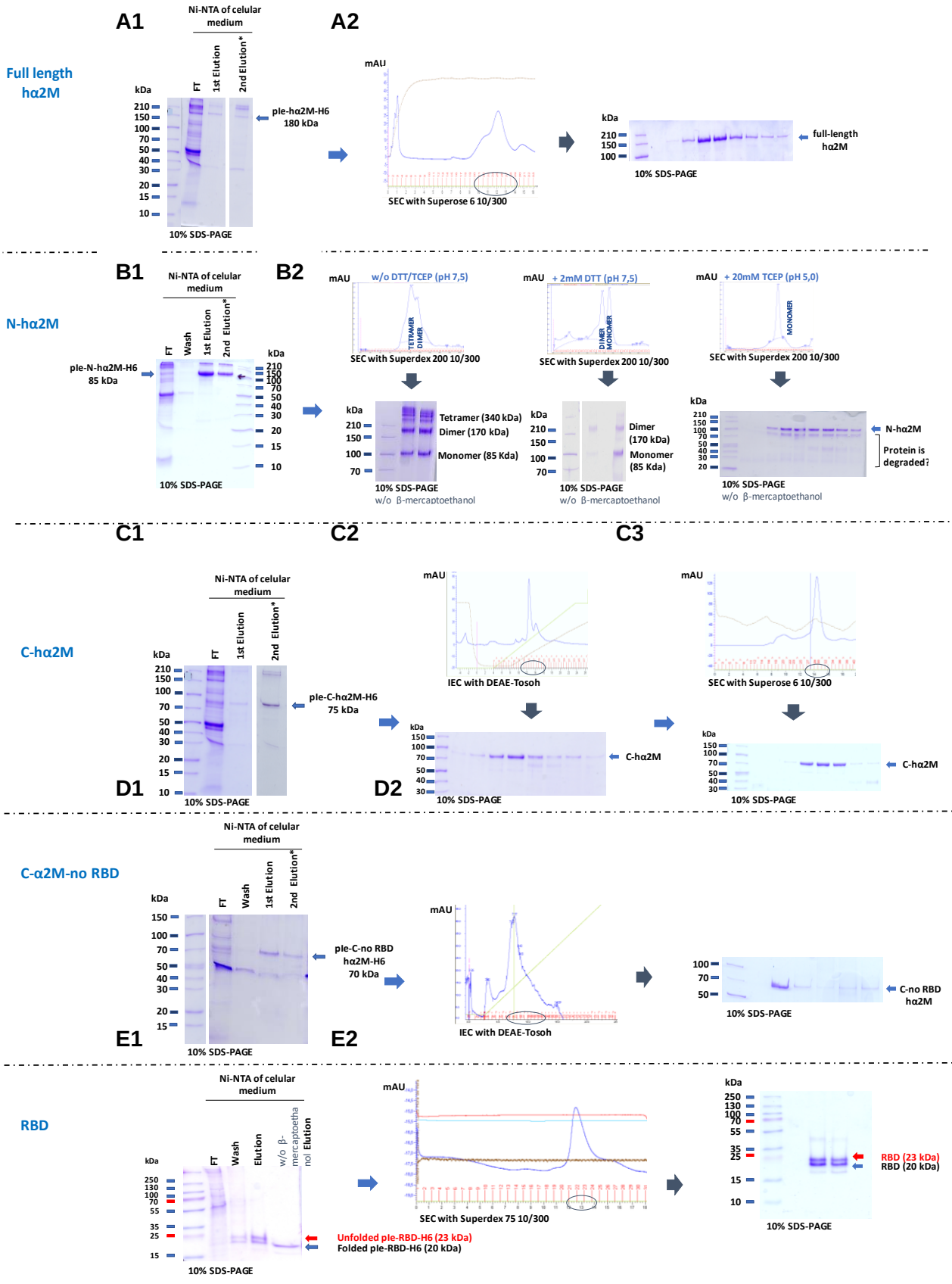


Figure 67. Recombinant protein production in insect cells and purification of the deglycosylated mutants of hα2M. The deglycosylated mutants of hα2M are indicated in blue and their corresponding wild type hα2M fusion protein are in black.

Figure 66. Overexpression and purification of human α2-macroglobulin fusion-proteins in *Drosophila* Schneider 2 Cells (pIE-H6 constructs). (A) Overexpression of hα2M. **(A1)** Purification by affinity and **(A2)** size-exclusion chromatographies. **(B)** Overexpression of N-hα2M. **(B1)** Purification by affinity and **(B2)** size-exclusion chromatographies with different reducing agents into the SEC running buffer to break the disulphate bond. **(C)** Overexpression of C-hα2M. **(C1)** Purification by affinity, **(C2)** ionic-exchange and **(C3)** size-exclusion chromatographies. **(D)** Overexpression of pIE-C-no RBD . **(D1)** Purification by affinity and **(D2)** ionic exchange chromatographies. **(E)** Overexpression of RBD. **(E1)** Purification by affinity and **(E2)** size-

exclusion chromatographies. The second elution corresponds to the second nickel resin incubation and it is marked with as asterisk. SDS-PAGE and samples without β-mercaptoethanol are also indicated.



3.4.2.2.2 Overexpression in mammalian cells

We next developed a transient expression system based on Expi293F cells, which derived from the HEK293 human embryonic kidney cell line and were grown in suspension in high density and maintained in FreeStyle™ F17 expression medium (Gibco). Expi293F cells were transfected with the corresponding pCMV-Sport6 construct at 1 × 10^6 cells/mL with linear 25 kDa polyethylenimine (PEI). Protein expressions were harvested after three days and further collected for protein purification. Cells were removed by centrifugation and the supernatants were used for subsequent purification steps.

The proteins produced in the Expi293F cell system were easily purified and the elutions from the affinity chromatographies much cleaner than in the insect expression system (Fig. 68). Thus, an IEC was most of the times enough to get pure protein. Protein yields were also higher using this protein expression system (Table 21). In this case ,a treatment with kifunensine, which is an inhibitor of class I glycoprotein-processing α-mannosidases (Chang et al., 2007; Elbein et al., 1990), was performed to remove the glycosylations (Fig. 68B).

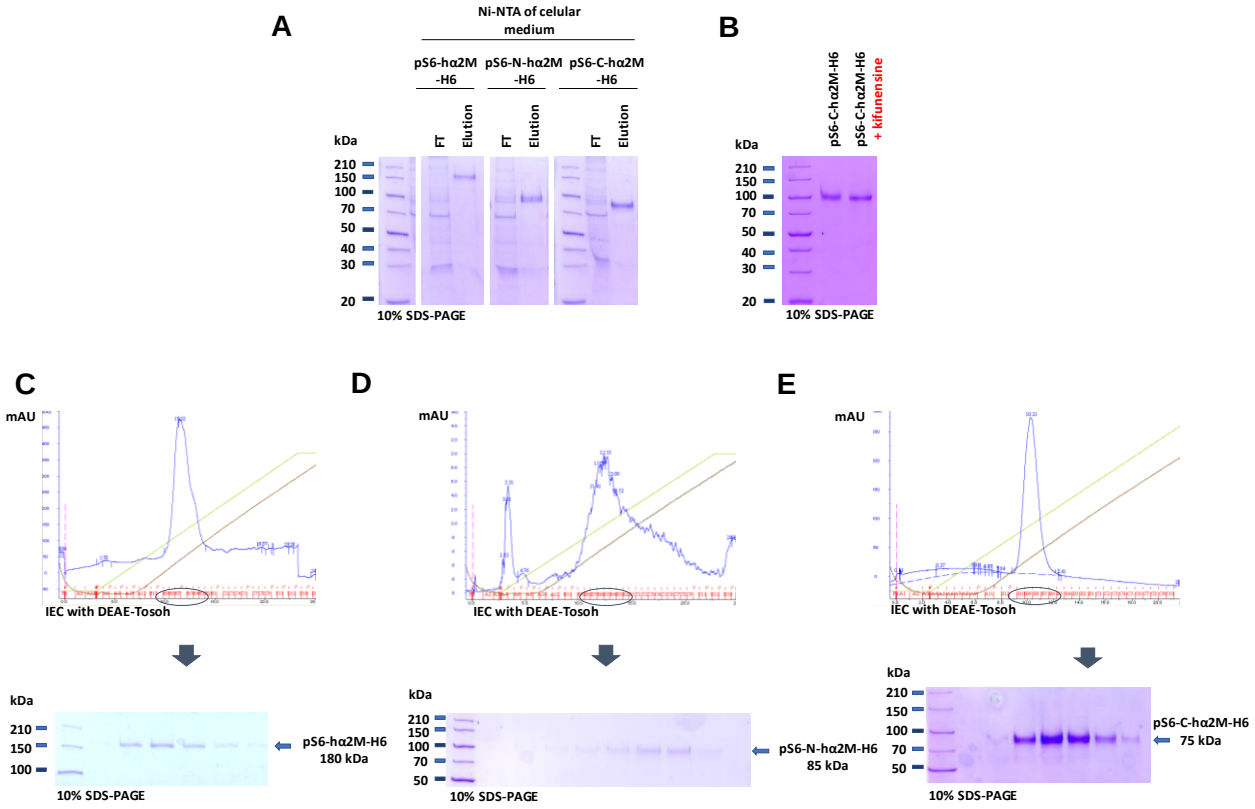


Figure 68. Overexpression and purification of hα2M fusion-proteins in human EXPI293F cells (pCMV-Sport6-H6 constructs). **(A)** Protein purification by affinity chromatography (Ni-NTA). **(B)** Deglycosylation treatment by kifunensine of C-hα2M. **(C)** Anion exchange chromatography of hα2M. **(D)** The same as (C) for N-hα2M. **(E)** The same as (C) for C-hα2M.

3.4.2.3 Analysis of the sequence of hα2M ligands and generation of their constructs

We were interested in studying the interactions of hα2M with three proteins: GRAB, pro-TGFβ2 and LRP1 receptor.

GRAB

The human pathogen *Streptococcus pyogenes* (group A *Streptococcus*) expresses a surface protein called GRAB (G-related α2M binding protein) with 217 residues (UP, entry Q7DAL7). This 22.8 kDa protein has a signal peptide of 33-residues, a highly conserved N-terminal domain A responsible for the interaction with hα2M (where Arg^{42} and, specially, Arg^{64} play a critical role for the binding) (Godehardt et al., 2004), a variable number of 28-residue repeats, a cell-wall attachment site with the Gram-positive membrane anchor motif ($L^{183}PTTG^{187}$) and the membrane anchor domain (Fig. 69A1) (Godehardt et al., 2004). Mature protein of GRAB is predicted to be from 34 to 187 residues and have a molecular mass of 15.8 kDa, but apparently it migrates much slower in SDS-PAGE. (Rasmusen et al., 1999) (Fig. 70).

GRAB was cloned from a synthetic gene, optimized for expression in *Escherichia coli* and inserted into the pCri8a vector (Goulas et al., 2014a). Two constructs were generated with two different affinity tags (6x His-tag for Ni-NTA purification and 2x Strep-tag for the SPR assays) (Fig. 69A2) (Table 22)..

Pro-TGFβ2

Human TGFβ2 (Transforming growth factor β2 proprotein) is biosynthesized as a latent 394-residue precursor (pro-TGFβ2) (UP, entry P61812), arranged as a glycosylated homodimer of 91 kDa containing three glycan chains attached to each LAP. Peptidolytic activation at bond R^{302}-A^{303} (De Martin et al., 1987; Marquardt et al., 1987) produces the mature TGFβ2, which spans 112 residues and is arranged as a disulphide-linked 25 kDa homodimer noncovalently associated with the respective LAP moieties (Del Amo-Maestro et al., 2019) (Fig. 69B1 and 70).

Pro-TGFβ2 was cloned from human c-DNA, optimized for insect and mammalian expressions (the endogenous signal peptide was replaced for AKH and Ig κ signal peptides, respectively) and inserted into pIEx and CMV-Sport6 vectors for insect and mammalian transient expressions, respectively. Four constructs were generated with two different affinity tags (6x or 8x His-tag for Ni-NTA purification in insect and mammalian expressions, respectively, and 2x Strep-tag for the surface plasmon resonance (SPR) assays (Fig. 69B2) (Table 22).

LRP1

The human LRP1 (Pro-low-density lipoprotein receptor-related protein 1) is a 600 kDa endocytic receptor involved in endocytosis and in phagocytosis of apoptotic cells. The ligand binding domains of LRP1 are called complement-like repeats (CR). Each CR is constituted by 40-42 residues, with three disulphide bonds and binds one calcium ion (Daly et al., 1995) (Fig. 69C1).These disulphide bonds are critical for the correct fold and function of the protein (Dolmer and Gettins, 2006). It has been also reported that hα2M has high affinity for CR3-4-5 (more than for isolate CR3 or CR3-4 and CR5-6-7) in cluster 2 (Dolmer and Gettins, 2006), being probably the binding site for RBD on human LRP1 receptor.

LRP1 was cloned from human c-DNA, optimized for bacteria, insect and mammalian expressions and inserted into pCri8a, pIEx and CMV-Sport6 vectors. Five constructs were generated with 6x His-tag for Ni-NTA purification and in the case of pCri vector, with a TEV cleavage as well (Fig. 69C2 and 70) (Table 22).

3.4.2.4 Overexpression of hα2M ligands

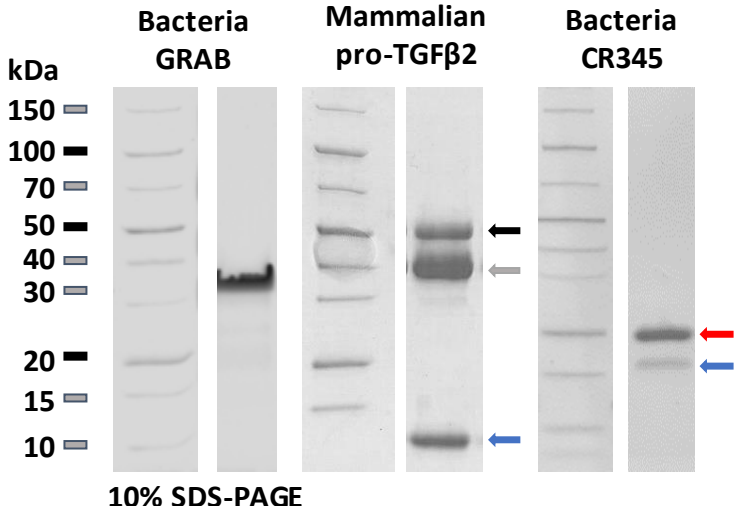


Figure 70. Recombinant protein production and purification of hα2M ligands. 10% SDS-PAGE analysis of recombinant mature GRAB (~15.8 kDa) expressed in *E.coli*, of pro-TGFβ2 (arrows indicate pro-TGFβ2 in black (~50 kDa), LAP in grey (~40 kDa) and mature TGFβ2 in blue (~12.5 kDa)) expressed in Expi293F cells and CR345 (20 and 17 kDa bands, unfolded in red arrow and folded CR345 in blue arrow, respectively) expressed in *E.coli*.

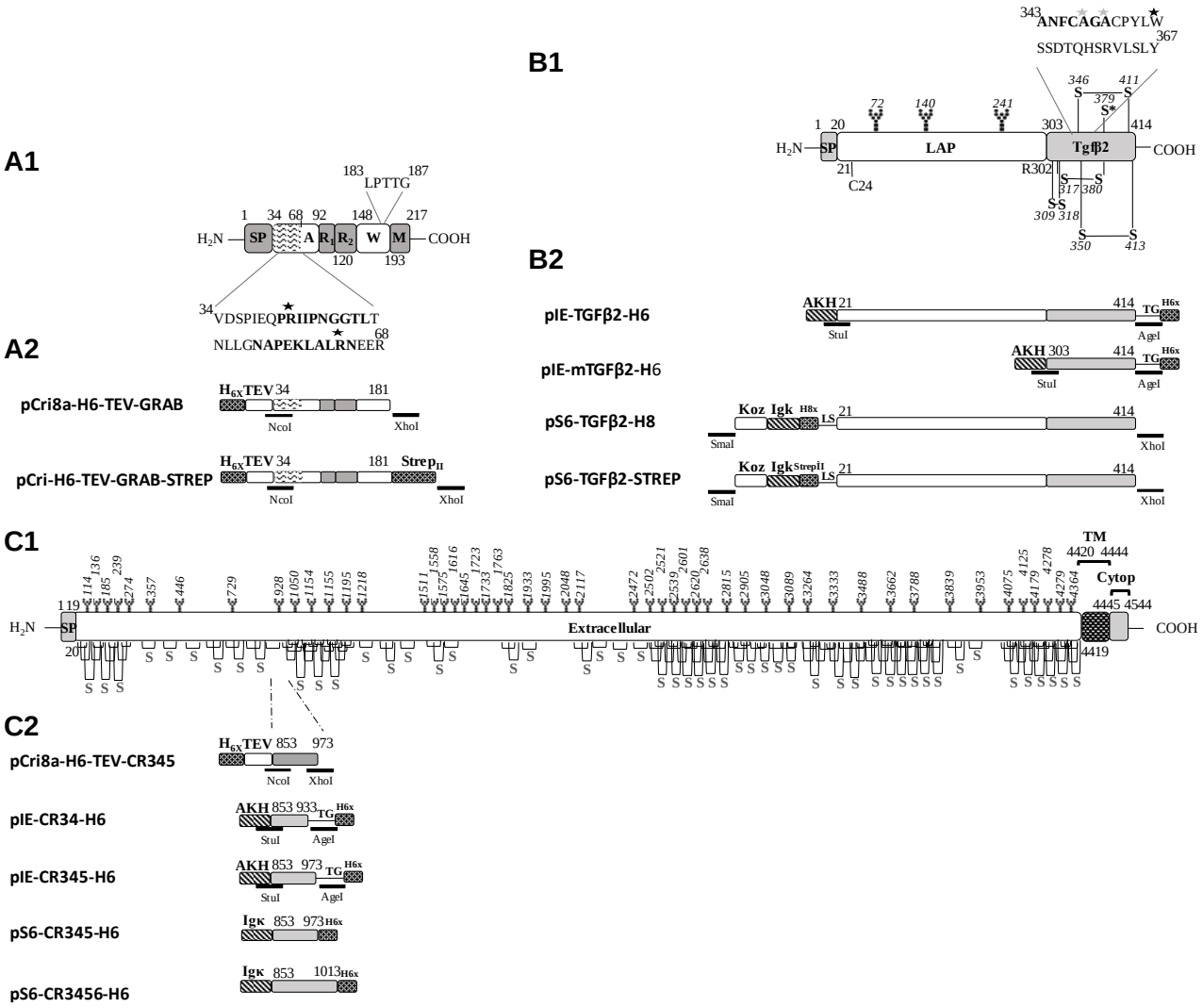


Figure 69. Overview of studied ligands of the hα2M and fusion proteins produced for their expression in bacteria, insect and mammalian cells. (A1) Scheme depicting the domain structure of GRAB **(A2)** and the constructs studied**.** The residue numbers correspond to UP Q7DAL7. **(A1)** Functional regions and domains are the signal peptide (SP); domain A, with the binding regions of hα2M hatched and in the inset; repeat regions R1 and R2; the cell-wall attachment site (W), with the cell-wall anchor motif shown in magnification; and the membrane anchor (M). Critical arginine residues for hα2M-binding are indicated by a star (R42 and R64). **(A2)** GRAB fusion proteins in pCri8a with His6-tag, TEV site and Strep-tag. **(B)** Same as (A) for pro-TGFβ2 (UP P61812). **(B1)** Functional domains and regions are the SP; the latency associated peptide (LAP); and the mature growth factor moiety (TGFβ2). Critical residues in LAP are C24, which is involved in the binding of LTBP, and R302, required for furin cleavage. Mature TGFβ2 segment A343-Y367 is involved in hα2M binding, important and critical residues are indicated by a grey (A347 and A349) and a black star (W354), respectively. Disulphide bonds are shown in black and linked cysteines are labelled. An interchain disulphide is pinpointed with an asterisk and N-linked glycosylation sites are highlighted with a sugar chain. **(B2)** Human pro-TGFβ2 fusion proteins produced with plasmids pIEx and pCMV-Sport6. The AKH SP, the Kozac (Koz), the mouse Igκ-chain SP, affinity tags (His6 and Strep) and restriction sites are graphically represented. **(C)** Same as (A) for pro-LRP1 (UP Q07954). **(C1)** Functional domains and regions are the SP, the extracellular topological domain, the transmembrane domain (TM) and the cytoplasmatic topological domain. **(C2)** Human pro-LRP1 fusion proteins produced with plasmids pCri8a, pIEx and pCMV-Sport6. The AKH SP, the mouse Ig κ-SP, the TEV cleavage, His6x-affinity tags and restriction sites are graphically represented. Disulphide bonds are shown in black and linked cysteines are labelled and N-linked glycosylation sites are highlighted with a sugar chain.

Table 22. Constructs, primers and plasmids of hα2M ligands for expression in bacteria, insect and mammalian cells							
Plasmid name	**Protein**	**Parental DNA**	**Forward -primer ***	**Reverse-primer ***	**Protein sequence ****	**Tags *****	**Comments**
pCri8a-H6-TEV-GRAB	GRAB	Synthetic full-length GRAB DNA pCri8a vector	CAATCCATGGTTGATAGCCCG ATTG	CAATCTCGAGTTAATTAACGTTC TGACGTT	**GAM**+V^{34}-N^{181}	N-t H_{6x}+TEV	Synthetic gene of GRAB optimized for expression in *Escherichia coli* inserted into the pCri8a vector (Goulas et al., 2014a) by directional cloning between the *Nco*I and *Xho*I restriction sites.
pCri8a-H6-TEV-GRAB-STREP	GRAB	pCri-H6-TEV-GRAB	ATGCCCATGGTTGATAGCCCG	CGAATTGTGGATGGCTCCAACCT CCATTAACGTTCTGACGTTC; CTTCCACCTCCAGAACCTCCACCC TTTTCGAATTGTGGATGGCTCC; GTGGATGGCTCCATGCGCTACCT CCACTTCCACCTCCAGAACC; GCATCTCGAGTTACTTTTCGAATT GTGGATGGCTCCATGCGC	**GAM**+V^{34}-N^{181}+**GGWSHPQ FEKGGGSGGGS GGSAWSHPQFE K**	N-t H_{6x}+TEV-(protein)-Strep	This construct was obtained from pCri-H6-TEV-GRAB by four consecutive PCR reactions to introduce a C-terminal Strep-tag.
pIE-TGFβ2-H6	pro-$TGFβ_2$	Human pro-$TGFβ_2$ c-DNA pIEx vector	CAATAGGCCTTGTCTACCTGCA GCACACTC	CAATACCGGTGCTGCATTTGCAA GACTTTAC	L^{21}-S^{414}+**TG**+$\mathbf{H_{6x}}$	C-t H_{6x}	Pro-$TGFβ_2$ in S2 cells. The gene was inserted by directional cloning (between *Stu*I and *Age*I) into the pIEx vector in frame with AKH signal peptide sequence
pIE-mTGFβ2-H6	$TGFβ_2$	Human pro-$TGFβ_2$ c-DNA pIEx vector	CAATAGGCCTCAGCTTTGGAT GCGGCCTATTG	CAATACCGGTGCTGCATTTGCAA GACTTTAC	A^{303}-S^{414}+**TG**+$\mathbf{H_{6x}}$	C-t H_{6x}	As above for mature $TGFβ_2$.
pS6-TGFβ2-H8	pro-$TGFβ_2$	pIE-TGFβ2-H6 pCMV-Sport 6 vector	TCACCACCACCATCATCTCAGC CTGTCTACCTGCAGCA; GGTTCCACTGGTGACCACCAC CATCACCACCACCATC; GGGTACTGCTGCTCTGGGTTC CAGGTTCCACTGGTGAC;	CAATCTCGAGCTAGCTGCATTTG CAAGACTTTAC	$\mathbf{H_{8x}}$+**LS**+L^{21}-S^{414}	N-t H_{8x}	Pro-$TGFβ_2$ in Expi293F cells. The coding sequence of human pro-TGFβ2 without its signal peptide, spanning the LAP (L21-R302) and the GF (A303-S414), was inserted in five consecutive PCR steps into the pCMV-Sport6 vector in frame with the Kozak sequence, the signal peptide from the mouse Ig κ-chain and the 8x-His-tag to the protein of interest. The final PCR product was digested with

			GACAGACACACTCCTGCTATG GGTACTGCTGCTC; CAATCCCGGGGCCACCATGGA GACAGACACACTCC				SmaI and XhoI restriction enzymes and ligated into pre-digested pCMV-Sport6 (Del Amo-Maestro et al., 2019).
pS6-TGFβ2-STREP	pro-$TGFβ_2$	pS6-TGFβ2-H8	GGTGGAGGTTCTGGAGGTGG AAGTGGAGGTAGCGCATGGA GCCATCCACAATTCGAAAAGC TCAGCCTGTCTACCTGC	CTTTTCGAATTGTGGATGGCTCC AGTCACCAGTGGAACCTGGAAC CCAGAGCAG	**WSHPQFEKGGG SGGGSGGSAWS HPQFEKLS**+L^{21}-S^{414}	N-t Strep	Pro-$TGFβ_2$ in Expi293F cells. The parental plasmid was modified by opposite primers to replace the N-terminal histidine-tag with a Strep-tag (Del Amo-Maestro et al., 2019).
pCri8a-H6-TEV-CR345	CR345	Human LRP1 c-DNA pCri8a vector	CAATCCATGGGACAGTGCCAG CCAGGCG	CAATCTCGAGTTAATAGGCACAC GAAGC	**GAMG**+Q^{853}-Y^{973}	N-t H_{6x}+TEV	CR345 of human LRP1 in *Escherichia coli* inserted into the pCri8a vector (Goulas et al., 2014a) by directional cloning between the *Nco*I and *Xho*I restriction sites.
pIE-CR34-H6	CR34	Human LRP1 c-DNA pIEx vector	CAATAGGCCTCACAGTGCCAG CCAGGCGAGTTTG	CAATACCGGTGGCTGAACAAGT GGCATTG	Q^{853}-A^{933}+**TG+H_{6x}**	C-t H_{6x}	CR34 of human LRP1 in S2 cells. The gene was inserted by directional cloning (between *Stu*I and *Age*I) into the pIEx vector in frame with the AKH signal peptide sequence. Partial digestion with Age I (there are 2 restriction sites for AgeI into the sequence).
pIE-CR345-H6	CR345	Human LRP1 c-DNA pIEx vector	CAATAGGCCTCACAGTGCCAG CCAGGCGAGTTTG	CAATACCGGTATAGGCACACGA AGCAGAC	Q^{853}-Y^{973}+**TG+H_{6x}**	C-t H_{6x}	CR345 of human LRP1 in S2 cells. The same as pIE-CR34-H6.
pS6-CR345-H6	CR345	Human LRP1 c-DNA pCMV-Sport 6 vector	TGGGTTCCAGGTTCCACTGGT GACTCACAGTGCCAGCCAGGC GAGTTT	CGCCTAATGGTGATGGTGATGG TGATAGGCACACGAAGCAGACT CATC	Q^{853}-Y^{973}+**H_{6x}**	C-t H_{6x}	CR345 of human LRP1 in Expi293F cells. The gene was inserted by restriction-free cloning into the pCMV-Sport 6 vector in frame with the Ig κ leader sequence.
pS6-CR3456-H6	CR3456	Human LRP1 c-DNA pCMV-Sport 6 vector	TGGGTTCCAGGTTCCACTGGT GACTCACAGTGCCAGCCAGGC GAGTTT	CGCCTAATGGTGATGGTGATGG TGGTGGCTGCAGCCGGCTTCGTC ACT	Q^{853}-H^{1013}+**H_{6x}**	C-t H_{6x}	CR3456 of human LRP1 in Expi293F cells. The gene was inserted by restriction-free cloning into the pCMV-Sport 6 vector in frame with the Ig κ leader sequence.

* Restriction-site sequences and overhangs for restriction-free cloning are underlined.

** Peptide sequence of the expressed protein after fusion-tag removal. Amino acids derived from the construct are in bold. See also **Fig**. 69

*** Fused tags at the carboxy-terminus (C-t).

3.4.2.4.1 GRAB fusion proteins

Full-length GRAB was expressed without the cell-wall anchoring region with two different affinity tags (6x-His and 2x-Strep)(Fig. 69A2) in a bacterial system yielding ~4 mg of pure protein per liter of expression medium after direct and reverse affinity chromatographies (elutions from first Ni-NTA were incubated O/N with TEV protease to be cleaved and then, the cleaved and free GRAB was recovered in the reverse Ni-NTA from the FT), IEC and SEC steps (Fig. 71). The protein migrated as a ~55 kDa species in SEC and as a ~33 kDa species in SDS-PAGE, but the values determined by SEC-MALLS (15.5 kDa; Table 23) were closer to the theoretical mass (15.8 kDa). We attributed this abnormal migration, which was described previously (Rasmusen et al., 1999), to the highly unstructured character of the protein.

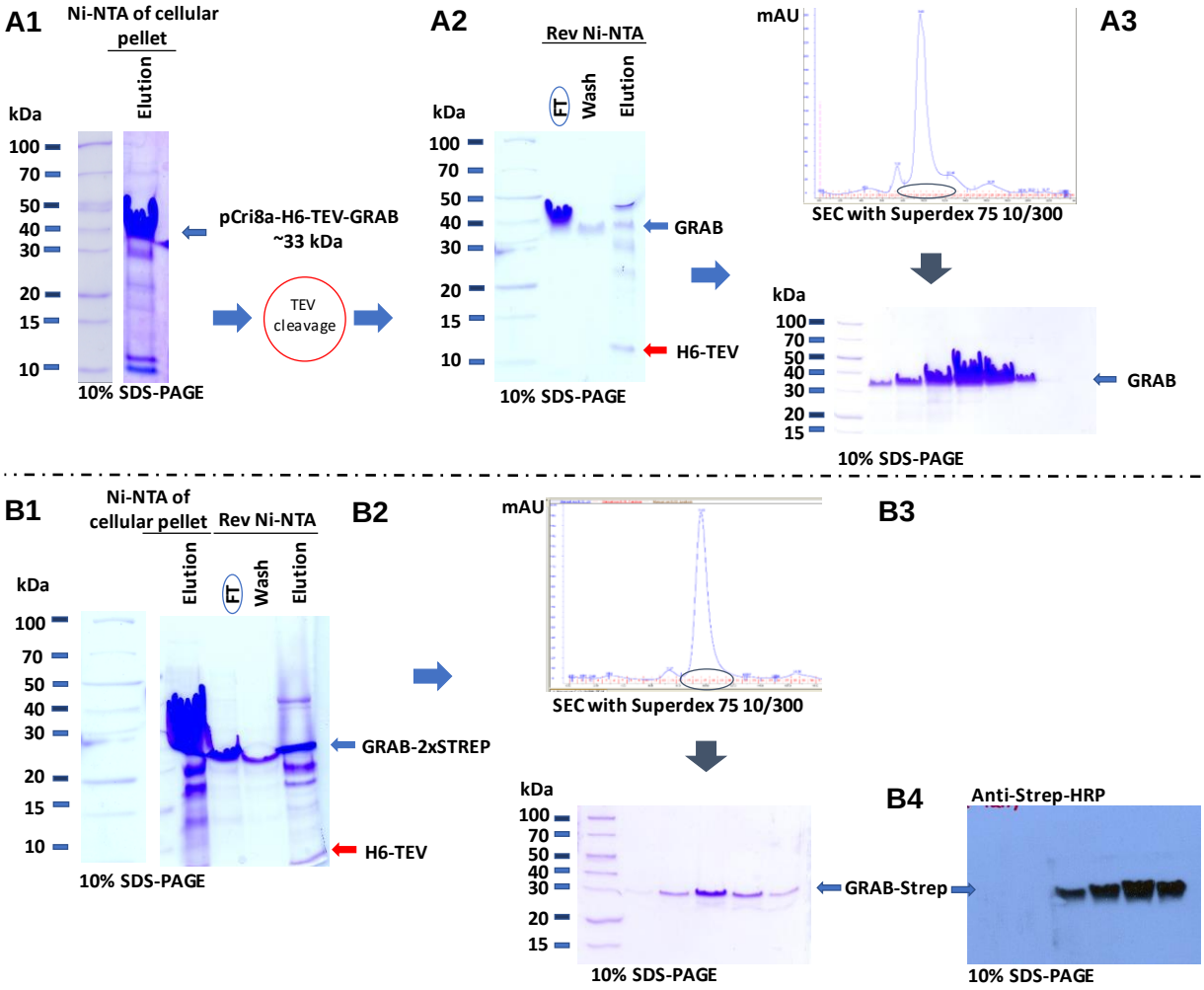


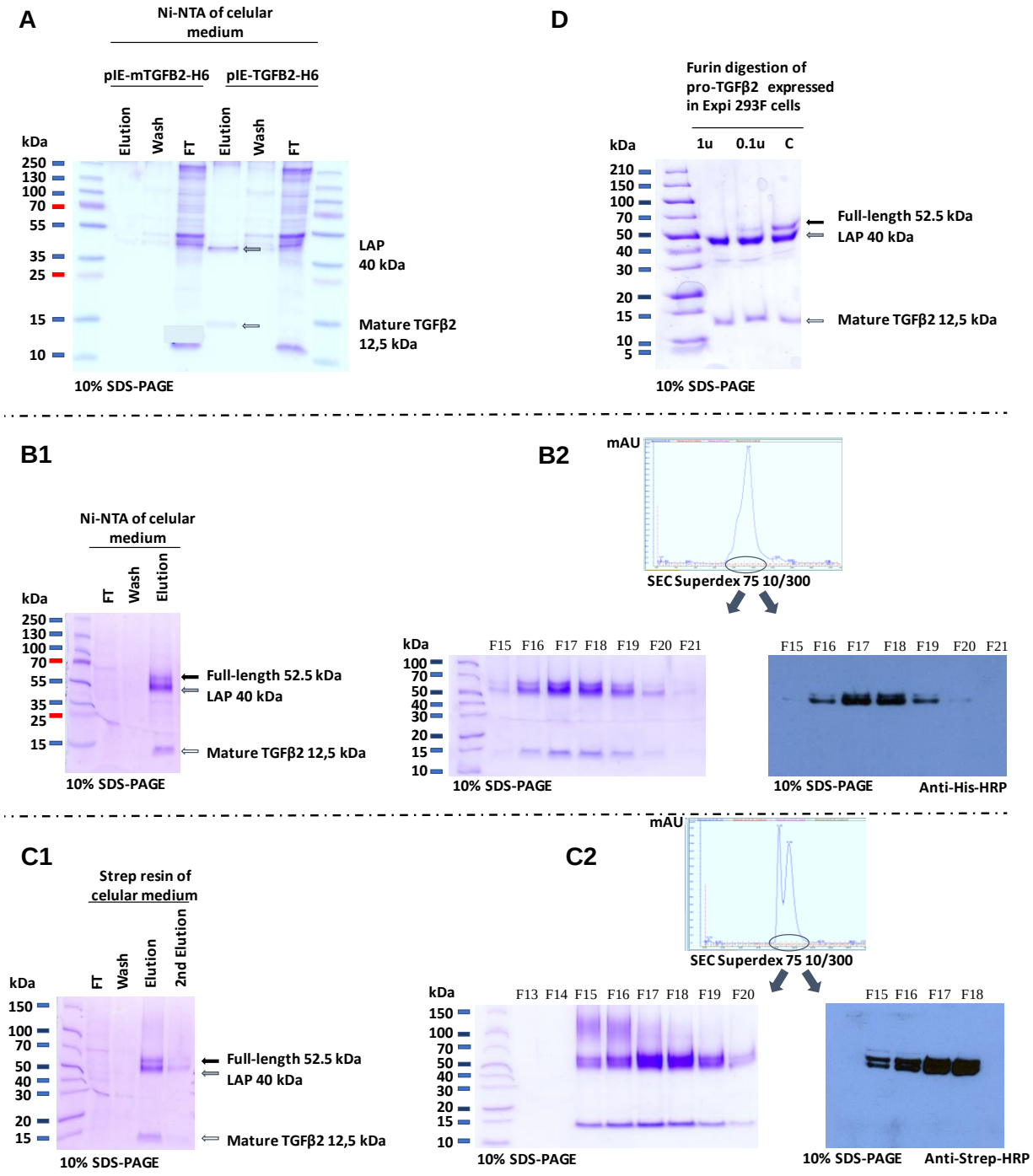
Figure 71. Overexpression of GRAB in BL21 (DE3) cells with 0.4 mM IPTG, O/N and at 20°C and purification. (A) Purification of GRAB **(A1)** direct and **(A2)** reverse affinity chromatographies (Ni-NTA) and **(A3)** SEC. **(B)** The same as (A), for GRAB-STREP. **(B4)** After the TEV cleavage, GRAB still have 2x-Strep tag which is used to be detected by Western blot against the Strep-tag after the SEC step. All proteins were analysed and identified by 10% SDS-PAGE with Coomassie staining. Arrows and circles indicate purified GRAB proteins of selected samples, respectively.

3.4.2.4.2 Pro-TGFβ2 fusion proteins

The insect and mammalian systems were also assayed for expression of pro-TGFβ2 (Fig. 69B2). Protein purifications were performed as for the α2M fusion proteins. In *Drosophila* insect cells, the mature TGFβ2 (pIE-mTGFB2-H6, from A^{303} to S^{414}) was no expressed, whereas the full-length construct (pIE-TGFB2-H6, from L^{21} to S^{414}) achieved expression but without noticeable yields (Fig. 72A). Therefore, full-length pro-TGFβ2 encompassing LAP and mature TGFβ2 was expressed and purified in Expi293F cells (Fig. 70 and 72B, C), with a final yield of ~2.7 mg and ~2.3 mg of pS6-TGFB2-H8 and pS6-TGFB2-STREP, respectively, per liter of mammalian cell culture. All the proteins were identified as TGFβ2 by peptide mass fingerprinting.

The protein migrated as a dimer of ~105 kDa in SEC (value determined by SEC-MALLS; Table 23), which indicates that the characteristic disulphide bonds were formed between the LAP and the mature TGFβ2 moieties. Our efforts to separate them by SEC in the presence of high salt contents (1 M sodium chloride), chaotropic agents (1 M/4 M urea), detergents (0.05% SDS), reducing agents (TCEP), and low-pH buffers (glycine pH 3.0) failed. We could not isolate them either by IEC with 20 mM sodium acetate pH 4.0 as buffer and a sodium chloride gradient. The purified protein was only partially cleaved before residue A^{303} by host peptidases. Subsequent treatment with the physiological activating endopeptidase furin produced a homogenously sample consisting of LAP associated with the mature cytokine (Fig. 72D). Under physiological conditions, TGFβ2 maturation is a complex process that involves a cascade of events with the participation of several proteins that interact with the initial complex of pro-TGFβ2 and the latent TGF-β binding protein (LTBP). LTBP participates as a localizer of pro-TGFβ2 to the extracellular matrix, whereas LAP senses the changes and releases mature TGFβ2 (Annes et al., 2003). Previous studies with a Chinese hamster ovary cell expression system benefited from the sensitivity of the LAP domain towards denaturing conditions at very low pH to separate it from mature TGFβ2 (Annes et al., 2003; Zou and Sun, 2006). In our case, this was unsuccessful, probably due to different post-translation modifications introduced by Expi293F cells in the highly glycosylated LAP (Croset et al., 2012).

Figure 72. Overexpression and purification of pro-TGFβ2 fusion-proteins in S2 cells and human Expi293F™ cells and purification. (A) Protein purification by affinity chromatography (Ni-NTA) of constructs expressed in insect cells. **(B)** Purification of pro-TGFB2-by **(B1)** Ni-NTA and **(B2)** SEC. Protein analysed and identified by 10% SDS-PAGE with Coomassie staining and Western blot against the His-tag. **(C)** Purification of pS6-TGFB2-STREP by **(C1)** Strep-tag resin and **(C2)** SEC**.** Protein analyzed and identified by 10% SDS-PAGE with Coomassie staining and Western-blot against the Strep-tag. **(D)** Pro-TGFβ2 digested by furin. Arrows indicate the different fragments of pro-TGFβ2 (full-length, LAP and mature protein).



3.4.2.4.3 LRP1 fusion proteins

The CR 3-4, CR 3-4-5 and CR 3-4-5-6 of LRP1 cluster 2 were expressed in bacteria, insect and mammalian expression systems (Fig. 73). In every system, CR constructs were expressed with two bands (Fig. 70 and 71), because they were partially unfolded (Dolmer and Gettins, 2006). Each CR

domain contains three disulphide bridges that needed to be folded to be functional, thus we performed a multi-step refolding protocol to ensure correct folding (Fig. 73B).

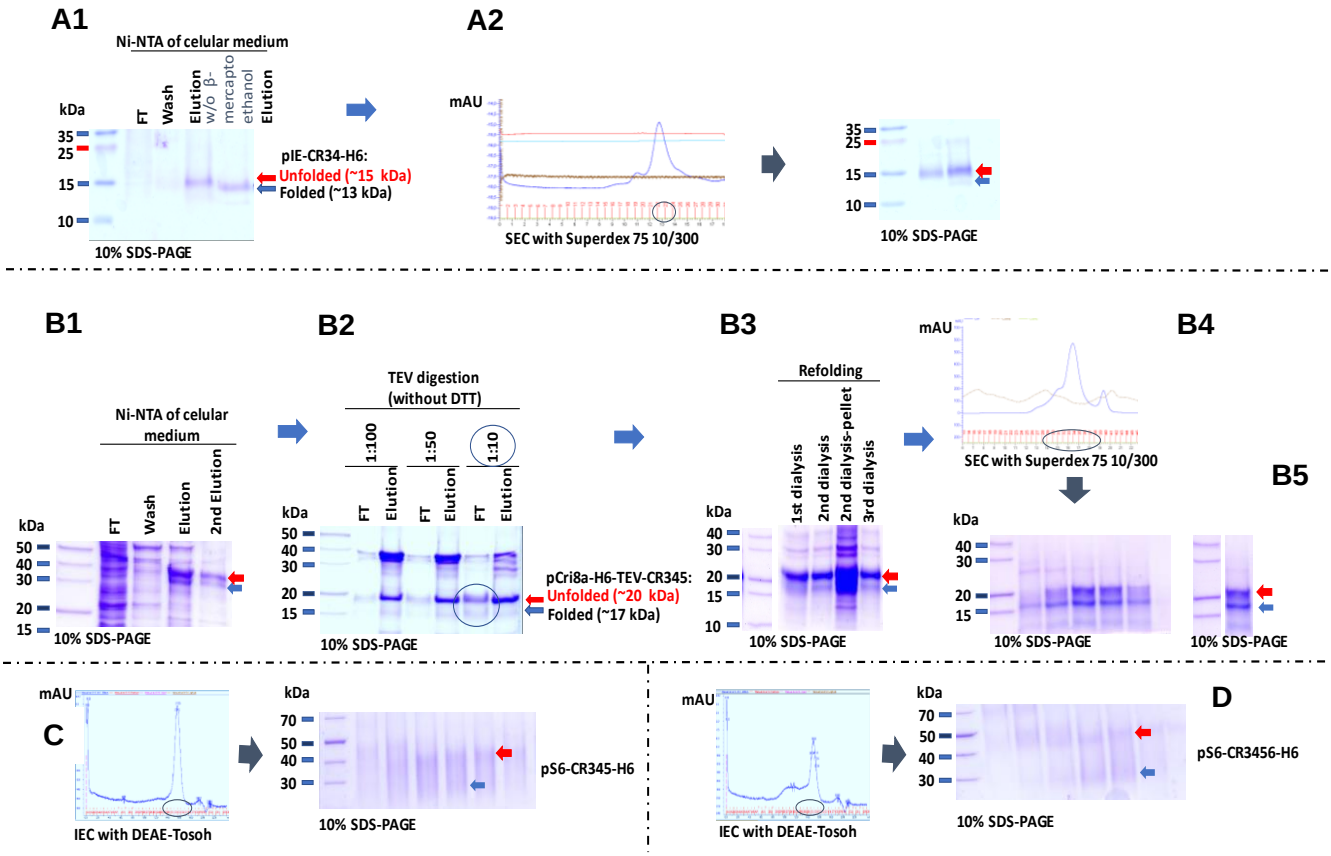


Figure 73. Overexpression of complement repeats (CR) fusion-proteins in bacteria, S2 and human Expi293F™ cells and protein purification. (A) Purification CR34 expressed in insect cells by **(A1)** affinity chromatography (Ni-NTA) and **(A2)** SEC. **(B)** Purification of pCri8a-H6-TEV-CR345-by **(B1)** direct and **(B2)** reverse Ni-NTA after TEV cleavage**, (B3)** multistep refolding**, (B4)** SEC and **(B5)** final refolded protein. **(C)** Purification of pS6-CR345-H6 by Ni-NTA and IEC. **(D)** The same as (C) for pS6-CR3456-H6. Red arrows indicate the unfolded proteins and blue arrows the folded ones.

3.4.2.5 Purification of wild-type hα2M from plasma blood

Wild-type hα2M protein was isolated from blood plasma (mixture of donors from Barcelona Blood Bank) and purified as described previously (Goulas et al., 2014b; Imber and Pizzo et al., 1981; Sottrup-Jensen et al., 1980) (Fig. 74A, B, C). Because of the heterogeneous source, every batch of protein behaved a bit different, with different ratios of contaminants (Andersen et al., 1994; Goulas et al., 2014b) and intermediate induced states. Once wild-type hα2M protein was purified, a change of its conformational state to the induced one, was done by treating part of it with methylamine (MA). Samples were then analyzed by native-PAGE to verify the conformational state change of hα2M from nascent to induced (Barret et al., 1979; Leuven et al., 1981). The native hα2M migrated slower than MA-induced hα2M in native-PAGE (Fig. 74D). We also found differences in SEC elution volumes and

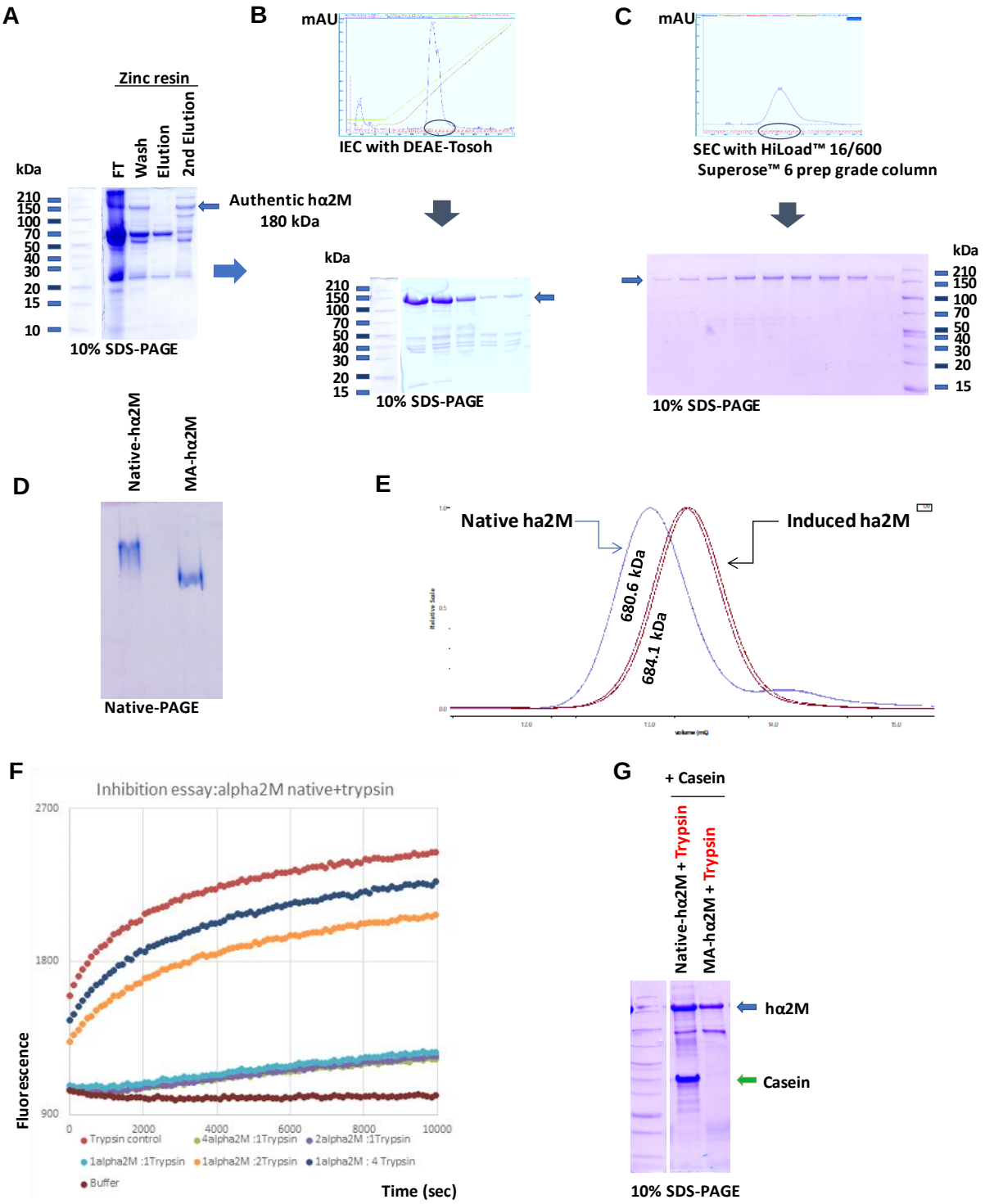


Figure 78. Purification and validation of wild-type hα2M from blood plasma. First, hα2M was isolated by precipitation with PEG 4000 and dissolved in Na_2HPO_4 and then, **(A)** purified by affinity chromatography with Zinc resin**.** Next, best fractions were pooled, then dialyzed against buffer A of IEC, and finally purified by **(B)** IEC and **(C)** SEC**. (D)** Part of the protein was treated with methylamine (MA) and analyzed by native-PAGE. **(E)** SEC-MALLS of hα2M species, with native (left and blue) and MA-induced (right and red) wild-type hα2M showing the measured molecular mass distribution. **(F)** and **(G)** proteolytic inhibition assays. **(F)** Inhibition assay performed in a fluorimeter with BODIPY FL-casein as fluorogenic substrate, and trypsin as peptidase against different ratios of tetrameric native hα2M. **(G)** The same as (F) but in 10% SDS-PAGE and with no-fluorogenic casein.

molecular mass (kDa) depending on the conformational state (~680 kDa and ~684 kDa for native and MA-induced hα2Ms, respectively) (Fig. 74E) (values were determined by SEC-MALLS, Table 23). We attributed the difference in migration to the more compact conformation of the MA-induced hα2M (Sottrup-Jensen, 1989), and the increase of molecular mass to the MA. To validate the activity of the native hα2M, we performed two types of proteolytic inhibition assays. The first inhibition assay against protein substrates was performed in a microplate fluorimeter (Infinite M200, TECAN) with BODIPY FL-casein as fluorogenic substrate and trypsin as peptidase against different ratios of tetrameric hα2M. One tetramer of native wild-type hα2M was able to inhibit one molecule of trypsin (Fig. 74F). The second inhibition assay was also with casein as substrate and trypsin as peptidase and allowed us to see the protector effect of native wild-type hα2M against casein degradation by trypsin in 10% SDS-PAGE (Fig. 74G).

3.4.2.6 Analysis of the protein interaction between hα2M and GRAB

Interaction of streptococci with hα2M has been reported to be highly specific (Godehardt et al., 2004; Rasmussen et al., 1999). Group A, G and C streptococci all bind the native form, whereas only the latter interact with the induced form. This result was attributed to the types of surface proteins, which are specific for each strain. GRAB is found in group A streptococci, and we studied its interaction with native wild-type hα2M by size exclusion chromatography, pull-down assays, surface plasmon resonance (SPR), SEC-MALLS and fluorogenic Sulfo-NHS-AMCA. The complex was first detected by size exclusion chromatography, where it was eluting together after a 15 min incubation at RT in the same elution volume (Fig. 75A1). It was also detected by pull-down with Strep-Tactin® resin under the same incubation conditions. GRAB-Strep was first bound to the resin and then, native wild-type hα2M (with no-tags) was passed through. After being extensively washed, native hα2M was still bound to GRAB (Fig. 75B1). To confirm this complex, GRAB was also immobilized as a ligand through a Strep-tag on a chip with covalently bound streptavidin. In a multicycle experiment, saturation of the ligand was reached with the two highest analyte concentrations, which gave on- and off-rate kinetic constants and results from affinity analysis (Fig. 76, A). From the sensorgrams during the sequential injections of different analyte concentrations, we observed fast association and slow dissociation of hα2M from GRAB, which indicated stable complex formation. Therefore, the ligand was removed in a regeneration step to make sure that all bound hα2M was eliminated between injections with different analyte concentrations. The group of curves in Fig. 76, B*i*, were fitted to a 1:1 Langmuir interaction model. These calculations revealed a $\chi 2$ value < 10% of Rmax, which is indicative of a good fit (Sinha-Datta et al., 2015). Consistently with the sensorgrams, the association rate constant (ka) and the dissociation

rate constant (kd) were 1.32 × 10^5 $M^{-1}s^{-1}$ and 1.90 × 10^{-3} s^{-1}, respectively, with an estimated dissociation halftime ($t_{1/2}$ = ln2/k_{off}) of 365 sec. The equilibrium dissociation constants (KD) from the kinetic and affinity analysis were 1.43 × 10^{-8} M and 3.45 × 10^{-8} M (Tables 23 and 24), respectively, which indicates high affinity and stable complex formation. The complex was also detected by SDS-PAGE and native-PAGE employing fluorophore-labelled GRAB (Fig. 77B). Finally, SEC-MALLS analysis (Fig.77A and Table 23) showed a molecular mass difference of 27.3 kDa over free hα2M, which corresponds to 1.7 molecules of GRAB per α2M tetramer . Hence, we assume that two molecules of GRAB bind one hα2M tetramer.

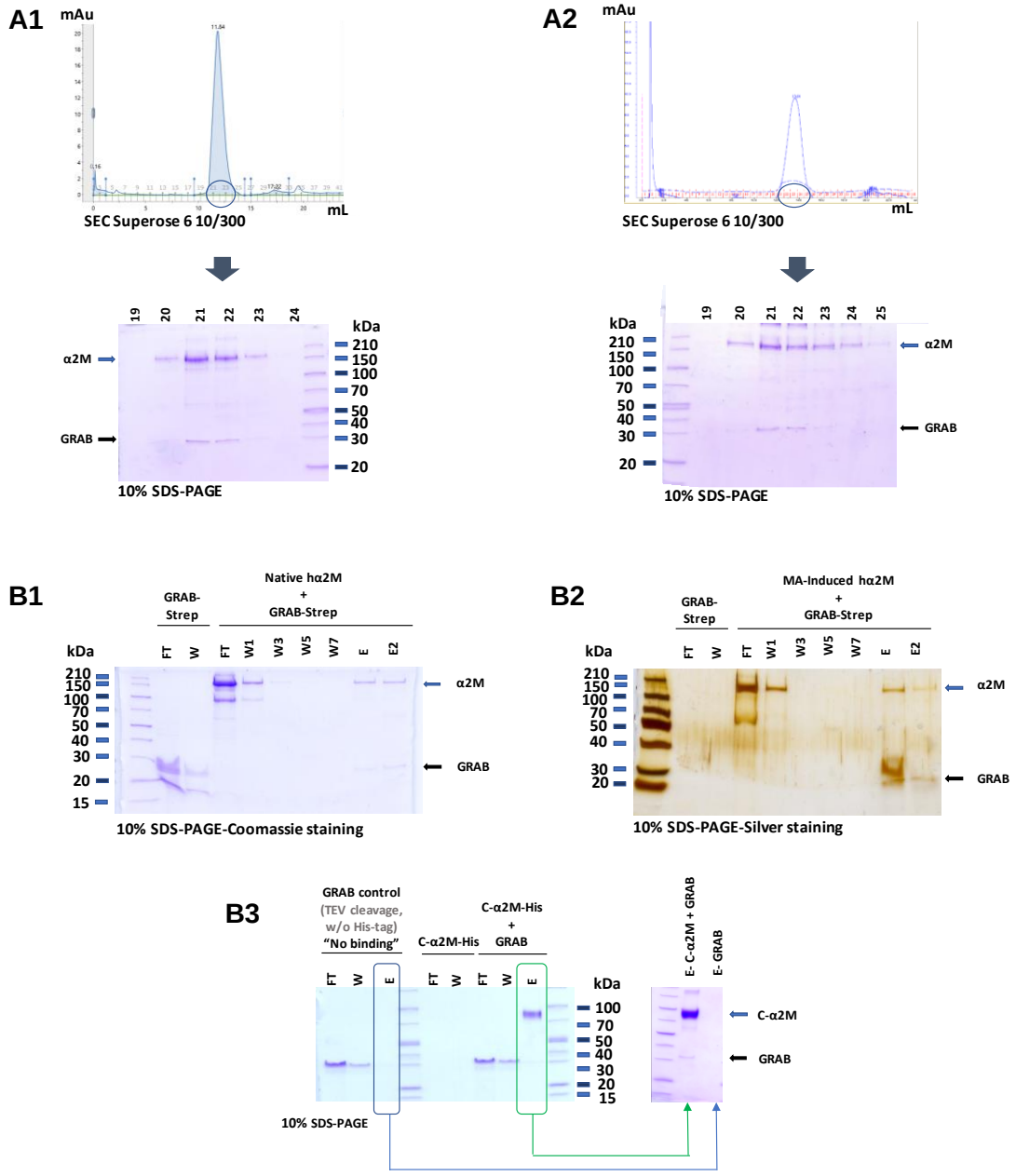


Figure 75. Preliminary studies of the interaction of GRAB with hα2M variants. (A) SEC analysis of complex formation between GRAB and native **(A1)** and induced **(A2)** wild-type hα2M. **(B)** Pull-down assays with Step-Tactin resin of GRAB-Strep and native **(B1)** and induced **(B2)** wild-type hα2M and with nickel resin of cleaved GRAB (without His-tag) and recombinant C-hα2M-His **(B3)** (concentrated elutions in squares). Results were analyzed by SDS-PAGE with Coomassie and Silver staining.

The interaction of MA-induced wild-type hα2M with GRAB was first studied by size exclusion chromatography and pull-down with O/N incubations at 4°C or 1h at 37°C (Fig. 75A2, B2). Under a similar experimental setup, MA-induced wild-type hα2M was also injected over immobilized GRAB to reach equilibrium and saturation, which enabled analysis by affinity. The affinity data permitted calculation with confidence (χ^2 < 10% of R_{max}) of the equilibrium dissociation constant (9.46×10^{-8} M), which was three times higher than that of native hα2M (Fig. 76B*i* and B*ii*) and Table 25). This was consistent with published results, which indicated that GRAB shows preference for native over protease-induced hα2M (Rasmusen et al., 1999). The complex was likewise analyzed by SDS-PAGE and native-PAGE with fluorophore-labelled GRAB (Fig. 77B). The SEC-MALLS results showed an increase in the molecular mass of 26.5 kDa over noncomplexed induced hα2M, which was equivalent to the results for native hα2M (Fig.77A and Table 23).

Protein sample	Molecular mass (kDa)
Native hα2M	680.6 ± 1.8
Native hα2M + GRAB	707.8 ± 3.4
Native hα2M + pro-TGFβ2	690.3 ± 0.4
Induced hα2M	684.1 ± 2.7
Induced hα2M + GRAB	710.6 ± 1.5
Induced hα2M + pro-TGFβ2	694.5 ± 0.4
GRAB	15.5 ± 0.0
Pro-TGFβ2	105.4 ± 0.6

Table 23. Molecular masses determined by SEC-MALLS. Values are represented as means and standard deviations of three replicates.

Protein sample	k_a (M^{-1} s^{-1})	k_d (s^{-1})	K_D (M)	R_{max} (RU)	χ^2
Native $h\alpha_2M$ + GRAB	$1.32 \times 10^{+5}$	1.90×10^{-3}	1.43×10^{-8}	18.51	1.23

Table 24. Kinetic rates and equilibrium constants of the interaction between native wild-type hα2M and GRAB. Constants were calculated from the corresponding plot assuming a 1:1 interaction model (two GRAB molecules per hα2M dimer), see Fig. 72A; ka, association rate constant; kd, dissociation rate constant; KD, equilibrium dissociation constant.

Protein sample	K_D (M)	R_{max} (RU)	χ^2
Native $h\alpha_2M$ + GRAB	3.45×10^{-8}	18.94	1.17
Induced $h\alpha_2M$ + GRAB	9.46×10^{-8}	6.82	0.01

Table 25. Equilibrium constants of the interaction between native or induced wild-type hα2M and GRAB. Values were derived from the corresponding plot of steady state response against concentration assuming a 1:1 model (one GRAB molecule per hα2M dimer), see Fig. 76, B.

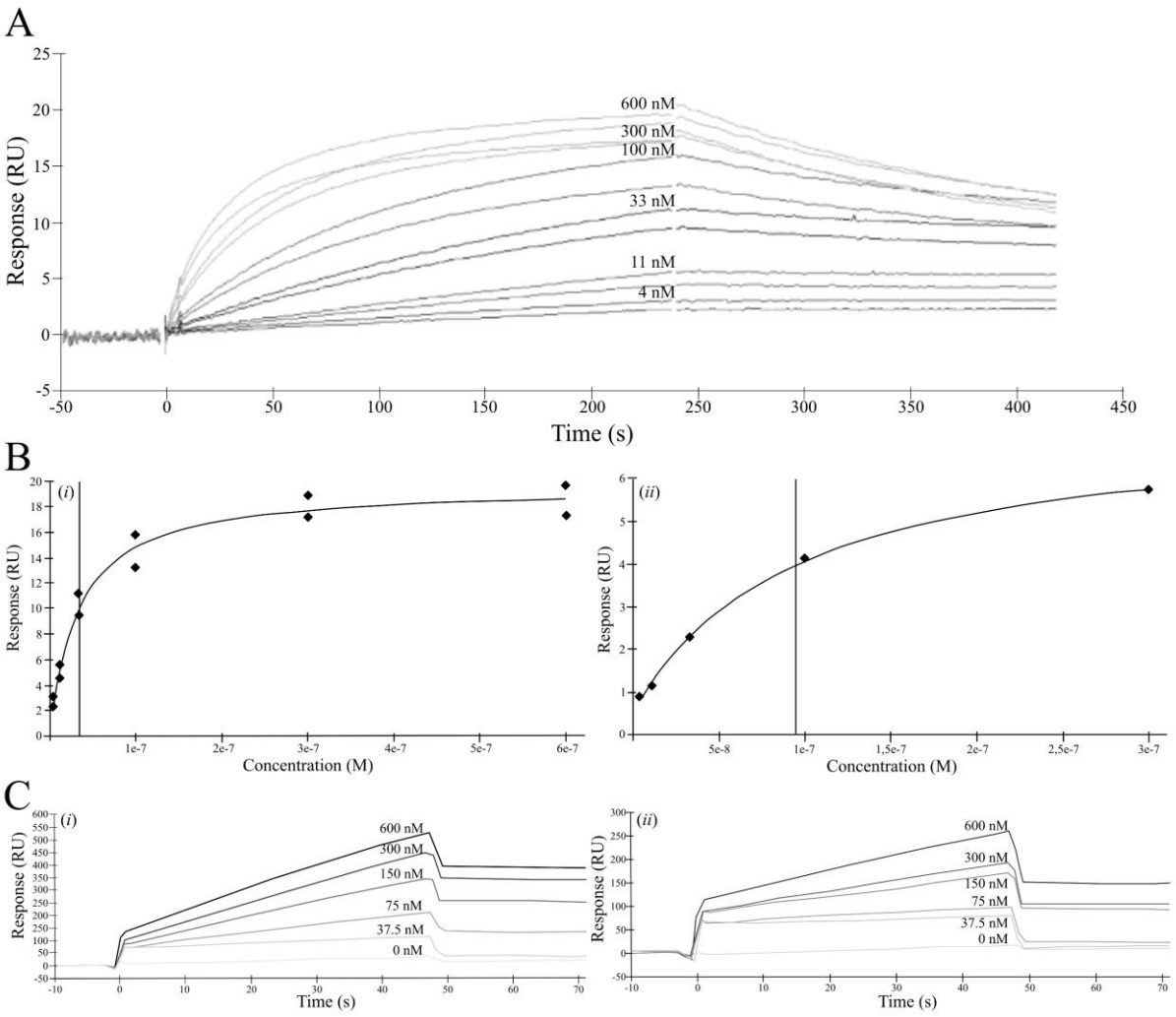


Figure 80. Surface-plasmon resonance studies of the interaction between GRAB with hα2M variants. (A, B) Surface-plasmon resonance sensorgrams of the interaction of native or induced wild-type hα2M with GRAB. Multi-cycle run for native hα2M with GRAB **(A)** and corresponding plot of the steady-state response **(B**, i and ii, for native and induced hα2M, respectively). Different hα2M concentrations were assayed to determine the rate constants that describe the kinetics and the equilibrium constants for complex strength (see also Tables 24 and 25). The vertical line in the plots of steady-state response indicates the value of the calculated equilibrium dissociation constant KD. **(C)** Sensorgrams of the interaction of N-hα2M (i) and C-hα2M (ii) with GRAB. Proteins were premixed, incubated at 37 °C for 1 h, injected over the chip, and the response was measured.

To map down the region of hα2M engaged in GRAB binding, we repeated the above experiments with N-hα2M and C-hα2M (Fig. 75B3). In a similar multicycle experimental setup, we could not detect any interaction. However, a previous incubation of the proteins at 37°C for one hour apparently enabled complex formation. Protein remained complexed over time after injection and washing of the chip (Fig. 76 C*i* and Cii), but in this case we could not determine the affinity constants due to the experimental setup. The complexes were subsequently evaluated in native-PAGE using fluorophore-

labelled GRAB (Fig. 77B). In this case, we detected the interaction of GRAB with C-hα2M and N-hα2M, but the signal was stronger for the C-hα2M.

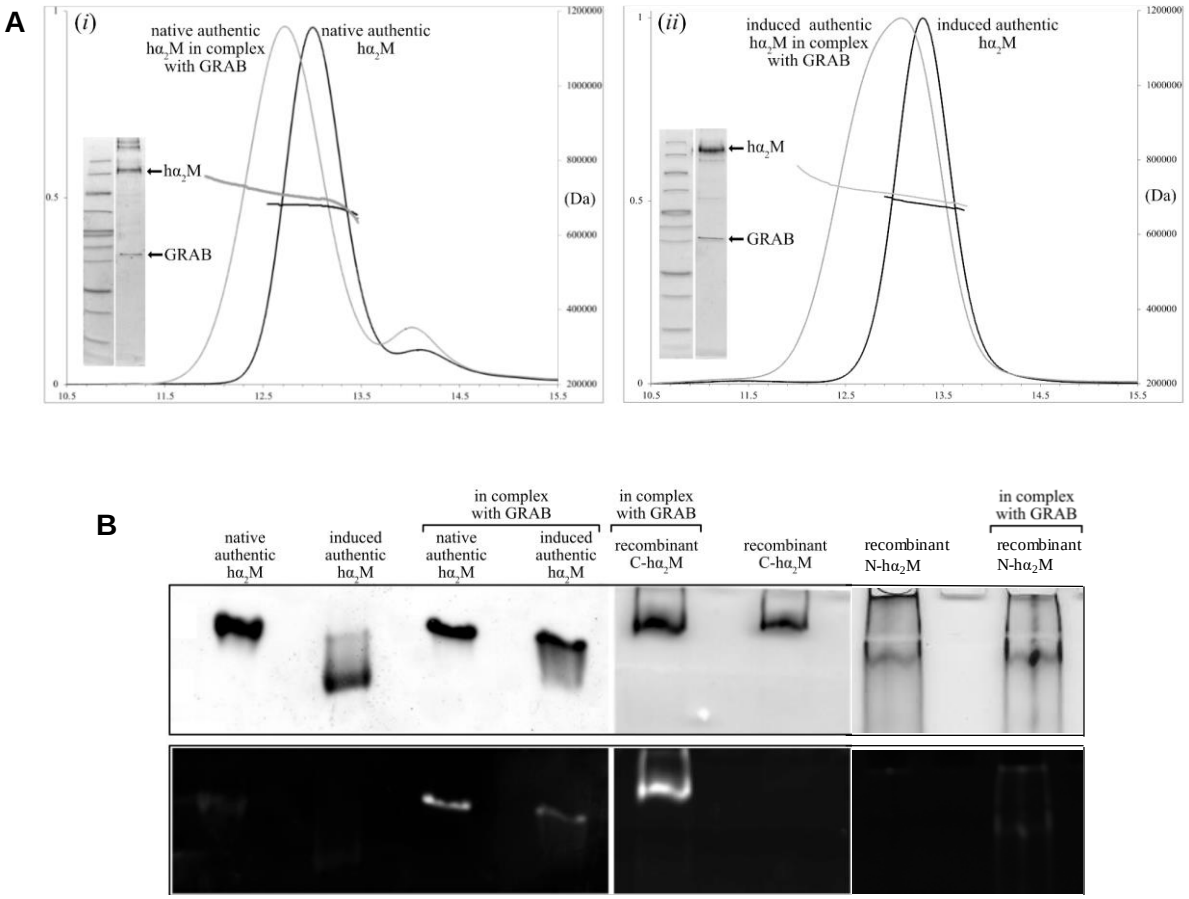


Figure 81. Last analysis of complex formation between hα2M variants and GRAB. (A) SEC-MALLS analysis of complex formation between GRAB and native (left) and induced (right) wild-type hα2M showing the measured molecular mass distribution. Inserted figures within graphs show the SDS-PAGE analysis of the respective purified complexes. **(B)** Complexes were separated by native-PAGE. GRAB was labelled with fluorogenic Sulfo-NHS-AMCA, visualized in a gel reader (lower panels) and then stained with Coomassie Brilliant Blue (upper panels).

3.4.2.7 Analysis of the protein interaction between human α2-macroglobulin and pro-TGFβ2

Previous biochemical data had revealed that hα2M binds pro-TGFβ2 mainly through a mature cytokine segment spanning residues A^{343}-Y^{367}, in which W^{354} plays a major role (Liu, Q. et al., 2001). No data have been reported on the role of the LAP domain. However, inspection of the crystal structure of homologous pro-TGFβ1 (see Protein Data Bank code 3RJR37) reveals that the interacting segment is partially shielded by LAP. Other studies employing a library of overlapping glutathione S-transferase fusion proteins ascribed the potential binding site for TGFβ1 to segment V^{723}-T^{761} of hα2M (Webb et al., 2000), which was subsequently narrowed down to E^{737}-V^{756} employing synthetic peptides

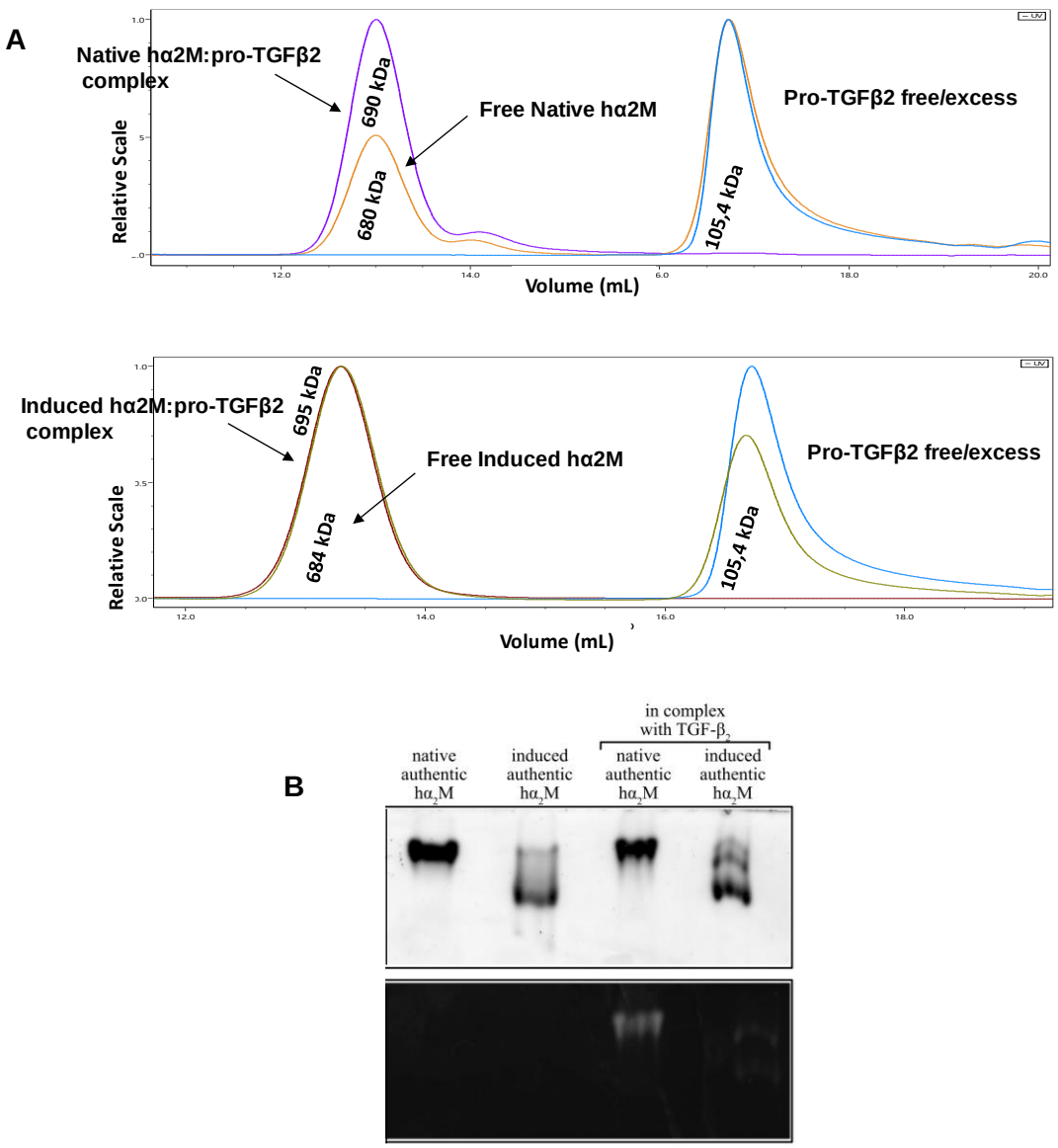


Figure 82. Analysis of complex formation between hα2M variants and pro-TGFβ2. (A) SEC-MALLS analysis of complex formation between pro-TGFβ2 and native (left) and induced (right) wild-type hα2M. **(B)** Complexes were separated by native-PAGE. TGFβ2 was labelled with fluorogenic Sulfo-NHS-AMCA and was visualized in a gel reader (lower panels) and then stained with Coomassie Brilliant Blue (upper panels).

(Arandjelovic et al., 2003). However, further details on the mechanism are unknown. To shed further light, we set out to characterize the binding of pro-TGFβ2 to hα2M. We studied the interaction by SPR, SEC-MALLS and Sulfo-NHS-AMCA. We only detected a difference of 10 KDa between the free native/induced hα2M and in complex with pro-TGFβ2 by SEC-MALLS (Fig. 76A and Table 23). These minimal differences could be because one molecule of mature TGFβ2 is interacting with both species of tetrameric hα2M. In SPR we performed multicycle experiments with immobilized pro-TGFβ2 as ligand but could not detect complex formation. Only after analysis by native-PAGE using fluorophore-

labelled pro-TGFβ2 we observed interaction with native wild-type hα2M and with the induced form (less intense) but not with the short variants (Fig. 78B). Given that the pro-TGFβ2 sample contained a mixture of cleaved and intact protein, we assayed the BS³ crosslinking technique with pro-TGFβ2 with Strep-tag linked to the N-terminal part of the LAP domain. After O/N incubation at 4°C of the complex with native wild-type hα2M and then, 30 min incubations at RT with BS³ crosslinker at different concentrations, a Western blot against the Strep-tag was performed in 10% SDS-PAGE. No interaction was detected in this case, because the Strep-tag was in the LAP domain (Fig. 79A), confirming that the interaction seen with the Sulfo-NHS-AMCA assay was only with the mature TGFβ2. The same experiment was performed with GRAB-Strep and native wild-type hα2M and the interaction was detected by WB, validating the experiment (Fig. 79B). Thus, we conclude that LAP prevents hα2M from binding mature TGFβ2 as suggested by structural studies on pro-TGFβ1.

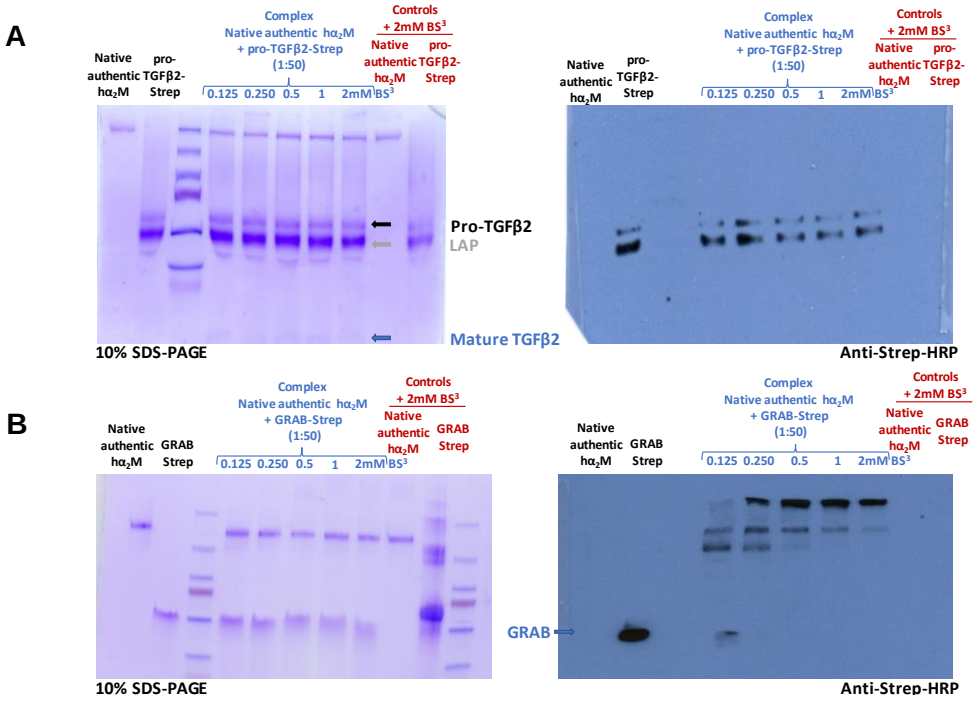


Figure 79. Validation of complex formation between wild-type hα2M and mature TGFβ2. BS³ crosslinking of complexes between native wild-type hα2M and **(A)** pro-TGFβ2-Strep and **(B)** GRAB-Strep. Results of the interactions were analyzed by Coomassie staining and Western blot against the Strep-tag.

3.4.2.8 Analysis of the protein interaction between human α2-macroglobulin and LRP1

The interaction of LRP1 with hα2M has already been reported and is highly specific for induced hα2M species (Dolmer and Gettins, 2006). For this binding, the exposure of the hidden RBD domain of hα2M is required. It is also known that induced hα2M can bind to cluster 2 and cluster 4 of LRP1, but with higher affinity for cluster 2 (Williams et al., 1994). Dolmer and Gettins reported in 2006 that the

binding between LRP1 and RBD was localized in the CR345 of cluster 2 in LRP1. Thus, there are available biochemical studies of RBD in complex with CR, but not structural models to confirm it. To shed further light, we set out to perform the binding of CRs to hα2M and later, crystallize it. We tried to perform the complex between C-hα2M or RBD expressed in S2 insect cells with several CR constructs (C34 and CR345 expressed in insect and bacteria cells, respectively) (Fig. 69C2). We preliminary studied the interaction by SEC, but we could not detect any interaction, even on presence of $CaCl_2$ (Fig.80a2M). Given that both proteins involved in this interaction contained disulphide bridges, and even more important for LRP1, which contains three disulphide bonds per CR domain (Daly et al., 1995) (Fig. 69C1), a proper refolding was seen to be crucial for the binding studies. RBD and CR construct were expressed into two different species, one well folded and the other unfolded and, at least the half of the proteins were unfolded (Fig. 66E and 73). A long refolding protocol was incorporated to fold CR345 expressed in bacteria, but we were still detecting the unfolded protein (Fig.73B). We proceeded with the binding studies, but we could not replicate the already published interaction (Dolmer and Gettins, 2006).

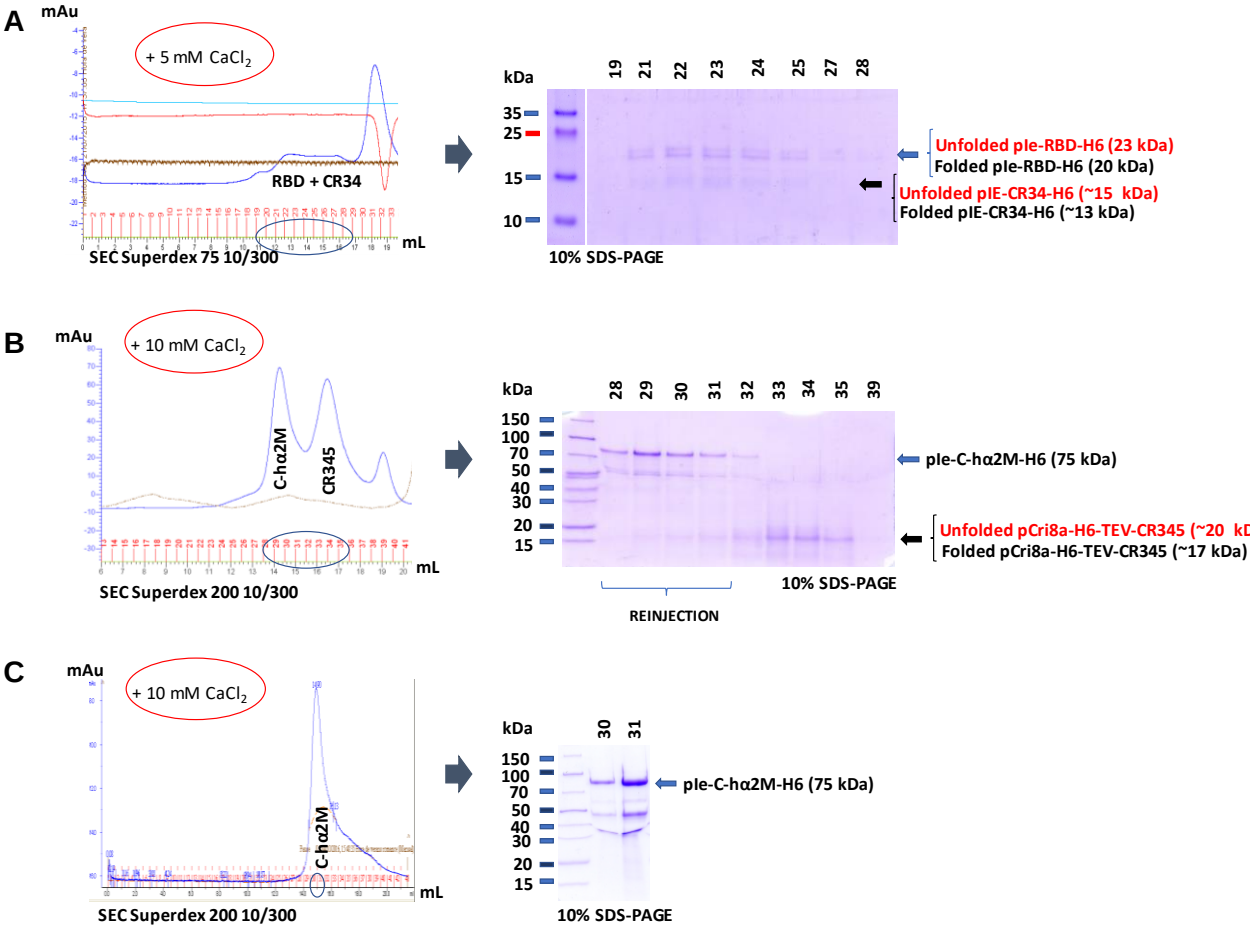


Figure 84. Analysis of complex formation between hα2M variants and CR constructs. (A) SEC analysis of complex formation between RBD and CR34. **(B)** SEC analysis of complex formation between C-hα2M and CR345. **(C)** Reinjection of first peak from (B) and we could not see complex with CR345. Results analyzed by 10% SDS-PAGE.

We conclude that CR samples were not correctly folded and extra steps of refolding were necessary, like coexpression and/or copurification with receptor-associated protein (RAP), which has seen to help in the refolding (Dolmer and Gettins, 2006).

3.4.2.9 Biochemical characterization of the recombinant hα2M proteins

In order to elucidate the proper folding of the thioester bond in the recombinant hα2M proteins (full-length and C-hα2M) and their inhibitory activity, purified proteins were addressed with several biochemical techniques. We performed native electrophoretic mobility studies, detection of the free sulfhydryl groups by a fluorometric thiol assay quantification (qualitative) (ab112158 assay; Abcam) and proteolytic inhibition assays with BODIPY FL-casein as fluorogenic substrate, and trypsin as peptidase. Native and MA-induced wild-type hα2M were also included to compare the effects shown and to validate the results.

Full-length hα2M expressed in S2 insect cells migrated in native-PAGE in a similar way to the MA-induced wild-type hα2M (Fig. 81A1), which migrates faster than the native protein. Chemical treatment with methylamine, which mimics the transition from native to induced hα2M by opening the reactive thioester bond to produce a free cysteine, did not have any effect on protein mobility of full-length hα2M expressed in S2 insect cells. Consistently, the protein could not inhibit trypsin activity against the fluorogenic protein substrate, even at 10-fold molar excess (Fig. 81C1). We concluded that recombinant hα2M produced in insect cells was in the induced form.

Whereas, full-length hα2M expressed in Expi293F cells showed electrophoretic mobility in native-PAGE between native and MA-induced wild-type hα2M (Fig. 81A2). Consistently, its capacity to inhibit trypsin was 35% of native wild-type hα2M (Fig.81C2 and C3). Together, these data indicate that the recombinant protein expressed in mammalian cells is partly native but mainly induced.

C-hα2M expressed in mammalian cells was also treated with methylamine and it did not affect the content of free cysteines or electrophoretic mobility in native-PAGE (Fig. 81A3). To follow this up, we qualitatively assayed the content of free sulfhydryl groups by performing a fluorometric thiol assay kit. This thiol sensor generates a strong fluorescent signal when free cysteines are reacting with it (λex = 490/λem = 520 nm). C-hα2M gave a stronger fluorescent signal for native than for the treated sample with MA (Fig. 81B)., meaning that the thioester bond of C-hα2M was already open, possibly owing to a nucleophilic component of the undisclosed cell-growth medium. This contrasted with native full-hα2M expressed in mammalian cells, which behaved like native wild-type hα2M, giving less signal than the MA-induced samples, which likewise gave a strong signal (Fig. 81B).

These assays indicated that the thioester bond was opened in C-hα2M and also behaved in native-PAGE as the MA-induced form; whereas, recombinant full-hα2M expressed in mammalian cells has some inhibitory activity. In order to confirm that the partial induction is a result of the nucleophilic components of the cell-growth medium, these recombinant proteins should first be expressed in a known recipe medium without nucleophilic elements and these assays should then be repeated with the new proteins.

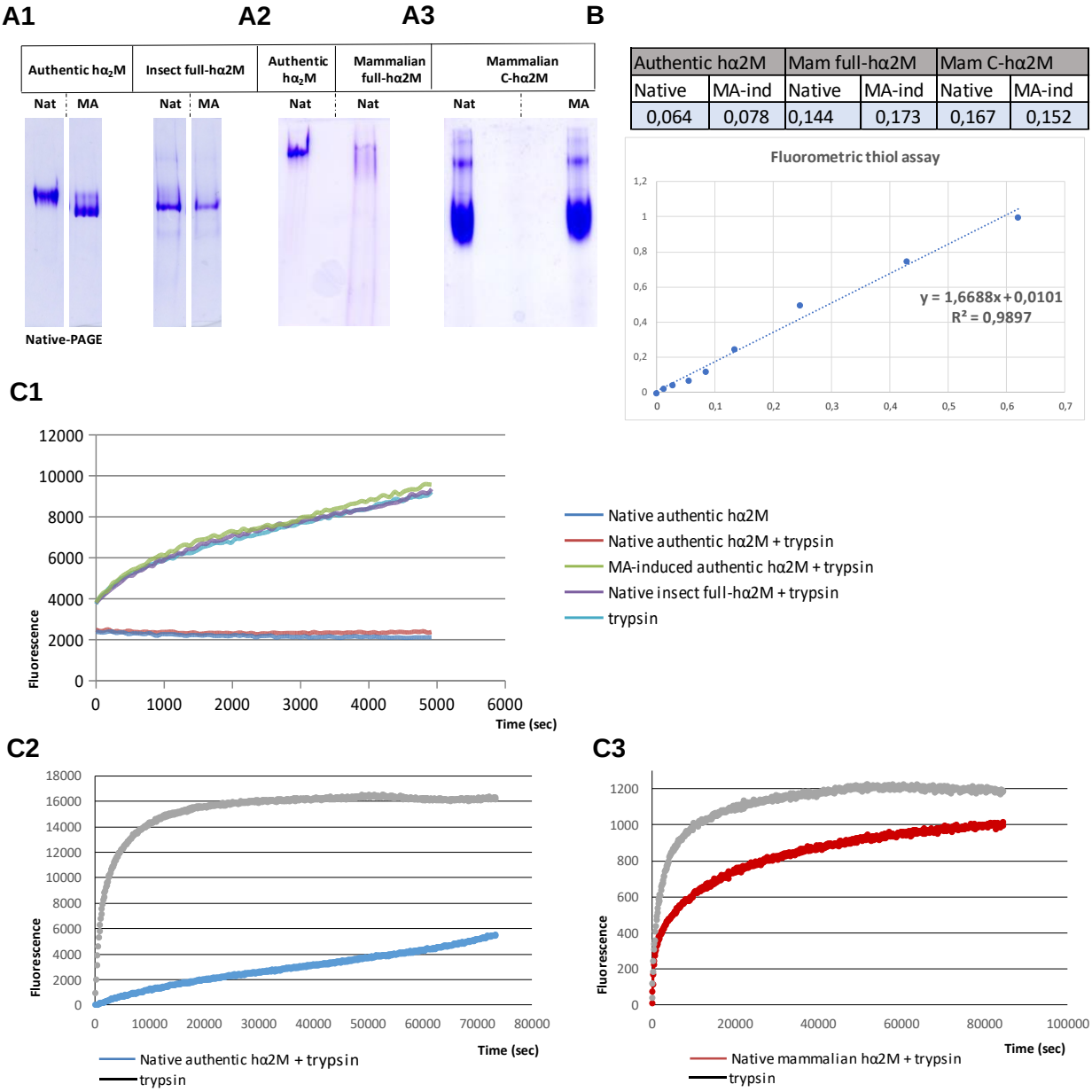


Figure 81. Biochemical characterization of the thioester bond in the recombinant hα2M proteins. (A) Native-PAGE analysis of wild-type and recombinant hα2M proteins, expressed in insect (S2) and mammalian (Expi293F) cells, untreated (native, nat) and treatred with methylamine (MA). **(B)** Detection of the free sulfhydryl groups by a fluorometric thiol assay in wild type and recombinant proteins (in native and induced states). **(C)** Fluorogenic proteolytic inhibition assays for wild-type and recombinant full-length hα2M proteins (expressed in insect and mammalian cells). **(C1)** Comparison of inhibitory protease activities of different types of hα2M (native and MA-induced wild-type and recombinant hα2M; expressed in S2 cells) against trypsin protease, whose substrate (BODIPY FL-casein) when it is cleaved by trysin releases fluorescent signal. There will

therefore not be fluorescence when hα2M is active and protecting the substrate to be degraded. **(C2)** The same as (C1) but only for native wild-type hα2M with trypsin & trypsin. **(C3)** The same as (C2) but for native mammalian hα2M.

3.4.2.10 Co-crystallization trials of the recombinant hα2M proteins with ligands

In order to improve the current atomic model of hα2M (crystal structure of MA-induced hα2M at 4.3 Å and published by Marrero and her colleagues in 2012 (PDB 4ACQ)), multiple efforts were made. Crystallization trials were first performed by screening of the conditions at different temperatures, protein concentrations, protein-precipitant ratios, buffers and additives at the IBMB/IRB Automated Crystallography Platform; and then, scaling up the best conditions to bigger volumes and condition range by performing the sitting-drop vapor-diffusion method.

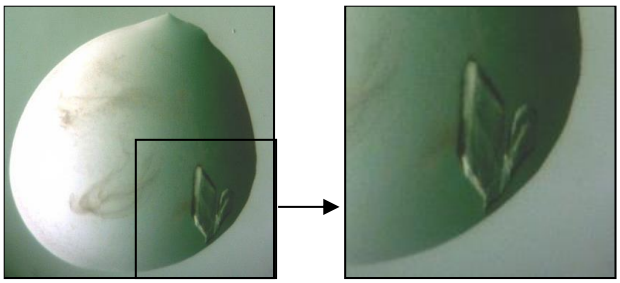
Pure protein of recombinant monomeric N-hα2M, treated with 10 mM TCEP and at pH 5.0 was promising scaled up by sitting drops at 4.7 mg/mL and 4°C in several plate conditions. They gave big crystals, but they turned out to be TCEP crystals (Fig. 82A). Afterwards, we also set up more drops for N-hα2M without TCEP in the buffer, which was substituted by DTT, but these crystals could not be reproduced.

Many efforts were made to achieve the crystallization of C-hα2M, but none resulted in crystals. This confirmed that this protein was still too flexible to be crystallized. Whereas, for C-α2M-no RBD, whose flexible RBD domain had been removed, one condition gave micro protein crystals at 20°C and 10 mg/mL (Fig. 82B). These crystals were transferred together into a mother liquor solution containing 15% (v/v) glycerol for cryo-protection and then, frozen in liquid nitrogen. They were diffracted at the synchrotron but unfortunately without diffraction because of the abundant ice surrounding the micro crystals.

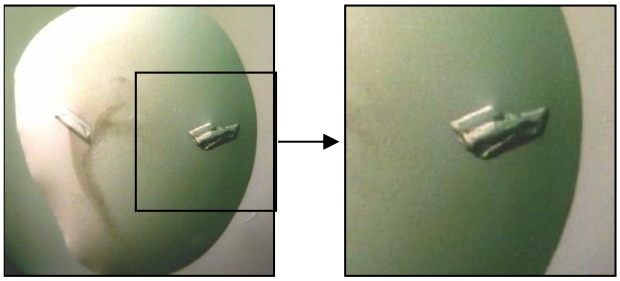
Wild-type hα2M (native and MA-induced) in complex with GRAB were also purified and crystal screens were set up at high concentrations and different temperatures. The complex 1 MA-hα2M molecule per 5 GRAB molecules crystallized into cubes and octahedral crystals after three months at 7.5 mg/mL and 20°C in several conditions. Best crystallization conditions were scaled up by sitting drop plates, flash-vitrified by liquid nitrogen in a cryo-protection drop containing 5% (v/v) glycerol and carried to the synchrotron (Fig. 83). Their diffractions were around 20 Å and they seemed to be only MA-induced hα2M.

During studies of the interactions between wild-type MA-induced hα2M and recombinant human pro-TGFβ2 a mixture of both proteins was set up for crystallization. Diffraction- grade protein crystals appeared after five days with 20% isopropanol, 0.2 M calcium

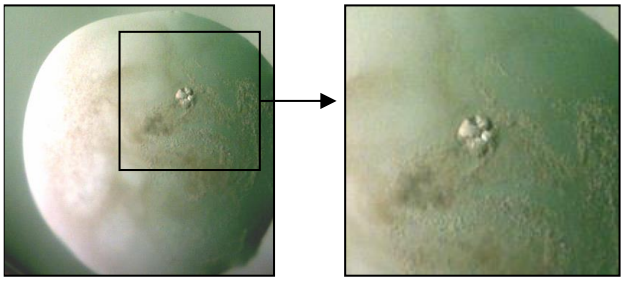
A



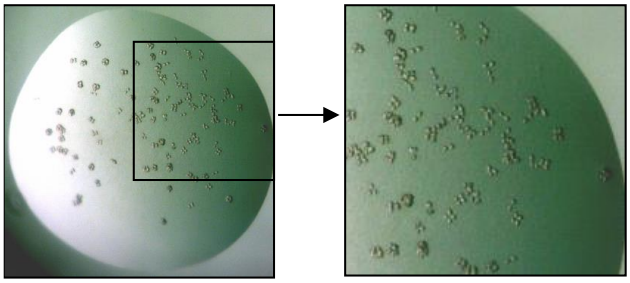
Crystallization conditions
20.0% (w/v) PEG 8000, 0.2M magnesium chloride, 0.1M hexahydrate tris pH 8.5 4°C 4.7 mg/mL of protein Protein with 10mM TCEP pH 5.0 1+1 Protein+precipitant ratio



Crystallization conditions
30.0% (w/v) PEG 4000, 0.2M magnesium chloride hexahydrate, 0.1M TRIS pH 8.5 4°C 4.7 mg/mL of protein Protein with 10mM TCEP pH 5.0 1+1 Protein+precipitant ratio

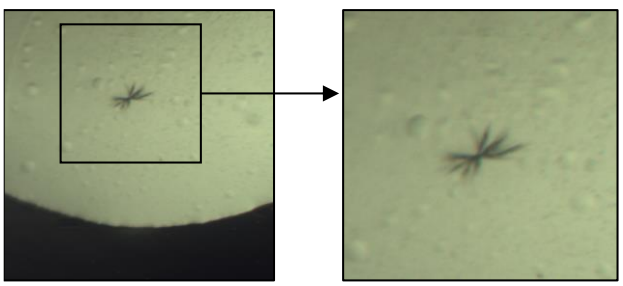


Crystallization conditions
15.0% (v/v) Ethanol, 0.2M zinc acetate dihydrate, 0.1M MES pH 6.0 4°C 4.7 mg/mL of protein Protein with 10mM TCEP pH 5.0 1+1 Protein+precipitant ratio



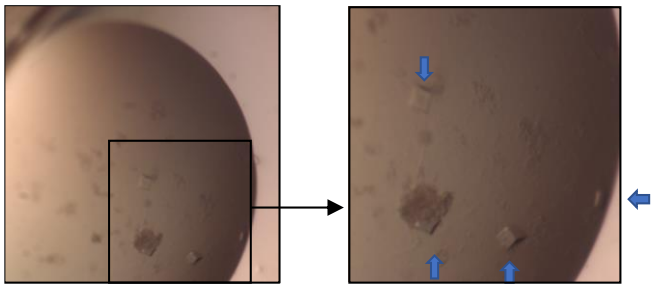
Crystallization conditions
1.0M Lithium sulfate monohydrate, 0.1M TRIS, 0.01M nickel chloride pH 8.5 4°C 4.7 mg/mL of protein Protein with 10mM TCEP pH 5.0 1+1 Protein+precipitant ratio

B

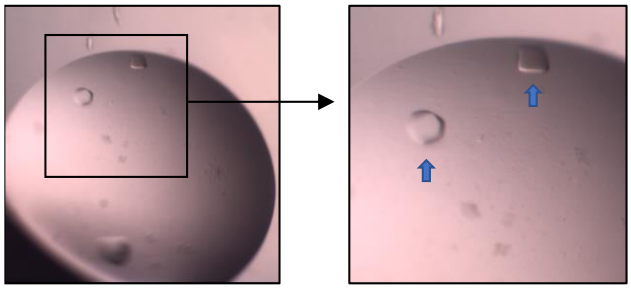


Crystallization conditions
1.6M Magnesium chloride heptahydrate, 0.1M MES pH 6.5 20°C 10 mg/mL of protein 2+1 Protein+precipitant

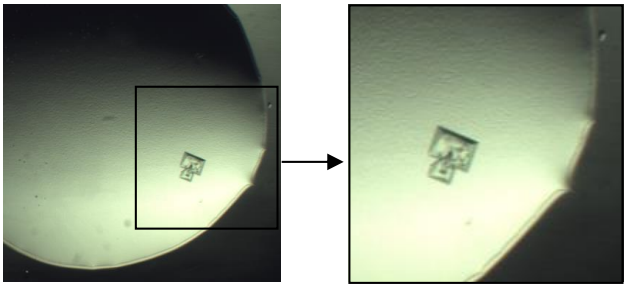
Figure 82. Crystallization trials of recombinant hα2M expressed in insect cells. (A) Crystallization conditions for N-hα2M and **(B)** for C-α2M-RBD.-



Crystallization conditions
0.1M Bicine, 2% dioxano, 10% (w/v) PEG 20000 20°C 7.5 mg/mL of protein Complex 1 MA-hα2M:5 GRAB 1+1 Protein+precipitant ratio



Crystallization conditions
0.2M Sodium chloride, 0.1M CHES pH 9.5, 10% (w/v) PEG 8000 20°C 7.5 mg/mL of protein Complex 1 MA-hα2M:5 GRAB 1+1 Protein+precipitant ratio



Crystallization conditions
5mM Magnesium sulfate hydrate, 0.05M Tris pH 8,5 35% 1,6-hexanediol 20°C 7.5 mg/mL of protein Complex 1 MA-hα2M:5 GRAB 1+1 Protein+precipitant ratio

Figure 83. Crystallization trials of wild-type MA-induced hα2M with GRAB (1:5 complex).

chloride, and 0.1 M sodium acetate pH 4.6 as reservoir solution (Fig. 84A). These crystals were transferred into a cryoprotecting drop containing the condition into the drop *plus* 15% (v/v) glycerol and then, they were flash-vitrified in liquid nitrogen. A complete dataset was collected at a beam line of the ALBA Synchrotron Radiation Facility (Barcelona, Spain). We grew more crystals under the same crystallization condition to perform a 10% SDS-PAGE (Fig. 84B) and a peptide mass fingerprinting of ~90 collected, carefully washed and dissolved crystals. This revealed that they contained mature TGFβ2 (GF). We also optimized the crystallization conditions with bigger crystals at 1.5 mg/mL and 20% ethanol in drops 2+1 and 3+1 (protein-precipitant). The crystals belonged to a hitherto undescribed, tightly packed tetragonal space group, diffracted to 2.0 Å (Fig. 85) resolution, and contained half a growth factor moiety (GF) disulphide-linked dimer per asymmetric unit. Diffraction data processing statistics are provided in Table 26. These crystals also appeared when pro-TGFβ2 alone was subjected to similar crystallization conditions. In this case, the crystals diffracted to lower resolution, i.e. hα2M apparently played a favorable role as an additive for crystallization. To pursue this further, we collected

the supernatant from crystallization drops that had given rise to crystals, purified it by size-exclusion chromatography, and then assayed a fraction that migrated according to the mass of hα2M by gelatin zymography. We detected gelatinolytic activity (Fig. 84C), which points to a contaminant present in the purified hα2M sample. Accordingly, we concluded that the low pH of the crystallization assay, together with the crystallization process, achieved the separation of the LAP and GF moieties that we could not obtain by chromatography. This process was probably facilitated by a peptidolytic contaminant present in the purified hα2M sample, as peptidases trapped within the induced tetrameric hα2M cage are known to still possess activity (Juanhuix et al., 2014; Lienart et al., 2018).

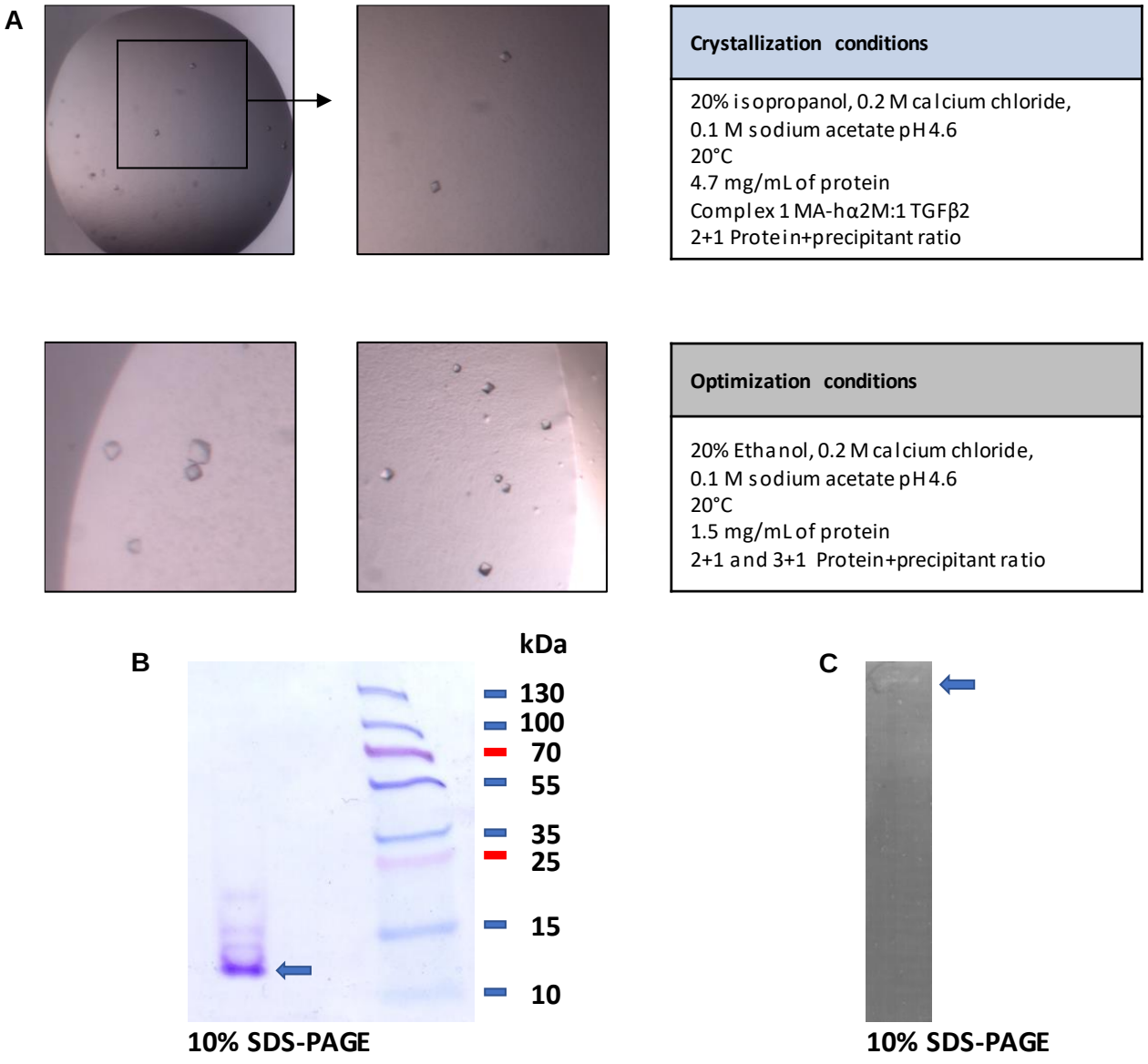


Figure 84. Crystallization of human TGFβ2 expressed in mammalian cells. (A) Representative tetragonal crystals of mature TGFβ2 of ~20 microns maximal dimension. Crystallization and optimization conditions are also shown. **(B)** 10% SDS-PAGE of ~90 collected, carefully washed and dissolved diffraction-grade crystals revealing they contain mature TGFβ2 (blue arrow). **(C)** Gelatin zymogram of pooled and purified crystallization drop supernatant showing a band pinpointed by an arrow corresponding to the mass of human α2-macroglobulin associated with gelatinolytic activity.

3.4.2.11 Structure of mature TGFβ2

The crystal structure was solved by maximum likelihood-scored molecular replacement, which gave a unique solution for one GF per asymmetric unit at final Eulerian angles and fractional cell coordinates (α, β, γ, x, y, and z) 348.4, 21.1, 119.9, 0.127, 0.696, and −0.899. The initial values for the rotation/translation function Z-scores were 6.0/14.0 and the final log-likelihood gain was 132. Taken together, these values indicated that P41212 was the correct space group, and the final model after model rebuilding and refinement contained residues A303-S414 and 23 solvent molecules. Segments P351-W354 and N371-S377 were flexible but clearly resolved in the final Fourier map for their main chains. Table 26 provides refinement and model validation statistics.

Dataset	Mature TGFβ2
Data processing	
Space group	$P4_12_12$
Cell constants (a and c, in Å)	55.57, 70.57
Wavelength (Å)	1.0332
No. of measurements/unique reflections	192,672/7,919
Resolution range (Å)	70.6–2.00 (2.12–2.00) [a]
Completeness (%)	100.0 (99.9)
R_{merge}	0.070 (2.495)
$R_{meas}/CC^{1/2}$	0.072 (2.546)/1.000 (0.870)
Average intensity	23.4 (1.7)
B-Factor (Wilson) (Å^2)/Aver. multiplicity	56.6/24.3 (24.8)
Structure refinement	
Resolution range used for refinement (Å)	43.7–2.00
No. of reflections used (test set)	7,511 (407)
Crystallographic R_{factor} (free R_{factor})	0.217 (0.253)
No. of protein residues/atoms/solvent molecules	112/890/23
Correlation coefficient F_{obs}-F_{calc} with all reflections/test set	0.943/0.938
***Rmsd* from target values**	
Bonds (Å)/angles (°)	0.010/1.18
Average B-factors (Å^2) (all/protein)	66.2/66.4
All-atom contacts and geometry analysis[b]	
Residues	
in favored regions/outliers/all residues	102 (93%)/0/110
outlying rotamers/bonds/angles/chirality/planarity	4/0/0/0/0
All-atom clashscore	1.7

Table 26. Crystallographic data of TGFβ2. [a]Data processing values in parenthesis are for the outermost resolution shell. [a]ccording to the wwPDB X-ray Structure Validation Report.

Human TGFβ2 GF is an elongated α/β-fold molecule consisting of an N-terminal helix α1 disemboguing into a fourfold antiparallel sheet of simple up-and-down connectivity (Fig. 86A). Each strand is subdivided into two, β1 + β2, β3 + β4, β5 + β6, and β7 + β8, and the sheet is slightly curled backwards with respect to helix α1. The most exposed segment of the moiety is the tip of hairpin β6β7, and β-ribbon β2β3 is linked by a second α-helix (α2) as part of a loop, whose conformation is mediated by the cis conformation of residue P338. Finally, a long loop segment connects strands β4 and β5 and contains helix α3, whose axis is roughly perpendicular to the β-sheet. TGFβ2 is internally crosslinked by four disulphides forming a cysteine knot (Fig. 86A) and a further intermolecular disulphide by symmetric C379 residues links two crystallographic symmetry mates to yield the functional dimer. Here, helix α3 of one protomer nestles into the concave face of the β-sheet of the other protomer (Fig. 86B). The respective tips of hairpins β6β7, as well as β-ribbons β2β3 with connecting helices α2, are exposed for functional interactions.

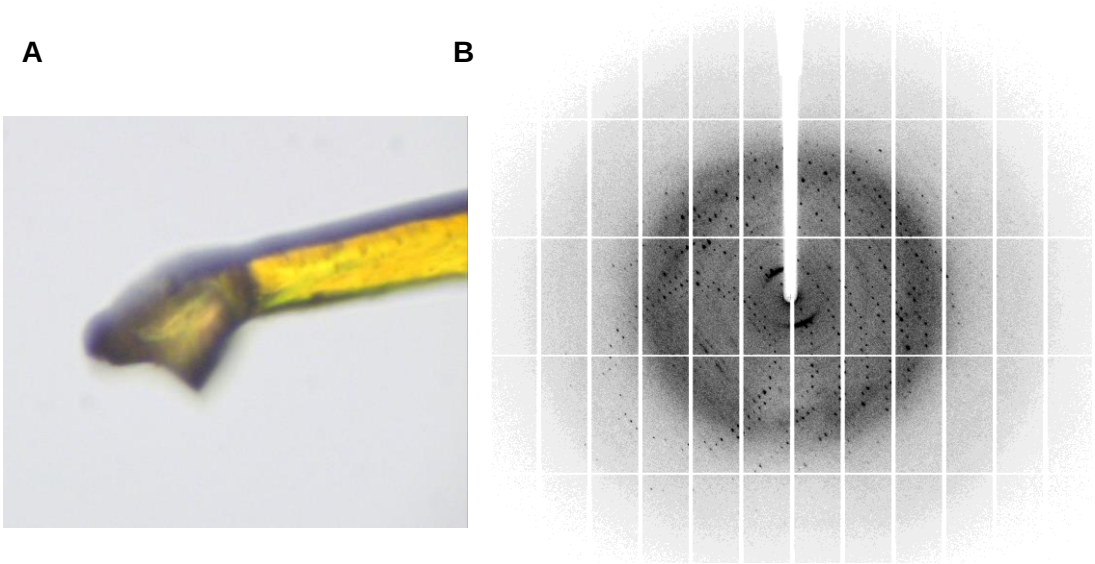


Figure 85. X-ray diffraction of mature TGFβ2. (A) Used loop to diffract the crystal**. (B)** The resolution at the horizontal borders of the detector is 1.8 Å.

3.4.2.11.1 Comparison with previous TGFβ2 structures

The structure of isolated human TGFβ2 GF was solved in 1992 by two groups simultaneously. They obtained the same trigonal crystal form with half a disulphide-linked dimer per asymmetric unit (Protein Data Bank access codes (PDB) 1TFG (Schlunegger and Grütter, 1992) and 2TGI (Daopin et al., 1992), see Table 27). In 2014, the complex of the GF with the Fab fragment of a neutralizing antibody was reported in an orthorhombic crystal form (PDB 4KXZ; Moulin et al., 2014). These crystals contained two GF dimers per asymmetric unit, each bound to two Fab moieties. Three more structures of TGFβ2

GF were reported in 2017 (Kim et al., 2017) (Table 27). These corresponded to engineered monomeric forms from human (PDB 5TX4; in complex with TGFR-II ectodomain) and mouse (PDB 5TX2 and 5TX6) that lacked helix α3 and encompassed several point mutations.

Superposition of the GF structures of PDB 2TGI, 1TFG, and 4KXZ (Fig. 86C) revealed very similar conformations of bound and unbound protomers and dimers, despite distinct chemical and crystallographic environments, which can generally contribute to different conformations owing to crystallographic artifacts (Janin and Rodier, 1995). The rmsd (root-mean-square distance) values with respect to PDB 2TGI (considered hereafter the reference structure) upon superposition of one protomer were 0.26 Å (1TFG), 1.05 Å (4KXZ dimer AB), and 1.22 Å (4KXZ dimer DE) for 112, 111, and 111 common Cα atoms, respectively. Only minor variations occurred at the tips of hairpins β6β7 and at helix α2 (Fig. 86C). Interestingly, also the engineered monomeric variants kept the overall structure of the native TGFβ2 protomer, with the exception of a loop that replaced helix α3 (Kim et al., 2017). This close similarity was also found with the unbound human ortholog TGFβ3 (PDB 1TGJ, Mittl et al., 1996).

Inversely, the present GF structure (PDB 6I9J; hereafter the current structure) showed deviations from the reference structure, which are reflected by a rmsd of 1.90 Å for 107 common Cα atoms. The central β-sheets nicely coincide, and only minor differences are found within the deviating regions of the above GF structures. However, a major rearrangement is found within segment Y352-C380, which encompasses helix α3 plus the flanking linkers to strands β4 and β5 (Fig. 86D, E). Owing to an outward movement of segment W354-D357, α3 is translated by maximally ~5.5 Å (at I370). This, in turn, causes downstream segment E373-C379 to be folded inward, with a maximal displacement of ~6.7 Å (at A374). This displacement cascades into a slight shift (~2 Å) of C379 and, thus, the intermolecular disulphide. These differences within protomers further result in variable arrangement of the dimers (Fig. 86F). The second protomer is rotated by ~10°, so that K396 at the tip of hairpin β6β7 is displaced by ~9 Å. In addition, the central β-sheets overlap but are laterally shifted with respect to each other.

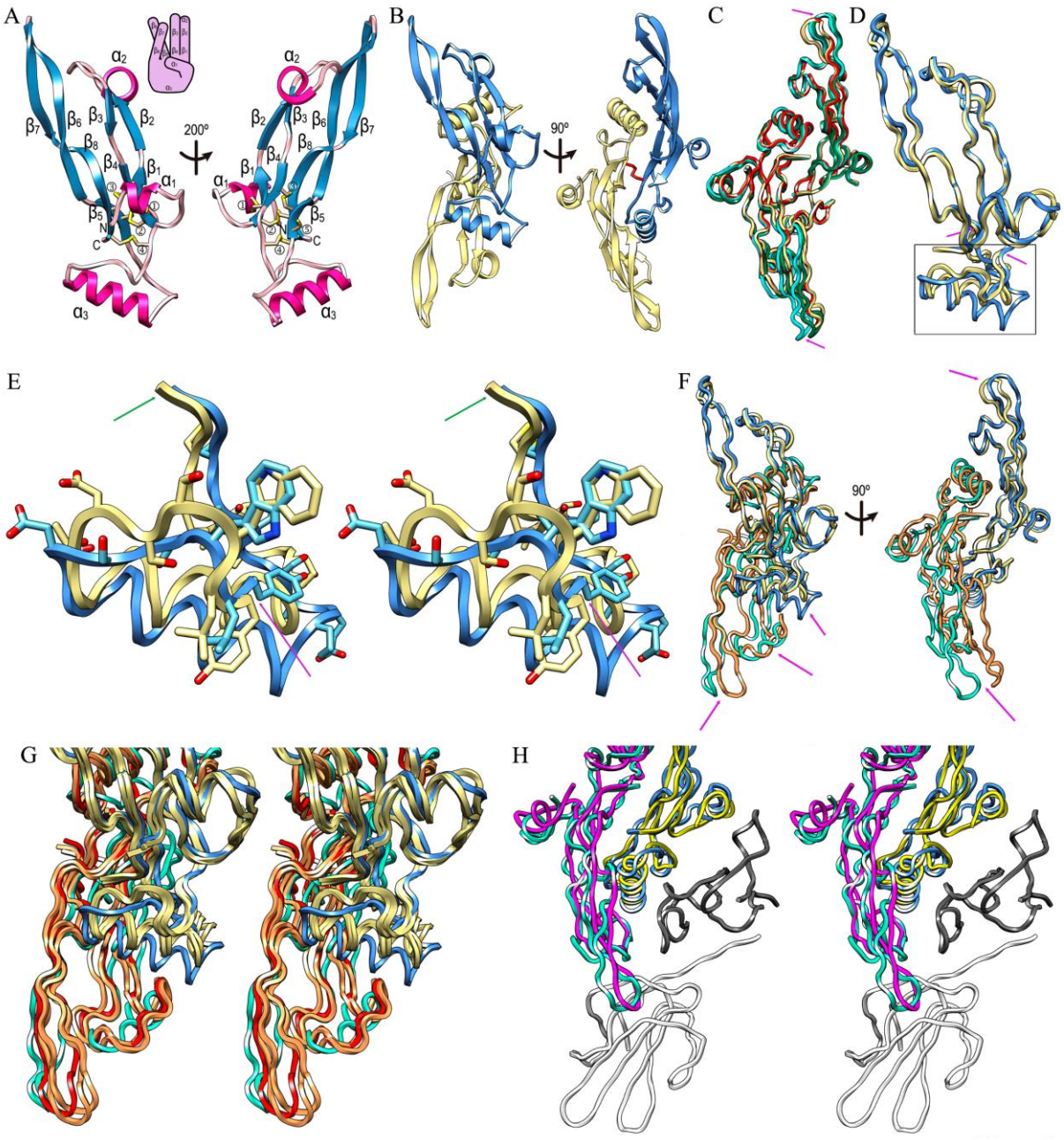
Finally, inspection of the respective crystal packings of the current and reference structures (Fig. 87A, B) reveals that while the latter is loosely packed in a hexagonal honeycomb-like arrangement, with large void channels of > 40 Å diameter and high solvent content (61%), the former is tightly packed (solvent content of 43%). In particular, the tip of hairpin β6β7 is not engaged in significant crystal contacts in either crystal form, which supports that both conformations are wild-type and do not result from crystallization artifacts.

PDB	Resolution (Å)	Residues [a]	No. of residues	No. of copies in a.u.	State (Isolated/ Complex)	Space group and cell constants (Å/°)	Organism	Reference	*rmsd* (Å) over residues[f]
1TFG	1.95	303–414	112	1	I	$P3_221$ a=b=60.60, c=75.20	Human	[57]	2.2/111
2TGI	1.80	303–414	112	1	I	$P3_221$ a=b=60.60, c=75.30	Human	[58]	2.3/111
4KXZ	2.83	303–414	112	4	C	$P2_12_12$ a=131.20, b=359.68 c=64.63	Human	[26]	2.0/111
5TX2	1.82	303–414	93[c]	2	I	C2 a=99.46, b=33.36 c=54.13, β=109.6	Mouse	[59]	1.3/91
5TX4	1.88	303–414	92[d]	1	C	$P2_12_12_1$ a=39.02, b=70.77 c=77.17	Human	[59]	1.2/91
5TX6	2.75	303–414	93[c]	3	I	$P3_121$ a=b=81.74, c=80.93	Mouse	[59]	0.8/88
5TY4	2.90[b]	317–413	65[e]	1	C	$P2_12_12_1$ a=41.53, b=71.33 c=79.51	Human	[86]	0.8/65
6I9J	2.00	303–314	112	1	I	$P4_12_12$ a=b=55.57, c=70.57	Human	This work	—

Table 27. Crystal structures of TGFβ2. [a]See UP entries P61812 and P27090 for human and mouse TGFβ2 sequences, respectively. [b]Obtained by crystal electron diffraction. [c]Mutant Δ354–373, K327R, R328K, L353R, A376K, C379S, L391V, I394V, K396R, T397K, and I400V. Contains an extra M at the N-terminus. [d]Mutant Δ354–373, K327R, R328K, L353R, A376K, C379S, L391V, I394V, K396R, T397K, and I400V. [e]Mutant Δ354– 378, K327R, and R328K. [f]Computed for the common Cα atoms of a protomer with the DALI program (Holm and Laakso, 2016) with respect to 6I9J.

Figure 86. **TGFβ2 in a new conformation.** (A) Ribbon-type plot of human TGFβ2 (left panel) and after vertical rotation (right panel). The eight β-strands β1 (residues C317-R320), β2 (L322-D325), β3 (G340-N342), β4 (F345-A347), β5 (C379-S382), β6 (D384-I394), β7 (T397-S404), and β8 (M406-S414), as well as the three helices α1 (A306-F310), α2 (F326-L330), and α3 (Q359-I370), are labeled, as are the N- and the C-terminus. The four intramolecular disulphides are depicted for their side chains and labeled ① (C309-C318), ② (C317-C380), ③ (C346-C411) and ④ (C350-C413). The cysteine engaged in a symmetric intermolecular disulphide (C379) is further labeled as ⑤. (B) Human mature TGFβ2 dimer with one protomer in the orientation of A (left panel) in blue and the second in pale yellow (left), which is related to the former through a horizontal crystallographic twofold axis. An orthogonal view is provided in the right panel, the intermolecular disulphide is depicted as red sticks. (C) Superposition of the Cα- traces in ribbon presentation of the dimers of previously reported structures of mature TGFβ2 (PDB 2TGI, pale yellow; PDB 1TFG, red; PDB 4KXZ dimer AB, green; and PDB 4KXZ dimer DE, aquamarine) after optimal superposition of the respective top protomers. Magenta arrows pinpoint the only points of significant deviation, i.e. the tips of respective β-ribbons β6β7. The view is that of B (right panel). (D) Superposition of the protomers of PDB 2TGI in pale yellow onto the current structure (PDB 6I9J) in the view of A (left panel). The region of largest deviation (Y352-C380) is pinpointed by magenta arrows and framed. (E) Close-up in cross-eye stereo of the framed region of (D), with ribbon and carbons in pale yellow for PDB 2TGI and in blue/cyan for PDB 6I9J. Y352 and C380 are pinpointed by a magenta and a green arrow, respectively. (F) Superposition of the dimers of PDB 6I9J (top protomer in blue, bottom protomer in aquamarine) and PDB 2TGI (top protomer in pale yellow, bottom protomer in orange) in the views of (B). Owing to the different chain traces of segment Y352-C380 (top magenta arrow in the left panel, see also [D]), substantial variations are observed at distal regions of the bottom protomers. (G) Close-up in stereo of F (left panel) depicting PDB 6I9J (dimer in blue/aquamarine); PDB 2TGI and 4KXZ dimer AB, both in pale yellow/orange; and human TGFβ3 as found in its complex with the ectodomains of human TGFR-I and -II (PDB 2PJY; protomers in pale yellow and red). (H) Superposition in stereo of the dimers of PDB 6I9J (dimer in blue/aquamarine) and human TGFβ3 (PDB 2PJY dimer in yellow/ magenta) in complex with the ectodomains of human TGFR-

I (dark grey) and -II (white) in the orientation of B (right panel). TGFβ2 in PDB 6I9J must rearrange to bind the receptors as performed by TGFβ3.



3.4.2.11.2 Potential implications of a new dimeric arrangement

Structural information on TGFR binding is available for human TGFβ1 and TGFβ3, whose wild-type forms were crystallized in complexes with the ectodomains of TGFR-I and -II (PDB 3KFD (Radaev et al., 2010); PDB 1KTZ (Hart et al., 2002); and PDB 2PJY (Groppe et al., 2008)). In the binary complex of TGFβ3 with TGFR-II (PDB 1KTZ), the cytokine was observed in a unique dimeric arrangement, termed “open”,

in which the respective helices α3 were completely disordered. This led the second protomer to be rotated by ~180° and the tip of hairpin β6β7 to be displaced by ~36 Å when compared with the reference structure (Hart et al., 2002). The functional significance of this structure, which differs from any other mature TGFβ dimer, is not clear.

By contrast, the other two complex structures, which contained both receptors, showed that the TGFβ1 and TGFβ3 isoforms dimerize very similarly to the unbound reference structure when bound to receptors (rmsd of 1.43 Å and 1.61 Å for 222 and 219 common Cα atoms for PDB 3KFD and PDB 2PJY, respectively). In contrast, they differed from the current structure (Fig. 86G), which shows a rmsd of 2.94 Å for 216 common Cα atoms when compared with the reference. In addition, both cytokine isoforms—and by similarity probably also TGFβ2—bound their receptors in an equivalent manner. TGFR-II was contacted by the tip of hairpin β6β7 and the linker between α2 and β3 of one protomer, and TGFR-I was liganded by the interface between protomers created by strands β5-β8 from one protomer and α1 plus the region around α3 of the other protomer. In addition, the TGFRs further contacted each other through the N-terminal extension of TGFR-II (Fig. 86H). Notably, the binding mode of TGFR-II was shared with the engineered monomeric variant of TGFβ2 (PDB 5TX4, Kim et al., 2017). In contrast, superposition of the cytokine dimers of the current structure and the triple complexes evinced that among other structural elements hairpin β6β7 is displaced away from interacting segment S52-E55 of TGFR-II (see PDB 2PJY) by ~2.7 Å. This hairpin rearrangement also impairs interaction of the β5β8 ribbon with TGFR-I segment I54-F60. Thus, TGFRs could not bind the current structure in the same way that they bind TGFβ1 and TGFβ3 without rearrangement

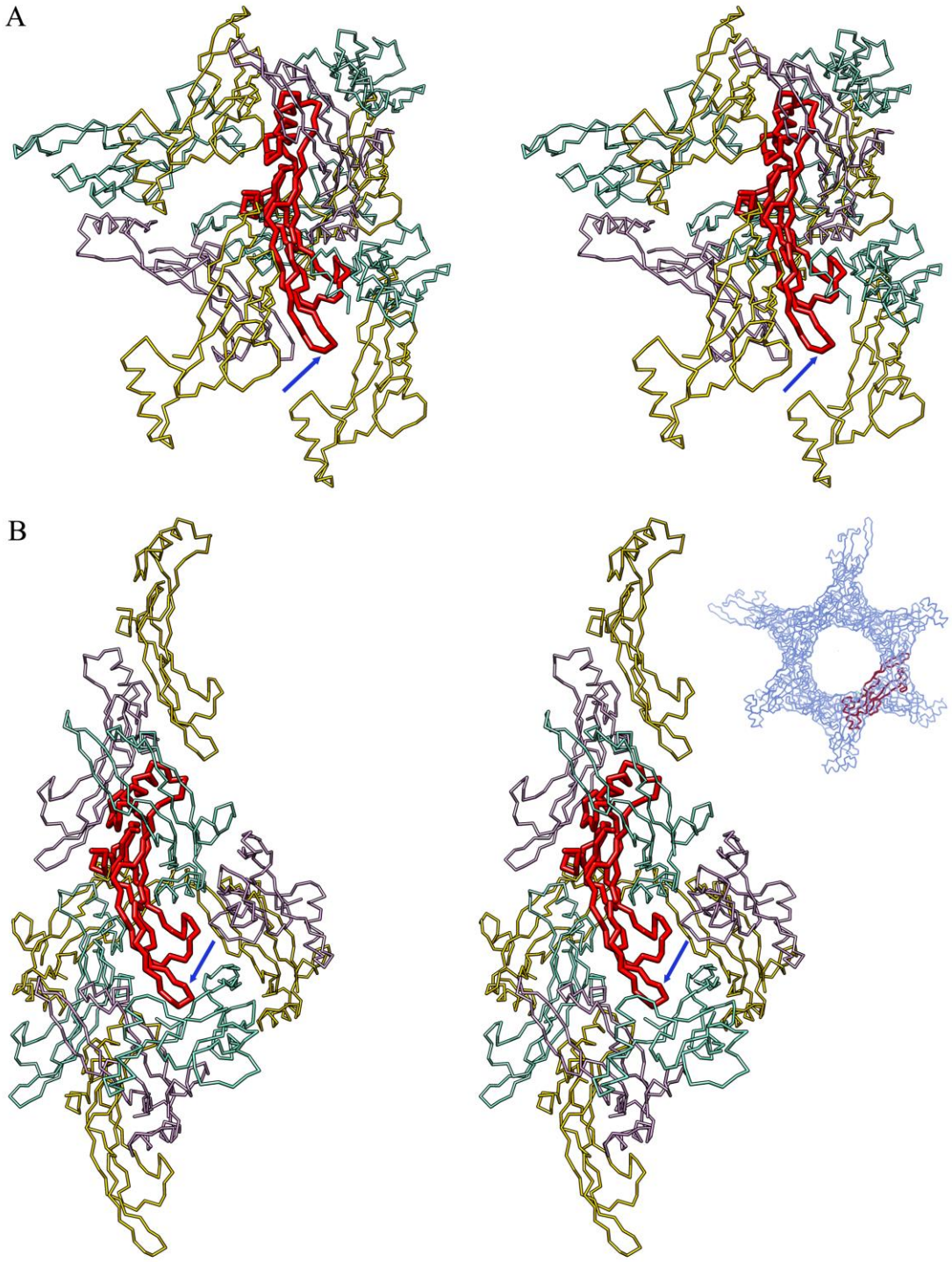


Figure 87. Crystal packing. (A) Cross-eye stereoplot showing the crystal packing of the current TGFβ2 structure (PDB 6I9J), with the protomer in the asymmetric unit in red and the surrounding symmetry mates in gold, plum and aquamarine. The tip of β-ribbon β6β7 is pinpointed by a blue arrow. **(B)** Same as (A) showing the crystal environment of the reference TGFβ2 structure (PDB 2TGI). The top right inset shows a view down of the crystallographic threefold axis to illustrate the solvent channels. The protomer in the asymmetric unit is in red, the symmetry mates setting up the crystal lattice in blue.

3.4.3 CONCLUSIONS AND FUTURE PERSPECTIVES

The hα2M protein is a sophisticated player to spatially and temporally restrict and regulate key physiological processes that control the distribution and activity of many proteins, including peptidases, cytokines, hormones and other physiological effectors (Garcia-Ferrer et al., 2017). Since the 1940s, several efforts have been made to understand its mechanism of action *in vivo* and *in vitro*, but they have been hampered by the unavailability of high-yield recombinant expression systems.

Here, we developed two eukaryotic systems, in insect (*Drosophila* Schneider 2) and mammalian (human Expi293F) cells, for expressing full-length protein and shorter fragments, and we obtained the highest yields reported to date. Pure recombinant hα2M yielded at ~1.0 and ~0.4 mg per liter of insect and mammalian cell culture, respectively. Expression of N-hα2M yielded ~1.6/~3.2 mg per liter of insect and mammalian cell culture, respectively, while the values for C-hα2M were ~1.3/~4.6 mg. The proteins were mainly produced in an induced state, possibly due to media components that cause induction during the time scale of expression. Thus, other media with a regulated composition should be assayed to re-evaluate the recombinant systems. The importance of glycosylation in the folding of recombinant hα2M proteins was also seen, losing yield and stability when they were mutated.

The insect and mammalian systems were also assayed for expressing the full-length pro-TGFβ2 encompassing LAP and mature TGFβ2 fragments which yielded ~2.7 mg and ~2.3 mg of N-terminally octahistidine-tagged and Strep-tagged forms, respectively, per liter of mammalian cell culture. The protein migrated as a dimer of ~110 kDa in SEC, which indicated that the disulphide bonds were formed between the LAP and the mature TGFβ2 moieties.

Whereas, full-length GRAB was expressed without the cell-wall anchoring region in a bacterial system yielding ~4 mg of pure protein per liter of expression medium and had an abnormal migration (Rasmusen et al., 1999), that was attributed to the highly unstructured character of the protein.

Regarding the studied protein interactions, the recombinant proteins *plus* wild-type hα2M (purified from blood) were analyzed by multiple techniques for binding with GRAB and pro-TGFβ2. We detected tightly complexes between native and methylamine-induced wild-type hα2M, with two molecules of GRAB per hα2M tetramer. The short variants, especially C-hα2M, likewise complexed GRAB, but apparently through a different mechanism from the full-length forms. In contrast, full-length pro-TGFβ2 did not complex any hα2M variant, probably owing to steric hindrance by the N-terminal LAP domain. However, we did detect the interaction of wild-type hα2M with mature TGFβ2. We also preliminary analyzed the interaction between recombinant C-hα2M and RBD proteins with CR34 and CR345 (from LRP1 receptor). However, the refolding of the numerous disulphide bonds in the CR domains was crucial to perform the binding assays. Therefore, other refolding strategies are suggested,

like coexpression and/or copurification with receptor-associated protein (RAP), which has seen to help in the refolding (Dolmer and Gettins, 2006).

We also performed many crystallization screens of recombinant hα2M fragments. Recombinant N-hα2M was forming dimers of ~170 kDa due to the presence of an intermolecular disulphide bond (C278–C431), which is also required for dimerization of the wild-type full-length protein, complicating its crystallization. C-hα2M was still too flexible to be crystallized; however, a crystal was obtained from the C-hα2M construct truncated before the flexible RBD domain. Thus, this approach should be continued with its expression in Expi293F cells, which give better protein yields, and it may be followed by deglycosylation of the protein with Kifunensine to help the crystallization.

The complexes between wild-type hα2M with GRAB and pro-TGFβ2 were also assayed by crystallization screens. For the first complex, we only crystallized induced wild-type hα2M at 20Å), and in the case of the second one, instead of crystallizing the complex, we crystallized the mature TGFβ2 fragment at 2Å). We attributed the separation of the LAP and GF moieties to the low pH of the crystallization assay, together with the crystallization process. This process was probably facilitated by a peptidolytic contaminant present in the purified hα2M sample, like peptidases trapped within the induced tetrameric hα2M cage, whose presence were demonstrated by gelatin zymography.

Mature TGFβ2 was in a different crystallographic space group, which mainly deviated in the region around helix α3 from current functional TGFβ1, TGFβ2 and TGFβ3 structures. Importantly, this region was the segment that showed the highest variability in sequence among TGFβs (Moulin et al. 2014). Although crystal packing artefacts could not be completely ruled out, the fact that several different crystal forms of the three TGFβ GFs had produced highly similar structures to date suggested that the new conformation may be wild-type and have functional implications. The divergent conformation of the region around helix α3 led to differences in the dimer, which could not bind cognate TGFR-I and -II in the same way as TGFβ1 and TGFβ3 do. Moreover, as revealed by pro-TGFβ1 structures from pig (PDB 5VQF) (Shi et al., 2011) and human (PDB 5VQP and PDB 6GFF) (Lienart S et al., 2018; Zhao et al., 2018), this region differed substantially between latent and mature moieties owing to the presence of the respective LAPs. Hence, the GF protomers associated differently within the respective dimers. Large differences were also found between the mature form and the pro-form of the more distantly related TGFβ family member activin A (PDB 2ARV and PDB 5HLY (Harrington et al., 2006; Wang et al., 2016). In contrast, bone morphogenetic factor 9 showed deviations in hairpin β6β7 but kept the dimer structure (PDB 5I05 and PDB 4YCG (Mi et al., 2015; Saito et al., 2017). Overall, we concluded that our mature structure may represent an inactive variant or be one of ensemble of conformational states, which may still undergo an induced fit or selection fit mechanism to form a functional ternary ligand-receptor complex.

Reproduced from todotelescopios.com

OVERALL CONCLUSIONS AND FUTURE PERSPECTIVES

In this thesis, two eukaryotic expression systems have been established. One from insect cells (*Drosophila melanogaster* Schneider 2) and another from mammalian cells (human cells, Expi293F). Many more have been tested (H5, HEK293T, CHO cells) in order to choose the best system that gives the highest protein yield performance. The cells were cultured in suspension to increase the density and the transient transfections with PEI were optimized. They expressed human glycoproteins with high yield, purity and homogeneity.

About the hSKI1's project , several fragments of hSKI1 (catalytic domain constructs of different lengths and the BTMD) were expressed in three protein expression systems (bacteria, insect and mammalian cells). In the case of bacteria, some constructs were successfully expressed, but they were unfortunately expressed intracellularly in inclusion bodies in all cases. Bacteria do not have the needed machinery to fold this protein. That happened independently of the signal peptide, tag, bacterial strain or culture conditions used for their expression. Despite of being expressed in inclusion bodies, we struggled to purify and fold the insoluble proteins with several protocols and with the last one (with L-arginine as additive), we set up crystallization plates, but protein crystals were never obtained.

In insects, the only good expression was by stably transfected lines of a truncated full-length of hSKI1 before the transmembrane domain (BTMD) and with a secretion signal peptide (pMT/GNBP3-hSKI1 (18-998)-H8 construct). It was secreted to the extracellular medium with two main bands of approximately 120 and 100 kDa and its estimated real protein yield was around 0.4 mg/L cell culture. This protein had many contaminants after affinity, size exclusion and anion chromatographies columns. In addition to that, it required a detergent to stabilise it and not precipitate into the chromatography columns. Thus, all these indicating that it was not well folded.

In mammalian cells, four constructs were engineered: one for the catalytic domain fragment and three for the BTMD expression. Regarding the BTMD constructs, two were ending in residue 998 (but with different signal peptides: the endogenous signal for intracellular expression and a secretion signal) and one ending in residue 947, which is before the shedding motif. Fortunately, this expression system was successful in obtaining soluble BTMD protein secreted to cellular medium. The catalytic domain fragment was once again not expressed, validating our theory about the luminal domain importance for folding and protein stability. When analysing the BTMD expressions in supernatants, we mainly detected differences in protein yield and purity. The only expressed construct with the endogenous signal peptide (pXLG40-hSKI1 (1-998)-H6) behaved with only one band about 100 kDa and with a 1 mg/L cell culture of very pure protein, which is the highest yield that has been published to date.

We concluded that a proper cleavage of the prodomain is crucial for trafficking and for obtaining this mature protein with full activity (which is the final step of this complex and the progressive

processing of the pro-hSKI1). In conclusion, the only way to obtain a proper folded hSKI1 protein secreted to the medium would be by expressing the full-length or the truncated form before the transmembrane domain with its endogenous signal peptide, therefore being, very promising for the pXLG40-hSKI1 (1-998)-H6 construct.

About the hα2M's project, we generated constructs for full-length protein and shorter fragments expression in insect and mammalian cells, with the highest yields reported to date. Pure recombinant hα2M yielded at ~1.0 and ~0.4 mg per liter of insect and mammalian cell culture, respectively. Expression of N-hα2M yielded ~1.6/~3.2 mg per liter of insect and mammalian cell culture, respectively, while the values for C-hα2M were ~1.3/~4.6 mg. However, the proteins were mainly produced in an induced state, possibly due to media components that cause induction during the time scale of expression.

The insect and mammalian systems were also assayed for expression of full-length pro-TGFβ2 encompassing LAP and mature TGFβ2 fragments and yielded ~2.7 mg and ~2.3 mg of N-terminally octahistidine-tagged and Strep-tagged forms, respectively, per liter of mammalian cell culture. Whereas, full-length GRAB was expressed without the cell-wall anchoring region in a bacterial system yielding ~4 mg of pure protein per liter of expression medium and had an abnormal migration in SDS-PAGE.

The protein interactions of the recombinant proteins *plus* wild-type hα2M (purified from blood) were analyzed by multiple techniques for binding with GRAB and pro-TGFβ2. We detected tight complexes between native and methylamine-induced wild-type hα2M, with two molecules of GRAB per hα2M tetramer. The short variants, especially C-hα2M, likewise complexed GRAB, but apparently through a different mechanism than with the full-length forms. In contrast, full-length pro-TGFβ2 did not complex any hα2M variant, probably owing to steric hindrance by the N-terminal LAP domain.

Many crystallization screens with recombinant hα2M fragments and wild-type hα2M in complex with GRAB and pro-TGFβ2 were performed. And finally, the mature TGFβ2 was crystallized in a different crystallographic space group at 2Å (PDB 6I9J), this was probably due to it being an inactive variant or being one of ensemble of conformational states of TGFβ2.

And finally, about the MMP's project, many potential MMP catalytic domains were found by database research in invertebrates, plants, fungi, viruses, protists, archaea and bacteria.

Altogether, this thesis has contributed substantially to the field at the expression, purification and crystallization levels on the serine protease hSKI1 and protease inhibitor hα2M, alone or interacting with GRAB and TGFβ2. In addition to this, a new variant structure of mature TGFβ2 and potential MMP

catalytic domains have been described, at atomic structure and database search levels, respectively. Moreover, two very efficient eukaryotic expression systems have been established at the laboratory.

GENERAL BIBLIOGRAPHY

A

Adachi H, Takano K, Morikawa M, Kanaya S, Yoshimura M, Mori Y, Sasaki T (2003). Application of a two-liquid system to sitting-drop vapour-diffusion protein crystallization. *Acta Crystallografica section D Biological Crystallography* 59(Pt 1):194-196.

Afonine PV, Grosse-Kunstleve RW, Echols N, Headd JJ, Moriarty NW, Mustyakimov M, Terwilliger TC, Urzhumtsev A, Zwart PH, Adams PD (2012). Towards automated crystallographic structure refinement with phoenix refine. Acta Crystallographica section D 68, 352–367.

Aggarwal RS (2014). What's fueling the biotech engine-2012 to 2013. *Nature Biotechnology, 32*(1): 32-39.

Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology, 37*(4): 420-423.

Altincicek B, Vilcinskas A (2008). Identification of a lepidopteran matrix metalloproteinase with dual roles in metamorphosis and innate immunity. *Developmental and Comparative Immunology,* 32: 400–409.

Altincicek B, Fischer M, Fischer M, Luersen K, Boll M, Wenzel U, Vilcinskas A (2010). Role of matrix metalloproteinase ZMP-2 in pathogen resistance and development in *Caenorhabditis elegans*. *Developmental and Comparative Immunology,* 34: 1160–1169.

Andersen OM, Christensen PA, Christensen LL, Jacobsen C, Moestrup SK, Etzerodt M, and Thøgersen HC (2000). Specific binding of α-macroglobulin to complement type repeat CR4 of the low-density lipoprotein receptor-related protein. *Biochemistry 39*, 10627-10633.

Andersen GR, Koch T, Dolmer K, Sottrup-Jensen L, Nyborg J (1995). Low resolution X-ray structure of human methylamine-treated alpha-2-macroglobulin. *Journal of Biological Chemistry 270*(42):25133–25141.

Andersen GR, Koch T, Sørensen AH, Thirup S, Nyborg J, Dolmer K, Jacobsen L, Sottrup-Jensen L (1994). Crystallisation of proteins of the α2-macroglobulin superfamily. *Annals of the New York Academy of Sciences 737,* 444–446.

Angerer L, Hussain S, Wei Z, Livingston BT (2006). Sea urchin metalloproteases: a genomic survey of the BMP-1/tolloid-like, MMP and ADAM families. *Developmental Biology,* 300: 267–281.

Annes, J. P., Munger, J. S. & Rifkin DB (2003). Making sense of latent TGF-β activation. *Journal of Cell Science 116*, 217–224.

Antoni C, Vera L, Devel L, Catalani MP, Czarny B, Cassar-Lajeunesse E, Nuti E, Rossello A, Dive V, Stura EA (2013). Crystallization of bi-functional ligand protein complexes. *Journal of Structural Biology,* 182: 246–254.

Apic G, Gough J, Teichmann SA (2001). Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *Journal of Molecular Biology,* 310 311–325.

Apte SS (2009). A disintegrin-like and metalloprotease (reprolysin-type) with thrombospondin type 1 motif (ADAMTS) superfamily: functions and mechanisms. *Journal of Biological Chemistry, 284*: 31493-31497.

Arandjelovic S, van Sant CL & Gonias SL (2006). Limited mutations in full-length tetrameric human α2-macroglobulin abrogate binding of platelet-derived growth factor-BB and transforming growth factor-β1. Journal of Biological Chemistry 281, 17061–17068.

Aravind L, Subramanian G (1999). Origin of multicellular eukaryotes - insights from proteome comparisons. *Current Opinion in Genetics and Development,* 9: 688–694.

Armstrong PB (2006). Proteases and protease inhibitors: a balance of activities in host-pathogen interaction. *Immunology, 211*: 263-281.

Armstrong PB (2010). Role of apha-2-macroglobulin in the immune responses of invertebrates. *Invertebrate Survival Journal 7*:165–180.

Armstrong PB and Quigley JP (1999). α2-macroglobulin: an evolutionarily conserved arm of the innate immune system. *Developmental & Comparative Immunology 23*, 375-390.

Arnold JN, Wallis R, Willis AC, Harvey DJ, Royle L, Dwek RA, Rudd PM, Sim RB (2006). Interaction of mannan binding lectin with alpha-2-macroglobulin via exposed oligomannose glycans: a conserved feature of the thiol ester protein family? *Journal of Biological Chemistry 281*(11):6955–6963.

Artenstein AW and Opal SM (2011). Proprotein convertases in health and disease. *The New England Journal of Medicine, 365*: 2507–2518.

Ashwell G, Harford J (1982). Carbohydrate-specific receptors of the liver. *Annual Review of Biochemistry, 51*: 531–554.

Axelsson P, Albandar JM, Rams TE (2002). Prevention and control of periodontal diseases in developing and industrialized nations. *Journal of Periodontology,* 2000 (29): 235–246.

B

Baca AM, Hol WG (2000). Overcoming codon bias: a method for high-level overexpression of Plasmodium and other AT-rich parasite genes in *Escherichia coli. International Journal of Parasitology*; 30(2):113-8.

Backliwal G, Hildinger M, Chenuet S, Wulhfard S, De Jesus M, Wurm FM (2008). Rational vector design and multi-pathway modulation of HEK 293E cells yield recombinant antibody titers exceeding 1g/l by transient transfection under serum-free conditions. *Nucleic Acids Research, 36*(15), e96.

Baldi L, Hacker DL, Adam M, Wurm FM (2007). Recombinant protein production by large-scale transient gene expression in mammalian cells: state of the art and future perspectives. Biotechnology Letters*, 29*(5): 677-684.

Barcelona PF and Saragovi HU (2015). A pro-nerve growth factor (proNGF) and NGF binding protein, a2-macroglobulin, differentially regulates p75 and TrkA receptors and is relevant to neurodegeneration ex vivo and in vivo. *Molecular and Cellular Biology*, vol. 35, no. 19, pp. 3396–3408.

Barrett AJ (1981). α2-Macroglobulin*. Methods in Enzymology. 80* (Pt C), 737-754.

Barrett AJ, Brown MA, Sayers CA (1979). The electrophoretically 'slow' and 'fast' forms of the alpha-2-macroglobulin molecule. *Biochemical Journal 181*:401–418.

Barrett AJ, McDonald JK (1985). Nomenclature: a possible solution to the 'peptidase anomaly'. *Biochemical Journal*231, 807.

Barrett AJ, Rawlings ND, Salvesen G, Fred Woessner J (2013). Introduction. *In Handbook of Proteolytic Enzymes*, Salvesen NDS (ed). Academic Press.

Barrett AJ and Starkey PM (1973). The interaction of α2-macroglobulin with proteinases. Characteristics and specificity of the reaction, and a hypothesis concerning its molecular mechanism. *Biochemical Journal 133*, 709-724.

Basak A, Bateman O, Slingsby C, Pande A, Asherie N, Ogun O, Benedek GB, Pande J (2003). High resolution X-ray crystal structures of human gammaD crystallin (1.25 A) and the R58H mutant (1.15 A) associated with aculeiform cataract. *Journal of Molecular Biology, 328*(5): 1137–1147.

Battesti A, Gottesman S (2013). Roles of adaptor proteins in regulation of bacterial proteolysis. *Current Opinion in Microbiology 16:* 140-147.

Becker-Pauly C, Stöcker W (2011). Insect Cells for Heterologous Production of Recombinant Proteins. In: Vilcinskas A. (eds) Insect Biotechnology. Biologically-Inspired Systems, vol 2. Springer, Dordrecht.

Bergmann M, Ross WF (1936). On proteolytic enzymes. The enzymes of papain and their activation. *Journal of Biological Chemistry, 114*: 717-726.

Berman H, Henrick K and Nakamura H (2003). Announcing the worldwide Protein Data Bank. *Nature Structural & Molecular Biology*. 10, 980–980.

Bertini I, Calderone V, Fragai M, Luchinat C, Mangan Si, Terni B (2004). Crystal structure of the catalytic domain of human matrix metalloproteinase 10. *Journal of Molecular Biology*. 336: 707–716.

Bertini I, Calderone V, Cosenza M, Fragai M, Lee YM, Luchinat C, Mangani S, Terni B, Turano P (2005). Conformational variability of matrix metalloproteinases: beyond a single 3D structure. *Proceedings of the National Academy of Sciences* U. S. A. 102: 5334–5339.

Betenbaugh MJ, Tomiya N, Narang S, Hsu JTA, Lee YC (2004). Biosynthesis of human-type N-glycans in heterollogous systems. *Current Opinion in Structural Biology, 14*(5): 601-606.

Bevec T, Stoka V, Pungercic G, Dolenc I, Turk V (1996). Major histocompatibility complex class II-associated p41 invariant chain fragment is a strong inhibitor of lysosomal cathepsin L. *Journal of Experimental Medicine, 183*, 1331–1338.

Beyer WR, Pöpplau D, Garten W, von Laer D and Lenz O (2003). Endoproteolytic processing of the lymphocytic choriomeningitis virus glycoprotein by the subtilase SKI-1/S1P. *Journal of Virology, 77*: 2866–2872.

Bieth J.G., Tourbez-Perrin M., and Pochon F (1981). Inhibition of α2-macroglobulin-bound trypsin by soybean trypsin inhibitor. *Journal of Biological Chemistry 256*, 7954-7957.

Birkenmeier G and Kopperschlager G (1994). *Methods in Enzymology 228*, 264–275.

Björk I and Fish WW (1982). Evidence for similar conformational changes in α2-macroglobulin on reaction with primary amines or proteolytic enzymes. *Biochemical Journal 207*, 347-356.

Blanchet M, Sureau C, Guévin C, Seidah NG and Labonté P (2015). SKI-1/S1P inhibitor PF-429242 impairs the onset of HCV infection. *Antiviral Research, 115*: 94–104.

Bloth B, Chesebro B, Svehag SE (1968). Ultrastructural studies of human and rabbit alpha-M-globulins. *Journal of Experimental Medicine 127*:749–762.

Bodvard K, Mohlin J, Knecht W (2006). Recombinant expression, purification, and kinetic and inhibitor characterisation of human site-1-protease. *Protein Expression and Purification, 51*(2): 308-319.

Bode W, Gomis-Rüth FX, Stöcker W (1993). Astacins, serralysins, snake venom and matrix metalloproteinases exhibit identical zinc-binding environments (HEXXHXXGXXH and Met-turn) and topologies and should be grouped into a common family, the ´metzincins'. *FEBS Letters*. 331: 134–140.

Bode W, Huber R (1992). Natural protein proteinase inhibitors and their interaction with proteinases. *European Journal of Biochemistry, 204*: 433–451.

Bode W, Huber R (2000). Structural basis of the endoproteinase–protein inhibitor interaction. *Biochimica et Biophysica Acta, 1477*: 241–252.

Boisset N, Taveau J-C, Pochon F, Lamy J (1996). Similar architectures of native and transformed human alpha-2-macroglobulin suggest the transformation mechanism. *Journal of Biological Chemistry 271*(42):25762–25769.

Borth W (1992). Alpha 2-macroglobulin, a multifunctional binding protein with targeting characteristics. *FASEB Journal; 6*(15):3345-53.

Brew K, Dinakarpandian D, Nagase H (2000). Tissue inhibitors of metalloproteinases: evolution, structure and function. *Biochimica et Biophysica Acta, 1477*: 267–283.

Broccanello C, Stevanato P, Biscarini F, Cantu D, Saccomani M (2015). A new polymorphism on chromosome 6 associated with bolting tendency in sugar beet. *BMC Genomics,* 16: 142.

Brooks SA (2004). Appropriate glycosylation of recombinant proteins for human use: implications of choice of expression system. *Molecular Biotechnology, 28*: 241–255.

Brooks SA (2006). Protein glycosylation in diverse cell systems: implications for modification and analysis of recombinant proteins. *Expert Review of Proteomics, 3*: 345–359.

Browner MF, Smith WW, Castelhano AL (1995). Matrilysin-inhibitor complexes: common themes among metalloproteases. *Biochemistry 34: 6602–6610.*

Bryan PN (2002). Prodomains and protein folding catalysis. *Chemical Reviews, 102*: 4805-4816.

Bryzek D, Ksiazek M, Bielecka E, Karim AY, Potempa B, Staniec D, Koziel J, Potempa J (2014). A pathogenic trace of Tannerella forsythia - shedding of soluble fully active tumor necrosis factor alpha from the macrophage surface by karilysin. *Molecular Oral Microbiology, 29: 294–306.*

Budd A, Blandin S, Levashina EA and Gibson TJ (2004). Bacterial α2-macroglobulins: colonization factors acquired by horizontal gene transfer from the metazoan genome? *Genome Biology. 5, R38.*

Bunch TA, Goldstein LS (1989). The conditional inhibition of gene expression in cultures *Drosophila* cells by antisense RNA. *Nucleic Acids Research, 17*: 9761-9782.

Bunch TA, Grinblat Y, Goldstein LSB (1988). Characterization and use of the *Drosophila* metallothionein promoter in cultured *Drosophila melanogaster* cells. *Nucleic Acids Research, 16*(3): 1043-1061.

Burda P, Aebi M (1999). The dolichol pathway of N-linked glycosylation. *Biochimica et Biophysica Acta, 1426*(2): 239-257.

Burri DJ, Pasqual G, Rochat C, Seidah NG, Pasquato A and Kunz S (2012). Molecular characterization of the processing of arenavirus envelope glycoprotein precursors by subtilisin kexin isozyme-1/site-1 protease. *Journal of Virology, 86:* 4935–4946.

Büssow K (2015). Stable mammalian producer cell lines for structural biology. *Current Opinion in Structural Biology, 32*: 81–90.

Butler M, Meneses-Acosta A (2012). Recent advances in technology supporting biopharmaceutical production from mammalian cells. *Applied Microbiology and Biotechnology, 96*(4): 885-894.

Byrne B, Donohoe GG, O'Kennedy R (2007). Sialic acids: carbohydrate moieties that influence the biological and physical properties of biopharmaceutical proteins and living cells. *Drug Discovery Today, 12*(7–8): 319–326.

C

Campbell C, Stanley P (1984). A dominant mutation to ricin resistance in Chinese hamster ovary cells induces UDP-GlcNAc:glycopeptide beta-4-N-acetylglucosaminyltransferase III activity. *Journal of Biological Chemistry*, 259: 13370–13378.

Cater JH, Wilson MR, Wyatt AR (2019). Alpha-2-Macroglobulin, a Hypochlorite-Regulated Chaperone and Immune System Modulator. *Oxidative Medicine and Cellular Longevity 2019*:5410657.

Cerdà-Costa N, Guevara T, Karim AY, Ksiazek M, Nguyen KA, Arolas JL, Potempa J, Gomis-Rüth FX (2011). The structure of the catalytic domain of Tannerella forsythia karilysin reveals it is a bacterial xenologue of animal matrix metalloproteinases. *Molecular Microbiology*, 79: 119–132.

Cerdà-Costa N, Gomis-Rüth FX (2014). Architecture and function of metallopeptidase catalytic domains. *Protein Science*, 23: 123–144.

Chang VT, Crispin M, Aricescu AR, Harvey DJ, Nettleship JE, Fennelly JA, Yu C, Boles KS, Evans EJ, Stuart DI, Dwek RA, Jones EY, Owens RJ, Davis SJ (2007). Glycoprotein structural genomics: solving the glycosylation problem. *Structure,* 15:267–273.

Cheng D, Espenshade PJ, Slaughter CA, Jaen JC, Brown MS, Goldstein JL (1999). Secreted site-1 protease cleaves peptides corresponding to luminal loop of sterol regulatory element-binding proteins. *The Journal of Biological Chemistry, 274*(32): 22805-22812.

Cherbas L, Cherbas P (2007). Transformation of *Drosophila* cell lines: an alternative approach to exogenous protein expression. *Methods in Molecular Biology, 388*: 317-340.

Cho CW, Chung E, Kim K, Soh HA, Jeong YK, Lee SW, Lee YC, Kim KS, Chung YS, Lee JH (2009). Plasma membrane localization of soybean matrix metalloproteinase differentially induced by senescence and abiotic stress. *The Journal of Plant Biology, 53 461–467.*

Chou CJ, Affolter M, Kussmann M (2012). A Nutrigenomics View of Protein Intake. Recent Advances in *Nutrigenetics and Nutrigenomics*, 51–74.

Chovar-Vera O, Valenzuela-Muñoz V, Gallardo-Escarate C (2015). Molecular characterization of collagen IV evidences early transcription expression related to the immune response against bacterial infection in the red abalone (Haliotis rufescens). *Fish and Shellfish Immunology*42: 241–248.

Chung YT, Keller EB (1990). Positive and negative regulatory elements mediating transcription from the *Drosophila melanogaster* actin 5C distal promoter. *Molecular and Cellular Biology, 10*(12): 6172-6180.

Claes H (1971). Autolysis of the cell wall of gametes of Chlamydomonas reinhardtii (German). *Archives of Microbiology*78: 180–188.

Clark IM, Thomas MD, de Vos S (2013). Chapter 177 — Plant Matrixins, in: N.D. Rawlings, G. Salvesen (Eds.), Handbook of Proteolytic Enzymes, vol. 1, *Academic Press, Oxford*, pp. 854–856.

Coates D, Siviter R, Isaac RE (2000). Exploring the Caenorhabditis elegans and *Drosophila melanogaster* genomes to understand neuropeptide and peptidase function. *Biochemical Society Transactions,* 28: 464–469.

Cohn EJ, Strong LE, Hughes WL, Mulford DL, Ashworth JN, Melin M, and Taylor HL (1946). Preparation and properties of serum and plasma proteins. IV. A system for the separation into fractions of the protein and lipoprotein components of biological tissues and fluids. *Journal of the American Chemical Society. 68*, 459-475.

Combier JP, Vernie T, de Billy F, El Yahyaoui F, Mathis R, Gamas P (2007). The MtMMPL1 early nodulin is a novel member of the matrix metalloendoproteinase family with a role in *Medicago truncatula* infection by *Sinorhizobium meliloti*. *Plant Physiology, 144: 703–716.*

Conrad HE (1998). Heparin-binding proteins, San Diego. USA: Academic Press.

Costa S, Almeida A, Castro A and Domingues L (2014). Fusion tags for protein solubility, purification and immunogenicity in *Escherichia coli:* the novel Fh8 system. *Frontiers in microbiology*, 5, 63.

Cotrin SS, Puzer L, de Souza Judice WA, Juliano L, Carmona AK, Juliano MA (2004). Positional-scanning combinatorial libraries of fluorescence resonance energy transfer peptides to define substrate specificity of carboxydipeptidases: assays with human cathepsin B. Analytical Biochemistry; 335(2):244-52.

Cray C, Zaias J, Altman NH (2009). Acute phase response in animals: a review. *Comparative Medicine 59*(6):517–526.

Cregg JM (2007). Introduction: distinctions between Pichia pastoris and other expression systems. *Methods in Molecular Biology*; 389:1-10.

Crookston KP, Webb DJ, Wolf B B and Gonias SL (1994). Classification of a2-Macroglobulin-Cytokine Interactions Based on Affinity of Noncovalent Association in Solution under Apparent Equilibrium Conditions. *Journal of Biological Chemistry 269*:1533-1540.

Croset A, Delafosse L, Gaudry JP, Arod C, Glez L, Losberger C, Begue D, Krstanovic A, Robert F, Vilbois F, Chevalet L, Antonsson B (2012). Differences in the glycosylation of recombinant proteins expressed in HEK and CHO cells. *Journal of Biotechnology*. 61, 336–348.

D

da Palma JR, Burri DJ, Oppliger J, Salamina M, Cendron L, de Laureto PP, Sidah NG, Kunz S and Pasquato A (2014). Zymogen activation and subcellular activity of subtilisin kexin isozyme 1/site 1 protease. *The Journal of Biological Chemistry, 289*(52): 35743–35756.

Daopin S, Piez KA, Ogawa Y and Davies DR (1992). Crystal structure of transforming growth factor-β 2: an unusual fold for the superfamily. *Science* 257, 369–373.

da Palma JR, Cendron L, Seidah NG, Pasquato A, Kunz S (2015). Mechanism of Folding and Activation of Subtilisin Kexin Isozyme-1 (SKI-1)/Site-1 Protease (S1P). *The Journal of Biological Chemistry*, *291*(5): 2055-66.

Daly NL, Scanlon MJ, Djordjevic JT, Kroon PA, Smith R (1995). Three-dimensional structure of a cysteine-rich repeat from the low-density lipoprotein receptor. *Proceedings of the National Academy of Sciences* U S A, 92: 6334–6338.

Daramola O, Stevenson J, Dean G, Hatton D, Pettman G, Holmes W, Field R (2014). A high-yielding CHO transient system: coexpression of genes encoding EBNA-1 and GS enhances transient protein expression. *Biotechnology Progress, 30*(1): 132–141.

Davis BJ (1964). Disc electrophoresis. II. Method and application to human serum proteins. *Annals of the New York Academy of Sciences*, 28;121:404-27.

Debanne MT, Bell R and Dolovich J (1976). Characteristics of the macrophage uptake of proteinase-α-macroglobulin complexes. *Biochimica et Biophysica Acta 428*, 466-475.

De Jesus M, Wurm FM (2011). Manufacturing recombinant proteins in kg-ton quantities using animal cells in bioreactors. *European Journal of Pharmaceutics and Biopharmaceutics, 78*(2): 184-188.

De Martin R, Haendler B, Hofer-Warbinek R, Gaugitsch H, Wrann M, Schlüsener H, Seifert JM, Bodmer S, Fontana A, and Hofer E (1987). Complementary DNA for human glioblastoma-derived T cell suppressor factor, a novel member of the transforming growth factor-β gene family. *EMBO Journal* 6, 3673–3677.

del Amo-Maestro L, Marino-Puertas L, Goulas T & Gomis-Rüth FX (2019). Recombinant production, purification, crystallization, and structure analysis of human transforming growth factor β2 in a new conformation. *Scientific Reports 9*, 8660.

Delorme VGR, McCabe PF, Kim DJ, Leaver CJ (2000). A matrix metalloproteinase gene is expressed at the boundary of senescence and programmed cell death in cucumber. *Plant Physiology, 123 917–927.*

Deu E, Verdoes M, Bogyo M (2013). New tools for dissecting protease function: implications for inhibitor design, drug discovery and probe development. *Nature Structural & Molecular Biology19*(1): 9–16.

Deveraux Q, Takahashi R, Salvesen GS, Reed JC (1997). X-linked IAP is a direct inhibitor of cell death proteases. *Nature 388*: 300–304.

Dewhirst FE, Chen T, Izard J, Paster BJ, Tanner AC, Yu WH, Lakshmanan A, Wade WG (2010). The human oral microbiome. *Journal of Bacteriology192 5002–5017.*

Doan N and Gettins PGW (2007). Human α2-macroglobulin is composed of multiple domains, as predicted by homology with complement component C3. *Biochemical Jounal, 407:* 23-30.

Doan N, Gettins PG (2008). Alpha-Macroglobulins are present in some gram-negative bacteria: characterization of the alpha-2-macroglobulin from Escherichia coli. *Jounal of Biological Chemistry 283*(42):28747–28756.

Dolmer K, Gettins PG (2006). Three complement-like repeats compose the complete alpha-2-macroglobulin binding site in the second ligand binding cluster of the low density lipoprotein receptor-related protein. Jounal of Biological Chemistry 281(45):34189–34196.

Doolittle RF, Feng DF, Anderson KL, Alberro MR (1990). A naturally occurring horizontal gene transfer from a eukaryote to a prokaryote. *Journal of Molecular Evolution,* 31: 383–388*.*

Drugmand JC, Schneider YJ, Agathos SN (2012). Insect cells as factories for biomanufacturing. *Biotechnology Advances, 30*(5): 1140-1157.

Dunn JT, Spiro RG (1967a). The alpha-2 macroglobulin of human plasma. Isolation and composition. *Journal of Biological Chemistry 242*(23):5549–5555.

Dyson MR, Shadbolt SP, Vincent KJ, Perera RL, McCafferty J (2004). Production of soluble mammalian proteins in *Escherichia coli*: identification of protein features that correlate with successful expression. *BMC Biotechnology*, 4: 32.

E

Ecker DM1, Jones SD, Levine HL (2015). The therapeutic monoclonal antibody market. *MAbs, 7*(1): 9-14.

Eder J, Fersht AR (1995). Pro-sequence-assisted protein folding. *Mol Microbiol 16*: 609-614.

Eigenbrot C, Kirchhofer D (2002). New insight into how tissue factor allosterically regulates Factor VIIa. *Trends Cardiovasc Med, 12*: 19–26.

Elagoz A, Benjannet S, Mammarbassi A, Wickham L and Seidah NG (2001). Biosynthesis and cellular trafficking of the convertase SKI-1/S1P: ectodomain shedding requires SKI-1 activity. *The Journal of Biological Chemistry, 277*: 11265–11275.

Elbein AD, Tropea JE, Mitchell M, Kaushal GP (1990). Kifunensine, a potent inhibitor of the glycoprotein processing mannosidase I. *Journal of Biological Chemistry,* 265(26): 15599-15605.

Emsle, P, Lohkamp B, Scott WG and Cowtan, K (2010). Features and development of Coot. *Acta Crystallographica*, section D 66, 486–501.

Enghild JJ, Salvesen G, Thøgersen IB and Pizzo SV (1989). Proteinase binding and inhibition by the monomeric α-macroglobulin rat α1-inhibitor-3. *Journal of Biological Chemistry 264*, 11428-11435.

Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A (2006). Comparative protein structure modeling using MODELLER. *Current Protocols in Bioinformatics* (Suppl. 15) 5.6.1–5.6.30.

F

Fanjul-Fernández M, Folgueras AR, Cabrera S, López-Otín C (2010). Matrix metalloproteinases: evolution, gene regulation and functional analysis in mouse models. *Biochimica et Biophysica Acta 1803: 3–19.*

Feige JJ, Negoescu A, Keramidas M, Souchelnitskiy S, Chambaz EM (1996). Alpha 2-macroglobulin: a binding protein for transforming growth factor-beta and various cytokines. *Hormone research 45*(3-5):227-322.

Farady CJ, Craik CS (2010). Mechanisms of Macromolecular Protease Inhibitors. *ChemBioChem, 11*: 2341-2346.

Fernandez FJ, Vega MC (2013). Technologies to keep an eye on: alternative hosts for protein production in structural biology. *Current Opinion in Structural Biology, 23*(3): 365–373.

Fernandez JM, Hoeffler JP (1999). Gene expression systems: using nature for the art of expression. San Diego; London: *Academic Press*. 480 p.

Flinn BS (2008). Plant extracellular matrix metalloproteinases. *Funct. Plant Biol. 35 1183–1193.*

French K, Yerbury JJ and Wilson MR (2008). Protease activation of α2-macroglobulin modulates a chaperone-like action with broad specificity. *Biochemistry*, vol. 47, no. 4, pp. 1176–1185.

Frenzel A, Hust M and Schirrmann T(2003). Expression of recombinant antibodies. *Frontiers in Immunology*, 4: 217

Fujinaga M, Cherney mm, Oyama H, Oda K, James MN (2004). The molecular structure and catalytic mechanism of a novel carboxyl peptidase from *Scytalidium lignicolum*. *Proceedings of the National Academy of Sciences USA, 101:* 3364-3369.

Fuller RS, Brake A and Thorner J (1989). Yeast prohormone processing enzyme (KEX2 gene product) is a Ca2+-dependent serine protease. *Proceedings of the National Academy of Sciences USA, 86:* 1434–1438.

G

Gaci N, Borrel G, Tottey W, O'Toole PW, Brugere JF (2014). Archaea and the human gut: new beginning of an old story. *World Journal of Gastroenterology,* 20: 16062–16078.

Gall AL, Ruff M, Kannan R, Cuniasse P, Yiotakis A, Dive V, Rio MC, Basset P, Moras D (2001). Crystal structure of the stromelysin-3 (MMP-11) catalytic domain complexed with a phosphinic inhibitor mimicking the transition-state. *Journal of Molecular Biology. 307: 577–586.*

Gallagher T, Gilliland G, Wang L and Bryan P (1995). The prosegment-subtilisin BPN' complex: crystal structure of a specific 'foldase'. *Structure*, 3(9):907-914.

Galliano MF, Toulza E, Gallinaro H, Jonca N, Ishida-Yamamoto A, Serre G, Guerrin M (2006). A novel protease inhibitor of the alpha-2-macroglobulin family expressed in the human epidermis. *Journal of Biological Chemistry 281*(9):5780–5789.

Ganrot P.O. and Schersten B. (1967). Serum α2-macroglobulin concentration and its variation with age and sex. *Clinica Chimica Acta 15*, 113-120.

Garcia-Ferrer I (2015). Structural and functional studies on *Escherichia coli* α2-macroglobulin: a snap-trap peptidase inhibitor (PhD tesis). Barcelona University, Barcelona, Spain.

Garcia-Ferrer I, Arêde P, Gómez-Blanco J, Luque D, Duquerroy S, Castón JR, Goulas T and Gomis-Rüth FX (2015). Structural and functional insights into Escherichia coli α2-macroglobulin endopeptidase snap-trap inhibition. *Proceedings of the National Academy of Sciences of the United States of America 112*, 8290-8295.

Garcia-Ferrer, I., Marrero, A., Gomis-Rüth, F X. & Goulas, T (2017). α2-Macroglobulins: structure and function. *Subcellular. Biochemistry. 83*, 149–183 (2017).

Garton MJ, Keir G, Lakshmi MV, and Thompson EJ (1991). Age-related changes in cerebrospinal fluid protein concentrations. *Journal of the Neurological Sciences*, vol. 104, no. 1, pp. 74–80, 1991.

Gaudry JP, Arod C, Sauvage C, Busso S, Dupraz P, Pankiewicz R, Antonsson B (2008). Purification of the extracellular domain of the membrane protein Glial-CAM expressed in HEK and CHO cells and comparison of the glycosylation. *Protein Expression and Purification, 58*, 94–102.

Geisse S, Voedisch B (2012). Transient expression technologies: past, present, and future. In: Voynov V, Caravella JA (eds) *Therapeutic proteins, vol 899*. Humana Press, New York, pp 203–219.

Gettins PG (2002). Serpin structure, mechanism, and function. *Chemical Reviews 102*(12): 4751-4804.

Ghetie MA, Uhr JW, Vitetta ES (1991). Covalent binding of human alpha-2-macroglobulin to deglycosylated ricin A chain and its immunotoxins. *Cancer Research 51*:1482–1487.

Ghiglione C, Lhomond G, Lepage T, Gache C (1994). Structure of the sea urchin hatching enzyme gene. *European Journal of Biochemistry,* 219: 845–854.

Gillich A (2005). Baculovirus-Insektenzellen-Expressionssystem. Vienna: University of natural resources and life sciences, Vienna. Bachelor: 26.

Godehardt A, Hammerschmidt S, Frank R & Chhatwal GS (2004). Binding of α2-macroglobulin to GRAB (Protein G-related α2-macroglobulin binding protein), an important virulence factor of group A streptococci, is mediated by two charged motifs in the ΔA region. *Biochemical Journal* 381, 877–885.

Golldack D, Popova OV, Dietz KJ (2002). Mutation of the matrix metalloproteinase At2-MMP inhibits growth and causes late flowering and early senescence in Arabidopsis. *Journal of Biological Chemistry*, 277: 5541–5547.

Gomis-Rüth FX (2003). Structural aspects of the metzincin clan of metalloendopeptidases *Molecular Biotechnology*, 24: 157–202.

Gomis-Rüth FX (2004). Hemopexin domains, in: A. Messerschmidt, W. Bode, M. Cygler (Eds.), Handbook of Metalloproteins, vol. 3, John Wiley & Sons, Ltd., Chichester (UK), pp. 631–646.

Gomis-Rüth FX (2009). Catalytic domain architecture of metzincin metalloproteases *Journal of Biological Chemistry,* 284: 15353–15357.

Gomis-Rüth FX, Botelho TO, Bode W (2012). A standard orientation for metallopeptidases. *Biochimica et Biophysica Acta* 1824: 157–163.

Gomis-Rüth FX, Trillo-Muyo S, Stöcker W (2012). Functional and structural insights into astacin metallopeptidases. *Biological Chemistry,* 393: 1027–1041.

Gomis-Rüth FX, Kress LF, Bode W (1993). First structure of a snake venom metalloproteinase: a prototype for matrix metalloproteinases/collagenases. *EMBO Journal*, 12: 4151–4157.

Gomis-Rüth FX, Maskos K, Betz M, Bergner A, Huber R, Suzuki K, Yoshida N, Nagase H, Brew K, Bourenkov GP, Bartunik H, Bode W (1997). Mechanism of inhibition of the human matrix metalloproteinase stromelysin-1 by TIMP-1. *Nature 4*; 389(6646): 77-81.

Gonçalves RN, Gozzini Barbosa SD, da Silva-López RE (2016). Proteases from Canavalia ensiformis: active and thermostable enzymes with potential of application in biotechnology. *Biotechnology Research International*, 3427098.

Gonias SL, Carmichael A, Mettenburg JM, Roadcap DW, Irvin WP and Webb DJ (2000) Identical or overlapping sequences in the primary structure of human alpha(2)-macroglobulin are responsible for the binding of nerve growth factor-beta, platelet-derived growth factor-BB, and transforming growth factor-beta. *The Journal of Biological Chemistry*, vol. 275, no. 8, pp. 5826–5831.

Gottesman MM (1987). Chinese hamster ovary cells. In: Gottesman M (ed) *Methods in enzymology*, vol 151. Academic Press Inc, San Diego, pp 3–8.

Goulas T, Arolas JL, Gomis-Rüth FX (2010). Structure, function and latency regulation of a bacterial enterotoxin potentially derived from a mammalian adamalysin/ADAMxenolog. Proceedings of the National Academy of Sciences, 108(5): 1856-1861.

Goulas T, Arolas JL, Gomis-Rüth FX (2011). Structure, function and latency regulation of a bacterial enterotoxin potentially derived from a mammalian adamalysin/ ADAM xenolog. *Proceedings of the National Academy of Sciences U.S.A.* 108: 1856–1861.

Goulas T, Cuppari A, Garcia-Castellanos R, Snipas S, Glockshuber R, Arolas JL, Gomis-Rüth FX (2014a). The pCri System: a vector collection for recombinant protein expression and purification. *PLoS One, 9*(11): e112643.

Goulas T, Garcia-Ferrer I, Garcia-Pique S, Sottrup-Jensen L & Gomis-Rüth FX (2014b). Crystallization and preliminary X-ray diffraction analysis of eukaryotic α2-macroglobulin family members modified by methylamine, proteases and glycosidases. *Molecular Oral Microbiology 29*, 354–364.

Goulas, T, Garcia-Ferrer I, Marrero A, Marino-Puertas L, Duqueroy S, Gomis-Rüth FX (2017). Structural and functional insight into pan-endopeptidase inhibition by α2-macroglobulins. *Journal of Biological Chemistry 398*, 975–994.

Goulas T, Gomis-Rüth FX (2013). 186. Fragilysin, in: N.D. Rawlings, G.S. Salvesen (Eds.), Handbook of Proteolytic Enzymes, Academic Press, Oxford, Great Britain, pp. 887–891.

Goulielmaki E, Siden-Kiamos I, Loukeris TG (2014). Functional characterization of Anopheles matrix metalloprotease 1 reveals its agonistic role during sporogonic development of malaria parasites. *Infection and Immunity* 82 (2014) 4865–4877.

Graham FL, Smiley J, Russell WC, Nairn R (1977). Characteristics of a human cell line transformed by DNA from human adenovirus type 5. *Journal of General Virology, 36*(1): 59–74.

Graham JS, Xiong J, Gillikin JW (1991). Purification and developmental analysis of a metalloendoproteinase from the leaves of Glycine max. *Plant Physiology,.* 97 786–792.

Granados RR, Li GX, Derksen ACG, Mckenna KA. 1994. A new insect cell line from *Trichoplusia ni* (BTI-Tn-5B1-4) susceptible to *Trichoplusia ni* single enveloped nuclear polyhedrosis virus. *Journal of Invertebrate Pathology, 64*(3): 260-266.

Grassmann W and Dyckerhoff H (1928). Hoppe-Seyler's Z. *Physiol Chem Journal.* 179: 41-78.

Green D (2006). Coagulation cascade. Hemodialysis International. *International Symposium on Home Hemodialysis, 10*, Suppl 2:S2-4.

Grishaev A, Tugarinov V, Kay LE, Trewhella J, Bax A (2008). Refined solution structure of the 82 kDa enzyme malate synthase G from joint NMR and synchrotron SAXS restraints. *Journal of Biomolecular NMR, 40*(2): 95–106.

Grishaev A, Ying J, Canny MD, Pardi A, Bax A (2008). Solution structure of tRNAVal from refinement of homology model against residual dipolar coupling and SAXS data. *Journal of Biomolecular NMR, 42*(2):99–109.

Groppe J, Hinck CS, Samavarchi-Tehrani P, Zubieta C, Schuermann JP, Taylor AB, Schwarz PM, Wrana JL, Hinck AP (2008). Cooperative assembly of TGF-β superfamily signaling complexes is mediated by two disparate mechanisms and distinct modes of receptor binding. *Molecular Cell* 29, 157–168.

Gross J, Lapière CM (1962). Collagenolytic activity in amphibian tissues: a tissue culture assay. Proceedings of the National Academy of Sciences U. S. A. 48: 1014–1022.

Guan JM, Bing L, Wang D, Liu CL, Shen WD (2009). Cloning, sequence analysis and expression of a matrix metalloproteinase gene (Bm-MMP) in the silkworm, Bombyx mori. *Acta Entomol. Sinica,* 52: 353–362.

Guevara T, Ksiazek M, Skottrup PD, Cerda-Costa N, Trillo-Muyo S, de Diego I, Riise E, Potempa J, Gomis-Rüth FX (2013). Structure of the catalytic domain of the *Tannerella forsythia* matrix

metallopeptidase karilysin in complex with a tetrapeptidic inhibitor. *Acta Crystallographica, Section F,* 69: 472–476.

H

Hacker DL, Kiseljak D, Rajendra Y, Thurnheer S, Baldi L, Wurm FM (2013). Polyethyleneimine-based transient gene expression processes for suspension adapted HEK-293E and CHO-DG44 cells. *Protein Expression and Purification, 92*(1): 67–76.

Haider SR, Sharp BL and Reid HJ (2011). A comparison of Tris-glycine and Tris-tricine buffers for the electrophoretic separation of major serum proteins. *Separation Science and Technology* 34, 2463–2467.

Hallmann A, Amon P, Godl K, Heitzer M, Sumper M (2001). Transcriptional activation by the sexual pheromone and wounding: a new gene family from Volvox encoding modular proteins with (hydroxy)proline-rich and metalloproteinase homology domains. *Plant Journal,* 26: 583–593.

Hanahan D (1983). Studies on transformation of Escherichia coli with plasmids. *Journal of Molecular Biology,* 166: 557-580.

Harrington AE, Morris-Triggs SA, Ruotolo BT, Robinson CV, Ohnuma S, Hyvönen M (2006). Structural basis for the inhibition of activin signalling by follistatin. *EMBO Journal* 25, 1035–1045.

Harrison RL, Jarvis DL (2007a). Transforming lepidopteran insect cells for continuous recombinant protein expression. *Methods in Molecular Biology, 388*: 299-316.

Harrison RL, Jarvis DL (2007b). Transforming lepidopteran insect cells for improved protein processing. *Methods in Molecular Biology, 388*: 341-56.

Hart PJ, Deep S, Taylor AB, Shu Z, Hinck CS, Hinck AP (2002). Crystal structure of the human TβR2 ectodomain–TGF-β3 complex. *Nature Structural Biology,* 9, 203–208.

Heitzer M, Hallmann A (2002). An extracellular matrix-localized metalloproteinase with an exceptional QEXXH metal binding site prefers copper for catalytic activity. *Journal of Biological Chemistry, 277 28280–28286.*

Hemsley A, Arnheim N, Toney MD, Cortopassi G and Galas DJ (1989). A simple method for site-directed mutagenesis using the polymerase chain reaction. *Nucleic Acids Research* 17, 6545–6551.

Holm L and Laakso LM (2016). Dali server update. *Nucleic Acids Research*. 44, W351–W355.

Huang SS, O'Grady P and Huang J S (1988). Human transforming growth factor-β-α2Macroglobulin complex is a latent form of transforming growth factor-β. *Journal of Biological Chemistry. 263*, 1535-1411.

Hunt I (2005). From gene to protein: a review of new and enabling technologies for multi-parallel protein expression. *Protein Expression and Purification, 40*(1): 1-22.

Huxley-Jones J, Clarke TK, Beck C, Toubaris G, Robertson DL, Boot- Handford RP (2007). The evolution of the vertebrate metzincins; insights from Ciona intestinalis and Danio rerio. *BMC Evol. Biol.* 7 63.

I

Imber MJ & Pizzo SV (1981). Clearance and binding of two electrophoretic "fast" forms of human α2-macroglobulin.*Journal of Biological Chemistry*. 256, 8134–8139.

Indra D, Ramalingam K, Babu M (2005). Isolation, purification and characterization of collagenase from hepatopancreas of the land snail *Achatina fulica*. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology.* 142 1–7.

Ishimwe E, Hodgson JJ, Passarelli AL (2015). Expression of the Cydia pomonella granulovirus matrix metalloprotease enhances Autographa californica multiple nucleopolyhedrovirus virulence and can partially substitute for viral cathepsin. *Virology,* 481: 166–178.

Isolani ME, Abril JF, Salo E, Deri P, Bianucci AM, Batistoni R (2013). Planarians as a model to assess in vivo the role of matrix metalloproteinase genes during homeostasis and regeneration. *PLoS One,* 8 e55649.

J

Janin J and Rodier F (1995). Protein-protein interaction at crystal contacts. *Proteins* 23, 580–587.

Jensen PE, Sottrup-Jensen L (1986). Primary structure of human alpha-2-macroglobulin. *Journal of Biological Chemistry 261*(34):15863–15869.

Jordan M, Schallhorn A, Wurm FM (1996). Transfecting mammalian cells: optimization of critical parameters affecting calcium-phosphate precipitate formation. *Nucleic Acids Research, 24*(4): 596–601.

Jordans S, Jenko-Kokalj S, Kuhl NM, Tedelind S, Sendt W, Bromme D, Turk D, Brix K (2009). Monitoring compartment-specific substrate cleavage by cathepsins B, K, L, and S at physiological pH and redox conditions. *BMC Biochemistry, 10*: 23.

Jozic D, Bourenkov G, Lim NH, Visse R, Nagase H, Bode W, Maskos K (2005). X-ray structure of human proMMP-1: new insights into procollagenase activation and collagen binding. *Journal of Biological Chemistry,* 280 9578–9585.

Juanhuix, J, Gil-Ortiz F, Cuní G, Colldelram C, Nicolás J, Lidón J, Boter E, Ruget C, Ferrer S, Benach J (2014). Developments in optics and performance at BL13-XALOC, the macromolecular crystallography beamline at the ALBA synchrotron. *Journal of Synchrotron Radiation* 21, 679-689.

Jusko M, Potempa J, Karim AY, Ksiazek M, Riesbeck K, Garred P, Eick S, Blom AM (2012). A metalloproteinase karilysin present in the majority of Tannerella forsythia isolates inhibits all pathways of the complement system. *Journal of Immunology.* 188 2338–2349.

K

Kang C, Han DY, Park KI, Pyo MJ, Heo Y, Lee H, Kim GS, Kim E (2014). Characterization and neutralization of Nemopilema nomurai (Scyphozoa: Rhizostomeae) jellyfish venom using polyclonal antibody. *Toxicon,* 86: 116–125.

Kantor AM, Dong S, Held NL, Ishimwe E, Passarelli AL, Clem RJ, Franz AWE (2017). Identification and initial characterization of matrix metalloproteinases in the yellow fever mosquito, *Aedes aegypti*. *Insect Molecular Biology.* 26 113–126.

Kao FT, Puck TT (1968). Genetics of somatic mammalian cells, VII. Induction and isolation of nutritional mutants in Chinese hamster cells. *Proceedings of the National Academy of Sciences USA, 60*(4): 1275–1281.

Karim AY, Kulczycka M, Kantyka T, Dubin G, Jabaiah A, Daugherty PS, Thogersen IB, Enghild JJ, Nguyen KA, Potempa J (2010). A novel matrix metalloprotease- like enzyme (karilysin) of the periodontal pathogen *Tannerella forsythia* ATCC 43037. *Journal of Biological Chemistry,* 391: 105–117.

Kato N, Kano M, Tani Y and Ogata K (1974). Purification and Characterization of Formate Dehydrogenase in a Methanol-utilizing Yeast, Kloeckera sp. No. 2201. *Agricultural and Biological Chemistry*, 38:1, 111-116, DOI: 10.1080/00021369.1974.10861128

Kim SK, Barron L, Hinck CS, Petrunak EM, Cano KE, Thangirala A, Iskra B, Brothers M, Vonberg M, Leal B, Richter B, Kodali R, Taylor AB, Du S, Barnes CO, Sulea T, Calero G, Hart PJ, Hart MJ, Demeler B, Hinck AP. (2017). An engineered transforming growth factor beta (TGF-β) monomer that functions as a dominant negative to block TGF-β signaling. *The Journal of Biological Chemistry*, 292: 7173–7188.

Kinoshita T, Fukuzawa H, Shimada T, Saito T, Matsuda Y (1992). Primary structure and expression of a gamete lytic enzyme in Chlamydomonas reinhardtii: similarity of functional domains to matrix metalloproteases. *Proceedings of the National Academy of Sciences U. S. A.* 89: 4693–4697.

Kirkpatrick RB1, Matico RE, McNulty DE, Strickler JE, Rosenberg M (1995). An abundantly secreted glycoprotein from Drosophila melanogaster is related to mammalian secretory proteins produced in rheumatoid tissues and by activated macrophages. *Gene, 153*(2): 147-154.

Kirti Rani, Rachita Rana, Sanchi Datt (2012). Review on latest overview of proteases. *International Journal of Current Life Sciences, 2*(1): 12– 18.

Kling J (2014). Fresh from the biotech pipeline--2013. *Nature Biotechnology, 32*(2): 121-124.

Knorr E, Schmidtberg H, Vilcinskas A, Altincicek B (2009). MMPs regulate both development and immunity in the tribolium model insect. *PLoS One,* 4: e4751.

Ko R, Okano K, Maeda S (2000). Structural and functional analysis of the *Xestia c-nigrum* granulovirus matrix metalloproteinase. *Journal of Virology.* 74 11240–11246.

Kohno T, Hochigai H, Yamashita E, Tsukihara T, Kanaoka M (2006). Crystal structures of the catalytic domain of human stromelysin-1 (MMP-3) and collagenase-3 (MMP-13) with a hydroxamic acid inhibitor SM-25453. Biochemical and Biophysical Research Communications. 344: 315–322.

Kolodziej SJ, Wagenknecht T, Strickland DK, Stoops JK (2002). The three-dimensional structure of the human alpha 2-macroglobulin dimer reveals its structural organization in the tetrameric native and chymotrypsin alpha 2-macroglobulin complexes. *Journal of Biological Chemistry, 277*(31): 28031-28037.

Kost TA, Condreay JP, Jarvis DL (2005). Baculovirus as versatile vectors for protein expression in insect and mammalian cells. *Nature Biotechnology*, 23(5):567-75.

Kost TA, Ignar DM, Clay WC, Andrews J, Leray JD, Overton L, Hoffman CR, Kilpatrick KE, Ellis B, Emerson DL (1997). Production of a urokinase plasminogen activator-IgG fusion protein (uPA-IgG) in the baculovirus expression system. *Gene, 190*(1): 139-144.

Kotnik T, Weaver JC (2016). Abiotic gene transfer: rare or rampant? *J. Membr. Biol.* 249 623–631.

Kovaleva ES, Masler EP, Skantar AM, Chitwood DJ (2004). Novel matrix metalloproteinase from the cyst nematodes *Heterodera glycines* and *Globodera rostochiensis*. *Molecular and Biochemical Parasitology,* 136: 109–112.

Koziel J, Karim AY, Przybyszewska K, Ksiazek M, Rapala-Kozik M, Nguyen KA, Potempa J (2010). Proteolytic inactivation of LL-37 by karilysin, a novel virulence mechanism of *Tannerella forsythia*. *Journal of Innate Immunity,* 2 288–293.

Krasnow MA, Saffman EE, Kornfeld K, Hogness DS (1989). Transcriptional activation and repression by Ultrabithorax proteins in cultures *Drosophila* cells. *Cells, 57*: 1031-1043.

Krimbou L, Tremblay M, Davignon J and Cohn JS (1998). Association of apolipoprotein E with α2-macroglobulin in human plasma. *Journal of Lipid Research*, vol. 39, no. 12, pp. 2373–2386.

Krysan DJ, Rockwell NC, Fuller RS (1999). Quantitative characterization of furin specificity. Energetics of substrate discrimination using an internally consistent set of hexapeptidyl methylcoumarinamides*. Journal of Biological Chemistry*, 274: 23229–23234.

Krissinel E and Henrick K (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica* section D 60, 2256–2268.

Krysan DJ, Rockwell NC, Fuller RS (1999). Quantitative characterization of furin specificity. Energetics of substrate discrimination using an internally consistent set of hexapeptidyl methylcoumarinamides. *Journal of Biological Chemistry,* 274 23229–23234.

Ksiazek M, Mizgalska D, Eick D, Thøgersen IB, Enghild JJ, Potempa J (2015). KLIKK proteases of *Tannerella forsythia*: putative virulence factors with a unique domain structure. *Frontiers in Microbiology,* 6: 312.

Kubelka V, Altmann F, Kornfeld G, Marz L (1994). Structures of the N-linked oligosaccharides of the membrane glycoproteins from three lepidopteran cell lines (Sf-21, IZD-Mb-0503, Bm-N). *Archives of Biochemistry and Biophysics, 308*(1): 148-157.

Kubo T, Saito T, Fukuzawa H, Matsud. Y (2001). Two tandemly-located matrix metalloprotease genes with different expression patterns in the Chlamydomonas sexual cell cycle. *Current Genetics, 40: 136–143.*

Kunz S, Edelmann KH, de la Torre JC, Gorney R and Oldstone MB (2003). Mechanisms for lymphocytic choriomeningitis virus glycoprotein cleavage, transport, and incorporation into virions. *Virology, 314*: 168–178.

Kyte J, Doolittle RF (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology, 157*(1): 105-132.

L

LaMarre J, Hayes AM, Wollenberg GK, Hussaini I, Hall SW, Gonias SL (1991a). An apha-2-macrolgobulin receptor-dependent mechanism for the plasma clearance of transforming growth factor-beta-1 in mice. *Journal of Clinical Investigation 87*:39–44.

LaMarre J, Wollenberg GK, Gonias SL, and Hayes MA (1991b). Cytokine binding and clearance properties of proteinase activated alpha 2-macroglobulins. *Laboratory Investigation*, vol. 65, no. 1, pp. 3–14.

Lang R, Braun M, Sounni NE, Noel A, Frankenne F, Foidart JM, Bode W, Maskos K (2004). Crystal structure of the catalytic domain of MMP-16/MT3-MMP: characterization of MT-MMP specific features. Journal of Molecular Biology. 336: 213–225.

Langer G, Cohen SX, Lamzin VS and Perrakis A (2008). Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nature Protocols* 3, 1171–1179.

Larsson AM, Ståhlberg J, Jones TA (2002). Preparation and crystallization of selenomethionyl dextranase from Penicillium minioluteum expressed in Pichia pastoris. Acta Crystallographica, section D, Biological Crystallography. 2002 Feb;58(Pt 2):346-8.

Laskowski MJ, Kato I (1980). Protein inhibitors of proteinases. *Annual Review of Biochemistry49*: 593–626.

Law SK, Dodds AW (1997). The internal thioester and the covalent binding properties of the complement proteins C3 and C4. *Protein Science 6*(2):263-274.

Lee SJ, Evers S, Roeder D, Parlow AF, Risteli J, Risteli L, Lee YC, Feizi T, Langen H, Nussenzweig MC (2002). Mannose receptor-mediated regulation of serum glycoprotein homeostasis. *Science, 295*: 1898–1901.

Lenz O, ter Meulen J, Klenk HD, Seidah NG and Garten W (2001). The Lassa virus glycoprotein precursor GP-C is proteolytically processed by subtilase SKI-1/S1P. *Proceedings of the National Academy of Sciences U.S.A. 98*: 12701–12705.

Leontovich AA, Zhang J, Shimokawa K, Nagase H, Sarras MP Jr. (2000). A novel hydra matrix metalloproteinase (HMMP) functions in extracellular matrix degradation, morphogenesis and the maintenance of differentiated cells in the foot process. *Development,* 127: 907–920.

Leuven FV, Cassiman JJ, Van den Berghe H (1981). Functional modifications of alpha 2-macroglobulin by primary amines. I. Characterization of alpha 2 M after derivatization by methylamine and by factor XIII. *Journal of Biological Chemistry 256*(17):9016-22.

Li D, Zhang H, Song Q, Wang L, Liu S, Hong Y, Huang L, Song F (2015). Tomato Sl3- MMP, a member of the matrix metalloproteinase family, is required for disease resistance against Botrytis cinerea and Pseudomonas syringae pv. tomato DC3000. *BMC Plant Biology.* 15 143.

Libson AM, Gittis AG, Collier IE, Marmer BL, Goldberg GI, Lattman EE (1995). Crystal structure of the haemopexin-like C-terminal domain of gelatinase A. *Nat. Struct. Biol.* 2 938–942.

Lienart S, Merceron R, Vanderaa C, Lambert F, Colau D, Stockis J, van der Woning B, De Haard H, Saunders M, Coulie PG, Savvides SN, Lucas S (2018). Structural basis of latent TGF-β1 presentation and activation by GARP on human regulatory T cells. *Science* 362, 952-956.

Lindsley CW (2013). The top prescription drugs of 2012 globally: biologics dominate, but small molecule CNS drugs hold on to top spots. *ACS Chemical Neuroscience, 4*(6): 905-907.

Liu C, Dalby B, Chen W, Kilzer JM, Chiou HC (2008). Transient transfection factors for high-level recombinant protein production in suspension cultured mammalian cells. *Molecular Biotechnology, 39*(2): 141–153.

Liu CY, Graham JS (1998). Cloning and characterisation of an Arabidopsis thaliana cDNA homologous to the matrix metalloproteinases (The Electronic Plant Gene register PGR 98–130). *Plant Physiology.* 117 1127.

Liu Q, Ling TY, Shieh HS, Johnson FE, Huang JS, Huang SS (2001). Identification of the high affinity binding site in transforming growth factor-β involved in complex formation with α2-macroglobulin. Implications regarding the molecular mechanisms of complex formation between α2-macroglobulin and growth factors, cytokines, and hormones. *Journal of Biological Chemistry 276*, 46212–46218.

Liu Y, Dammann C, Bhattacharyya MK (2001). The matrix metalloproteinase gene GmMMP2 is activated in response to pathogenic infections in soybean. *Plant Physiology.* 127 1788–1797.

Llano E, Pendas AM, Aza-Blanc P, Kornberg TB, López-Otín C (20000). Dm1-MMP, a matrix metalloproteinase from Drosophila with a potential role in extracellular matrix remodeling during neural development. *Journal of Biological Chemistry.* 275 35978–35985.

Llano E, Adam G, Pendas AM, Quesada V, Sánchez LM, Santamaria I, Noselli S, López-Otín C (2002). Structural and enzymatic characterization of Drosophila Dm2-MMP, a membrane-bound matrix metalloproteinase with tissue-specific expression. *Journal of Biological Chemistry.* 277 (2002) 23321–23329.

Long JZ and Cravatt BF (2011). The metabolic serine hydrolases and their functions in mammalian physiology and disease. *Chemical Reviews, 111*: 6022–6063.

López-Otín C, Bond JS (2008). Proteases: multifunctional enzymes in life and disease. *Journal of Biological Quemistry, 283*: 30433-30437.

López-Otín C, Palavalli LH, Samuels Y (2009). Protective roles of matrix metalloproteinases: from mouse models to human cancer. *Cell Cycle 8* 3657–3662.

López-Pelegrín M, Ksiazek M, Karim AY, Guevara T, Arolas JL, Potempa J, Gomis-Rüth FX (2015). A novel mechanism of latency in matrix metalloproteinases. *Journal of Biological Chemistry.* 290 4728–4740.

Lovejoy B, Welch AR, Carr S, Luong C, Broka C, Hendricks RT, JA Campbell, Walker KA, Martin R, Van Wart H, Browner MF (1999). Crystal structures of MMP-1 and -13 reveal the structural basis for selectivity of collagenase inhibitors. *Nature Structural & Molecular Biology*. 6: 217–221.

Lukassen MV, Scavenius C, Thøgersen IB, Enghild JJ (2016). Disulphide bond pattern of transforming growth factor β-induced protein. *Biochemistry, 55*(39): 5610-5621.

Lynch SV, Pedersen O (2016). The human intestinal microbiome in health and disease, N. Engl. J*. Med.* 375 2369–2379*.*

M

Maidment JM, Moore D, Murphy GP, Murphy G, Clark IM (1999). Matrix metalloproteinase homologues from Arabidopsis thaliana. Expression and activity. *Journal of Biological Chemistry.* 274 34706–34710*.*

Mandal MK, Fischer R, Schillberg S, Schiermeyer A (2010). Biochemical properties of the matrix metalloproteinase NtMMP1 from Nicotiana tabacum cv. BY-2 suspension cells. *Planta* 232 899–910.

Mannello F, Canesi L, Gazzanelli G, Gallo G (2001)/ Biochemical properties of metalloproteinases from the hemolymph of the mussel Mytilus galloprovincialis Lam. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology.* 128 507–515.

Marchal I, Jarvis DL, Cacan R, Verbert A (2001). Glycoproteins from insect cells: sialylated or not? *Journal of Biological Chemistry382*(2): 151–159.

Marino G, Funk C (2012). Matrix metalloproteinases in plants: a brief overview. *Physiol. Plant.* 145 196–202.

Marino G, Huesgen PF, Eckhard U, Overall CM, Schroder WP, Funk C (2014). Family-wide characterization of matrix metalloproteinases from Arabidopsis thaliana reveals their distinct proteolytic activity and cleavage site specificity. *Biochem. J*. 457 335–346.

Marquardt H., Lioubin MN and Ikeda T (1987). Complete amino acid sequence of human transforming growth factor type β 2. *Journal of Biological Chemistry* 262, 12127–12131.

Marrero A, Duquerroy S, Trapani S, Goulas T, Guevara T, Andersen GR, Navaza J, Sottrup-Jensen L, Gomis-Rüth XF (2012). The crystal structure of human alpha-2-macroglobulin reveals a unique molecular cage. *Angewandte Chemie International Edition 51*:3340–3344.

Marschner K, Kollmann K, Schweizer M, Braulke T and Pohl S (2011). A key enzyme in the biogenesis of lysosomes is a protease that regulates cholesterol metabolism. *Science, 333*: 87–90.

Martin W, Cerff R (1986). Prokaryotic features of a nucleus-encoded enzyme. Cdna sequences for chloroplast and cytosolic glyceraldehyde-3-phosphate dehydrogenases from mustard (Sinapis alba). *European Journal of Biochemistry.* 159 323–331.

Masroori N, Halabian R, Mohammadipour M, Roushandeh AM, Rouhbakhsh M, Najafabadi AJ, Fathabad ME, Salimi M, Shokrgozar MA, Roudkenar MH (2010). High level expression of functional recombinant human coagulation factor VII in insect cells. Biotechnology Letters*, 32*(6): 803-809.

Massova I, Kotra LP, Fridman R, Mobashery S (1998). Matrix metalloproeinases: structures, evolution, and diversification. *FASEB J.* 12 1075–1095.

Massova I, Mobashery S (1998). Kinship and diversification of bacterial penicillinbinding proteins and β-lactamases. *Antimicrob. Agents Chemother.* 42 1–17.

Matsuda Y (2013). Chapter 187 - Gametolysin, in: N.D. Rawlings, G. Salvesen (Eds.), Handbook of Proteolytic Enzymes, vol. 1, Elsevier Ltd., pp. 891–895.

Matthijs G, Devriendt K, Cassiman JJ, Van der Berghe H and Marynen P (1992). Structure of the human α2-acroglobulin gene and its promoter. *Biochemical and Biophysical Research Communications 184*, 596-603.

McCoy A, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ (2007). Phaser crystallographic software. *Journal of Applied Crystallography*, 40, 658–674.

McGeehan G, Burkhart W, Anderegg R, Becherer JD, Gillikin JW, Graham JS (1992). Sequencing and characterization of the soybean leaf metalloproteinase: structural and functional similarity to the matrix metalloproteinase family. *Plant Physiology.* 99 1179–1183.

McHenry JZ, Christeller JT, Slade EA, Laing WA (1996). The major extracellular proteinases of the silverleaf fungus, Chondrostereum purpureum, are metalloproteinases. *Plant Pathol.* 45 552–563.

Means JC, Passarelli AL (2010). Viral fibroblast growth factor, matrix metalloproteases, and caspases are associated with enhancing systemic infection by baculoviruses. *Proceedings of the National Academy of SciencesU. S. A.* 107 9825–9830.

Mi L.Z, Brown CT, Gao Y, Tian Y, Le VQ, Walz T, Springer TA (2015). Structure of bone morphogenetic protein 9 procomplex. *Proceedings of the National Academy of Sciences* USA 112, 3710–3715.

Mitchell KJ, Pinson KI, Kelly OG, Brennan J, Zupicich J, Scherz P, Leighton PA, Goodrich LV, Lu X, Avery BJ, Tate P, Dill K, Pangilinan E, Wakenight P, Tessier-Lavigne M, Skarnes WC (2001). Functional analysis of secreted and transmembrane proteins critical to mouse development. *Nature Genetics, 28*(3): 241-249.

Mittal R, Patel AP, Debs LH, Nguyen D, Patel K, Grati M, Mittal J, Yan D, Chapagain P, Liu XZ (2016). Intricate functions of matrix metalloproteinases in physiological and pathological conditions. *Journal of Cellular Physiology*. 231: 2599–2621.

Mittl PR, Priestle JP, Cox DA, McMaster G, Cerletti N, Grütter MG (1996). The crystal structure of TGF-β3 and comparison to TGF-β2: implications for receptor binding. Protein Science 5, 1261–1271.

Mivazono K, Hellman U, Wernstedt C and Heldin CH (1988). Latent high molecular weight complex of transforming growth factor-β1: Purification from human platelets and structural characterization. 1. *Journal of Biological Chemistry 263*, 6107-6415.

Miyamoto N, Yoshida MA, Koga H, Fujiwara Y. Genetic mechanisms of bone digestion and nutrient absorption in the bone-eating worm Osedax japonicus inferred from transcriptome and gene expression analyses. *BMC Evol. Biol.* 17 17.

Moncrief JS, Obiso RJ, Barroso LA, Kling JJ, Wright RL, Van Tassell RL, Lyerly DM, Wilkins TD (1995). The enterotoxin of Bacteroides fragilis is a metalloprotease. *Infection and Immunity* 63 175–181.

Moraes AM, Jorge SA, Astray RM, Suazo CA, Calderón Riquelme CE, Augusto EF, Tonso A, Pamboukian MM, Piccoli RA, Barral MF, Pereira CA (2012). *Drosophila melanogaster* S2 cells for expression of heterologous genes: From gene cloning to bioprocess development. *Biotechnology Advances, 30*(3): 613-628.

Morrison CJ, Butler GS, Rodríguez D, Overall CM (2009). Matrix metalloproteinase proteomics: substrates, targets, and therapy. Current Opinion in Cell Biology, 21, 645–653.

Moulin A, Mathieu M, Lawrence C, Bigelow R, Levine M, Hamel C, Marquette JP, Le Parc J, Loux C, Ferrari P, Capdevila C, Dumas J, Dumas B, Rak A, Bird J, Qiu H, Pan CQ, Edmunds T, Wei RR. (2014). Structures of a pan-specific antagonist antibody complexed to different isoforms of TGFβ reveal structural plasticity of antibody-antigen interactions. *Protein Science.* 23, 1698–1707.

Murphy GJ, Murphy G, Reynolds JJ (1991). The origin of matrix metalloproteinases and their familial relationships. *FEBS Lett.* 289 4–7.

Müller H-P, Rantamäki LK (1995). Binding of native alpha-2-macroglobulin to human group G Streptococci. *Infection and Immunity 63*(8):2833–2839.

N

Nagase H, Woessner JF Jr. (1999). Matrix metalloproteinases. *Journal of Biological Chemistry.* 274 21491–21494.

Nagase H, Visse R, Murphy G (2006). Structure and function of matrix metalloproteinases and TIMPs. *Cardiovasc. Res.* 69 562–573.

Nakajima T, Iwaki K, Kodama T, Inazawa J and Emi M (2000). Genomic structure and chromosomal mapping of the human Site-1 protease (S1P) gene. *Journal of Human Genetics, 45*: 212–217.

Nagase H, Visse R, Murphy G (2006) Structure and function of matrix metalloproteinases and TIMPs. *Cardiovascular Research*, 69 (2006) 562–573.

Nagase H, Woessner Jr JF (1999). Matrix metalloproteinases. *Journal of Biological Chemistry* 274: 21491–21494.

Nallet S, Fornelli L, Schmitt S, Parra J, Baldi L, Tsybin YO, Wurm FM (2012). Glycan variability on a recombinant IgG antibody transiently produced in HEK-293E cells. *New Biotechnol, 29*(4): 471–476.

Neves D, Estrozi LF, Job V, Gabel F, Schoehn G, Dessen A (2012). Conformational states of a bacterial a2-Macroglobulin resemble those of human complement C3. *PLoS ONE 7*:e35384.

Nielsen KL, Sottrup-Jensen L, Nagase H, Thøgersen HC, Etzerodt M (1994). Amino acid sequence of hen ovomacroglobulin (ovostatin) deduced from cloned cDNA. *DNA Sequencing 5*(2):111–119.

Nikapitiya C, McDowell IC, Villamil L, Muñoz P, Sohn S, Gómez-Chiarri M (2014). Identification of potential general markers of disease resistance in American oysters, Crassostrea virginica through gene expression studies. *Fish and Shellfish Immunology*41 27–36.

Nomura K, Shimizu T, Kinoh H, Sendai Y, Inomata M, Suzuki N (1997). Sea urchin hatching enzyme (envelysin): cDNA cloning and deprivation of protein substrate specificity by autolytic degradation. *Biochemistry* 36 7225–7238.

Nyberg P, Rasmussen M, Björck L (2004). α2-Macroglobulin-proteinase complexes protect Streptococcus pyogenes from killing by the antimicrobial peptide LL-37. *Journal of Biological Chemistry 279*(51):52820-52823.

O

O'Connor-McCourt MD. & Wakefield LM (1987). Latent transforming growth factor-β in serum. A specific complex with α2-macroglobulin. *Journal of Biological Chemistry 262*, 14090–14099.

Obiso RJJ, Bevan DR, Wilkins TD (1997). Molecular modeling and analysis of fragilysin, the Bacteroides fragilis toxin. *Clin. Infect. Dis. 25 (Suppl. 2) S153–S155.*

Ogay ID, Lihoradova OA, Azimova Sh S, Abdukarimov AA, Slack JM, Lynn DE (2006). Transfection of insect cell lines using polyethylenimine. *Cytotechnology, 51*(2): 89-98.

Ogata K, Nishikawa H and Ohsugi M(1969). A Yeast Capable of Utilizing Methanol. *Agricultural and Biological Chemistry*, 33:10, 1519-1520.

Ohta Y, Hojo H, Aimoto S, Kobayashi T, Zhu X, Jordan F, Inouye M (1991). Pro-peptide as an intramolecular chaperone: renaturation of denatured subtilisin E with a synthetic pro-peptide. *Molecular Microbiology*, 5(6):1507-1510.

Olmstead AD, Knecht W, Lazarov I, Dixit SB, Jean F (2012). Human subtilase SKI-1/S1P is a master regulator of the HCV Lifecycle and a potential host cell target for developing indirect-acting antiviral agents. *PLoS Pathogens, 8*(1): e1002468.

Olson ST, Björk I, Bock SC (2002). Identification of critical molecular interactions mediating heparin activation of antithrombin. *Trends in Cardiovascular Medicine12*,198–205.

Ornstein L (1964). Disc electrophoresis. I. Background and theory. *Annals of the New York Academy of Sciences*; 121:321-349.

Otlewski J, Jelen F, Zakrzewska M, Oleksy A (2005). The many faces of protease-protein inhibitor interaction. *The EMBO Journal 24*: 1303-1310.

Overall CM, López-Otín C (2002). Strategies for MMP inhibition in cancer: innovations for the post-trial era. *Nature Reviews Cancer2*: 657-672.

P

Pack DW, Hoffman AS, Pun S, Stayton PS (2005). Design and development of polymers for gene delivery. *Nature Reviews Drug Discovery, 4*(7): 581-593.

Page-McCaw A, Ewald AJ, Werb Z (2007). Matrix metalloproteinases and the regulation of tissue remodelling. *Nat. Rev. Mol. Cell Biol.* 8 221–233.

Page-McCaw A (2008). Remodeling the model organism: matrix metalloproteinase functions in invertebrates, Semin. *Cell Developmental Biology.* 19 14–23.

Paiva MM, Soeiro MN, Barbosa HS, Meirelles MN, Delain E, Araujo-Jorge TC (2010). Glycosylation patterns of human alpha-2-macroglobulin: analysis of lectin binding by electron microscopy. *Micron* 41(6):666–673.

Pak JH, Liu CY, Huangpu J, Graham JS (1997). Construction and characterization of the soybean leaf metalloproteinase cDNA. *FEBS Lett.* 404 283–288.

Pal LR, Guda C (2006). Tracing the origin of functional and conserved domains in the human proteome: implications for protein evolution at the modular level. *BMC Evol. Biol*. 6 91.

Pasquato A, Pullikotil P, Asselin MC, Vacatello M, Paolillo L, Ghezzo F, Basso F, Di Bello C, Dettin M and Seidah NG (2006). The proprotein convertase SKI-1/S1P: *in vitro* analysis of Lassa virus glycoprotein-derived substrates and *ex vivo* validation of irreversible peptide inhibitors. *Journal of Biological Chemistry, 281*: 23471-23481.

Passarelli AL (2011). Barriers to success: how baculoviruses establish efficient systemic infections. *Virology* 411 383–392.

Patino SF, Mancini RA, Pereira CA, Suazo CA, Mendonca RZ, Jorge SA (2014). Transient expression of rabies virus glycoprotein (RVGP) in *Drosophila melanogaster* Schneider 2 (S2) cells. Journal of Biotechnology*,* 192: 255-262.

Pavlovsky AG, Williams MG, Ye QZ, Ortwine DF, Purchase CF, White AD, Dhanaraj V, Roth BD, Johnson LL, Hupe D, Humblet C, Blundell TL (1999). X-ray structure of human stromelysin catalytic domain complexed with nonpeptide inhibitors: implications for inhibitor selectivity. *Protein Science* 8:1455–1462.

Pavlou AK, Reichert JM (2004). Recombinant protein therapeutics--success rates, market trends and values to 2010. *Nature Biotechnology, 22*(12): 1513-1519.

Pavlukova EB, Belozersky MA, Dunaevsky YE (1998). Extracellular proteolytic enzymes of filamentous fungi. *Biochemistry (Mosc.)* 63 899–928.

Peters TJ (1970). Intestinal peptidases. *Gut, 11*(8): 720–725.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. (2004). UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry* 25, 1605–1612.

Pisani D, Pett W, Dohrmann M, Feuda R, Rota-Stabelli O, Philippe H, Lartillot N, Worheide G (2015). Genomic data do not support comb jellies as the sister group to all other animals. *Proceedings of the National Academy of SciencesU. S. A.* 112 15402–15407.

Pomerantsev AP, Pomerantseva OM, Moayeri M, Fattah R, Tallant C, Leppla SH (2011). A Bacillus anthracis strain deleted for six proteases serves as an effective host for production of recombinant proteins. *Protein Expr. Purif.* 80 80–90.

Pop C, Salvesen GS (2009). Human caspases: activation, specificity and regulation. *Journal of Biological Chemistry 284*: 21777-21781.

Potempa J, Gomis-Rüth FX, Karim AY (2013). 185. Karilysin, N.D. Rawlings, G.S. Salvesen (Eds.), Handbook of Proteolytic Enzymes, vol. 1, Academic Press, Oxford, Great Britain, pp. 883–886.

Puck TT, Cieciura SJ, Robinson A (1958). Genetics of somatic mammalian cells: III. Long term cultivation of euploid cells from human and animal subjects. *Journal of Experimental Medicine, 108*(6): 945–956.

Puente XS, Sánchez LM, Overall CM, López-Otín C (2003). Human and mouse proteases: a comparative genomic approach. *Nat. Rev. Genet.* 4 544–558.

Pullikotil P, Benjannet S, Mayne J, Seidah NG (2007). The proprotein convertase SKI-1/S1P: alternate translation and subcellular localization. *The Journal of Biological Chemistry, 282*(37): 27402-13.

Pungercar JR, Caglic D, Sajid M, Dolinar M, Vasiljeva O, Pozgan U, Turk D, Bogyo M, Turk V, Turk B (2009). Autocatalytic processing of procathepsin B is triggered by proenzyme activity. *FEBS Journal 276*: 660-668.

Q

Qi RF, Song ZW, Chi CW (2005). Structural features and molecular evolution of Bowman-Birk protease inhibitors and their potential application. *Acta Biochimica et Biophysica Sinica 37*: 283-292.

Quesada V, Velasco G, Puente XS, Warren WC, López-Otín C (2010). Comparative genomic analysis of the zebra finch degradome provides new insights into evolution of proteases in birds and mammals. *BMC Genomics* 11 220.

Quiñones JL, Rosa R, Ruíz DL, García-Arrarás JE (2002). Extracellular matrix remodeling and metalloproteinase involvement during intestine regeneration in the sea cucumber Holothuria glaberrima. *Developmental Biology.* 250 181–197.

R

Radaev S, Zou Z, Huang T, Lafer EM, Hinck AP, Sun PD (2010). Ternary complex of transforming growth factor-β1 reveals isoform-specific ligand recognition and receptor recruitment in the superfamily. Journal of Biological Chemistry 285, 14806–14814.

Ragster LV, Chrispeels MJ (1979). Azocoll-digesting proteinases in soybean leaves: characteristics and changes during leaf maturation and senescence. *Plant Physiology. 64* 857–862.

Rajendra Y, Kiseljak D, Baldi L, Hacker DL, Wurm FM (2011). A simple high-yielding process for transient gene expression in CHO cells. *Journal of Biotechnology, 153*(1–2): 22–26.

Ramos OHP, Selistre-de-Araujo HS (2001). Identification of metalloprotease gene families in sugarcane. *Genet. Mol. Biol. 24* 285–290.

Ranasinghe S, McManus DP (2013). Structure and function of invertebrate Kunitz serine protease inhibitors. *Developmental and Comparative Immunology 39*: 219-227.

Rao MB, Tanksale AM, Ghatge MS, Deshpande VV (1998). Molecular and Biotechnological Aspects of Microbial Proteases. *Microbiology and molecular biology reviews*, 597–635.

Rasmussen M, Müller H-P, Björck L (1999). Protein GRAB of Streptococcus pyogenes regulates proteolysis at the bacterial surface by binding alpha-2-macroglobulin. *Journal of Biological Chemistry 274*(22):15336–15344.

Ratnaparkhe SM, Egertsdotter EM, Flinn BS (2009). Identification and characterization of a matrix metalloproteinase (Pta1-MMP) expressed during loblolly pine (Pinus taeda) seed development, germination completion, and early seedling establishment. *Planta 230* 339–354.

Rawlings ND (2009). A large and accurate collection of peptidase cleavages in the MEROPS database. *Database* (Oxford), bap015.

Rawlings ND (2010). Peptidase inhibitors in the MEROPS database. *Biochimie 92*: 1463-1483.

Rawlings ND, Barrett AJ (1993). Evolutionary families of peptidases. *Biochemical Journal290* (Pt 1): 205-218.

Rawlings ND, Barrett AJ (2013). Chapter 77- Introduction: Metallopeptidases and their clans. In *Handbook of Proteolytic Enzymes*, Salvesen NDR (ed): 325-370. Academic Press.

Rawlings ND, Barrett AJ, Bateman A (2010). MEROPS: the peptidase database. *Nucleic Acids Research, 38*: D227-D233.

Rawlings ND, Barrett AJ, Bateman A (2011). Asparagine peptide lyases: a seventh catalytic type of proteolytic enzymes. *Journal of Biological Chemistry, 286*: 38321-38328.

Rawlings ND, Barrett AJ, Finn R (2016). Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res. 44* D343–D350.

Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A & Finn RD (2018). The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Research 46*, D624-D632.

Rawlings ND, Tolle DP, Barrett AJ, Bateman A (2004). Evolutionary families of peptidase inhibitors. *Biochemical Journal378*: 705-716.

Rawlings ND, Waller M, Barrett AJ, Bateman A (2014). MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Research, 42*: D503-D509.

Rehman AA, Ahsan H, Khan FH (2013). α2-Macroglobulin: a physiological guardian*. Journal of. Cellular. Physiology. 228*, 1665-1675.

Reinemer P, Grams F, Huber R, Kleine T, Schnierer S, Piper M, Tschesche H, Bode W (1994). Structural implications for the role of the N terminus in the 'superactivation' of collagenases. A crystallographic study, *FEBS Letters*. 338: 227–233.

Rhee JS, Kim BM, Jeong CB, Horiguchi T, Lee YM, Kim IC, Lee JS (2012). Immune gene mining by pyrosequencing in the rockshell, Thais clavigera. *Fish and Shellfish Immunology32* 700–710*.*

Rio DC, Clark SG, Tjian R (1985). A mammalian host-vector system that regulates expression and amplification of transfected genes by temperature induction. *Science, 227*(4682): 23–28.

Robert-Genthon M, Casabona MG, Neves D, Coute Y, Ciceron F, Elsen S, Dessen A, Attree I (2013). Unique features of a Pseudomonas aeruginosa alpha-2-macroglobulin homolog. *MBio 4*(4):e00309–e00313.

Rojek JM, Lee AM, Nguyen N, Spiropoulou CF and Kunz S (2008). Site 1 protease is required for proteolytic processing of the glycoproteins of the South American hemorrhagic fever viruses Junin, Machupo, and Guanarito. *Journal of Virology, 82*: 6045–6051.

Rojek JM, Pasqual G, Sanchez AB, Nguyen NT, de la Torre JC and Kunz S (2010). Targeting the proteolytic processing of the viral glycoprotein precursor is a promising novel antiviral strategy against arenaviruses. *Journal of Virology, 84*: 573–584.

Roskov Y, Abucay L, Orrell T, Nicolson D, Bailly N, Kirk P, Bourgoin T, de Walt RE, Decock W, de Wever A, van Nieukerken E (2017) Species 2000 & ITIS Catalogue of Life., Species 2000, Leiden (The Netherlands).

Rost B (1999). Twilight zone of protein sequence alignments*. Protein Eng. 12: 85–94.*

Rowsell S, Hawtin P, Minshull CA, Jepson H, Brockbank SM, Barratt DG, Slater AM, McPheat WL, Waterson D, Henney AM, Pauptit RA (2002). Crystal structure of human MMP9 in complex with a reverse hydroxamate inhibitor. *Journal of Molecular Biology.* 319 173–181.

Rutishauser U (1996). Polysialic acid and the regulation of cell interactions. *Current Opinion in Cell Biology, 8*: 679–684.

S

Sabotič J, Kos J (2012). Microbial and fungal protease inhibitors-current and potential applications. *Applied Microbiology and Biotechnology 93*: 1351-1375.

Saiki R. et al (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 230, 1350–1354

Saito T, Bokhove M, Croci R, Zamora-Caballero S, Han L, Letarte M, de Sanctis D, Jovine L (2017). Structural basis of the human endoglin-BMP9 interaction: insights into BMP signaling and HHT1. *Cell Reports* 19, 1917–1928.

Sakai J, Rawson RB, Espenshade PJ, Cheng D, Seegmiller AC, Goldstein JL, and Brown MS (1998). Molecular identification of the sterol-regulated luminal protease that cleaves SREBPs and controls lipid composition of animal cells. *Molecular Cell, 2*: 505–514.

Salvesen GS, Sayers CA and Barrett AJ (1981). Further characterization of the covalent linking reaction of α2-macroglobulin. *Biochemical Journal 195*, 453-461.

Sallustio and Stanley (1989). Novel genetic instability associated with a developmentally regulated glycosyltransferase locus in Chinese hamster ovary cells. *Somatic Cell and Molecular Genetics*; 15:387–400.

Sambrook J, Russell DW (2001). Molecular cloning: A laboratory manual. 3rd Ed. Cold Spring Harbor Laboratory Press

Sand O, Folkersen J, Westergaard JG and Sottrup-Jensen L (1985). Characterization of human pregnancy zone protein. Comparison with human α2-macroglobulin. *Journal of Biological Chemistry 260*, 15723-15735.

Schechter I, Berger A (1967). On the size of the active site in proteases. I. Papain. *Biochem Biophys Res Commun, 27(*2): 157-62.

Schiermeyer A, Hartenstein H, Mandal MK, Otte B, Wahner V, Schillberg S (2009). A membrane-bound matrix-metalloproteinase from Nicotiana tabacum cv. BY-2 is induced by bacterial pathogens. *BMC Plant Biology, 9:* 83.

Schiestl M, Stangler T, Torella C, Cepeljnik T, Toll H, Grau R (2011). Acceptable changes in quality attributes of glycosylated biopharmaceuticals. *Nature Biotechnology, 29*: 310–312.

Schlombs K, Wagner T and Scheel J (2003). Site-1 protease is required for cartilage development in zebrafish. *Proceedings of the National Academy of Sciences USA, 100*: 14024–14029.

Schlunegger MP and Grütter MG (1992). An unusual feature revealed by the crystal structure at 2.2 Å resolution of human transforming growth factor-β2. *Nature* 358, 430–434.

Schmidt M, Finley D (2014). Regulation of proteasome activity in health and disease. *Biochimica et Biophysica Acta, 1843*(1): 13-25.

Schneider I (1972). Cell lines derived from late embryonic stages of Drosophila melanogaster. *Journal of embryology and experimental morphology, 27*(2): 353-65.

Schultze HE, Göllner I, Heide K, Schönenberger M, and Schwick G (1955). Zur Kenntnis der α-Globuline des menschlichen Normalserums. *Zeitschrift für Naturforschung, 10b*, 463-473.

Sears CL (2009). Enterotoxigenic Bacteroides fragilis: a rogue among symbiotes. *Clinical Microbiology Reviews, 22:* 349–369.

Seemuller E, Lupas A, Stock D, Lowe J, Huber R, Baumeister W (1995). Proteasome from *Thermoplasma acidophilum*: a threonine protease. *Science, 268*: 579-582.

Seidah NG (2011). The proprotein convertases, 20 years later. *Methods Molecular Biology 768*: 23–57.

Seidah NG, Mowla SJ, Hamelin J, Mamarbachi AM, Benjannet S, Touré BB, Basak A, Munzer JS, Marcinkiewicz J, Zhong M, Barale JC, Lazure C, Murphy RA, Chrétien M, Marcinkiewicz M (1999). Mammalian subtilisin/kexin isozyme SKI-1: A widely expressed proprotein convertase with a unique cleavage specificity and cellular localization. *Proceedings of the National Academy of Sciences U S A, 96*(4): 1321-1326.

Seidah ND and Prat A (2002). Precursor convertases in the secretory pathway, cytosol and extracellular milieu. *Essays in Biochemistry, 38*: 79-94.

Seidah ND and Prat A (2012). The biology and therapeutic targeting of the proprotein convertases. *Nature Reviews Drug Discovery, 11*: 367–383.

Sekiguchi R, Fujito NT, Nonaka M (2012). Evolution of the thioester-containing proteins (TEPs) of the arthropoda, revealed by molecular cloning of TEP genes from a spider, *Hasarius adansoni*. *Developmental and Comparative Immunology, 36*(2): 483-489.

Shen X, Hacker DL, Baldi L, Wurm FM (2014). Virus-free transient protein production in Sf9 cells. *Journal of Biotechnology, 171*: 61–70.

Shi M, Zhu J, Wang R, Chen X, Mi L, Walz T, Springer TA. (2011). Latent TGF-β structure and activation. *Nature* 474, 343–349.

Shimizu H, Zhang X, Zhang J, Leontovich A, Fei K, Yan L, Sarras MP Jr. (2002). Epithelial morphogenesis in hydra requires de novo expression of extracellular matrix components and matrix metalloproteinases. *Development, 129:* 1521–1532.

Shimizu T, Inoue T, Shiraishi H (2002). Cloning and characterization of novel extensinlike cDNAs that are expressed during late somatic cell phase in the green alga *Volvox carteri*. *Gene, 284:* 179–187.

Shiozaki EN, Chai J, Rigotti DJ, Riedl SJ, Li P, Srinivasula SM, Alnemri ES, Fairman R, Shi Y (2003). Mechanism of XIAP-mediated inhibition of caspase 9. *Molecular Cell, 11*: 519–527.

Shiryaev SA, Aleshin AE, Muranaka N, Kukreja M, Routenberg DA, Remacle AG, Liddington RC, Cieplak P, Kozlov IA, Strongin AY (2014). Structural and functional diversity of metalloproteinases encoded by the *Bacteroides fragilis* pathogenicity island. *FEBS Journal,* 281: 2487–2502.

Siddiqi A, Milne T, Cullinan MP, Seymour GJ (2016). Analysis of *P. gingivalis, T. forsythia* and *S. aureus* levels in edentulous mouths prior to and 6 months after placement of one-piece zirconia and titanium implants. *Clinical Oral Implants Research,* 27: 288–294.

Siezen RJ and Leunissen JA (1997). Subtilases: the superfamily of subtilisin-like serine proteases. *Protein Science*, *6*: 501–523.

Sinha-Datta U, Khan S and Wadgaonkar D (2015). Label-free interaction analysis as a tool to demonstrate biosimilarity of therapeutic monoclonal antibodies. *Biosimilars,* 15: 83–91.

Skornicka EL, Shi X and Koo PH (2002). Comparative binding of biotinylated neurotrophins to alpha(2)-macroglobulin family of proteins: relationship between cytokine-binding and neuro-modulatory activities of the macroglobulins. *Journal of Neuroscience Research*, vol. 67, no. 3, pp. 346–353.

Skottrup PD, Sorensen G, Ksiazek M, Potempa J, Riise E (2012). A phage display selected 7-mer peptide inhibitor of the *Tannerella forsythia* metalloprotease-like enzyme karilysin can be truncated to Ser-Trp-Phe-Pro. *PLoS One, 7: e48537.*

Small CD, Crawford BD (2016). Matrix metalloproteinases in neural development: a phylogenetically diverse perspective. Neural Regeneration Research, 11: 357–362.

Smart OS, Womack TO, Flensburg C, Keller P, Paciorek W, Sharff A, Vonrhein C, Bricogne G (2012). Exploiting structure similarity in refinement: automated NCS and target-structure restraints in BUSTER. *Acta Crystallographica* section D 68, 368–380.

Smith GE, Summers MD and Fraser MJ (1983). Production of human beta interferon in insect cells infected with a baculovirus expression vector. *Molecular and Cellular Biology, 3*: 2156-2165.

Söding J (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics,* 21: 951–960.

Song HK, Suh SW (1998). Kunitz-type soybean trypsin inhibitor revisited: refined structure of its complex with porcine trypsin reveals an insight into the interaction between a homologous inhibitor from *Erythrina caffra* and tissue-type plasminogen activator. *Journal of Molecular Biology,* 275(2): 347-363.

Spanier KI, Leese F, Mayer C, Colbourne JK, Gilbert D, Pfrender ME, Tollrian R (2010). Predator-induced defences in Daphnia pulex: selection and evaluation of internal reference genes for gene expression studies with real-time PCR. *BMC Molecular Biology,* 11: 50.

Sottrup-Jensen L (1987). Alpha-2-macroglobulin and related thiol ester plasma proteins. In the Plasma Proteins, (Putnam, F., ed) Vol. 5, pp. 191-292, Academic New York.

Sottrup-Jensen L (1989). Alpha-2-macroglobulins: structure, shape and mechanism of proteinase complex formation. *Journal of Biological Chemistry, 264*: 11539-11542.

Sottrup-Jensen L (1994). Role of internal thiol esters in the α-macroglobulin-proteinase binding mechanism. *Annals of the New York Academy of Sciences, 737*: 172-187.

Sottrup-Jensen L, Petersen TE, Magnusson S (1980). A thiol-ester in alpha 2-macroglobulin cleaved during proteinase complex formation. *FEBS Letters, 121*(2):275-9.

Sottrup-Jensen L, Petersen TE and Magnusson S (1981). Trypsin-induced activation of the thiol esters in α2-macroglobulin generates a short-lived intermediate ('nascent' α2-M) that can react rapidly to incorporate not only methylamine or putrescine but also proteins lacking proteinase activity. *FEBS Letters, 128*: 123-126.

Sottrup-Jensen L, Stepanik TM, Kristensen T, Lonblad PB, Jones CM, Wierzbicki DM, Magnusson S, Domdey H, Wetsel RA, Lundwall A (1985). Common evolutionary origin of α2-macroglobulin and complement components C3 and C4. *Proceedings of the National Academy of Sciences of the United States of America, 82*: 9-13.

Sottrup-Jensen L, Stepanik TM, Kristensen T, Wierzbicki DM, Jones CM, Lonbland PB, Magnusson S, Petersen TE (1984). Primary structure of human alpha-2-macroglobulin. V. The complete structure. *Journal of Biological Chemistry, 259*(13): 8318–8327.

Srivastava M, Begovic E, Chapman J, Putnam NH, Hellsten U, Kawashima T, Kuo A, Mitros T, Salamov A, Carpenter ML, Signorovitch AY, Moreno MA, Kamm K, Grimwood J, Schmutz J, Shapiro H, Grigoriev IV, Buss LW, Schierwater B, Dellaporta SL, Rokhsar DS (2008). The Trichoplax genome and the nature of placozoans. *Nature, 454:* 955–960*.*

Stanley P, Sundaram S, Tang J, Shi S. (2005). Molecular analysis of three gain-of-function CHO mutants that add the bisecting GlcNAc to N-glycans. *Glycobiology*, 15: 43–53.

Starkey PM and Barrett AJ (1973). Inhibition by α-macroglobulin and other serum proteins*. Biochemical Journal, 131*: 823-831.

Stevenson J, Krycer JR, Phan L, Brown AJ (2013). A practical comparison of ligation-independent cloning techniques. *PLOS ONE*, 8(12): e83888.

Strickland DK, Ashcom JD, Williams S, Burgess WH, Migliorini M and Argraves WS (1990). Sequence identity between the α2-macroglobulin receptor and low-density lipoprotein receptor-related protein suggests that this molecule is a multifunctional receptor. *Journal of Biological Chemistry, 265*: 17401-17404.

Stöcker W, Gomis-Rüth FX (2013). Astacins: Proteases in Development and Tissue Differentiation, in: K. Brix, W. Stöcker (Eds.), Proteases: Structure and Function, Springer Verlag, Vienna, Austria , pp. 235–263.

Stöcker W, Grams F, Baumann U, Reinemer P, Gomis-Rüth FX, McKay DB, Bode W (1995). The metzincins-topological and sequential relations between the astacins, adamalysins, serralysins, and matrixins (collagenases) define a superfamily of zinc-peptidases. *Protein Science,* 4: 823–840.

Stubbs MT, Laber B, Bode W, Huber R, Jerala R, Lenarcic B, Turk V (1990). The refined 2.4 A X-ray crystal structure of recombinant human stefin B in complex with the cysteine proteinase papain: a novel type of proteinase inhibitor interaction. *EMBO Journal, 9*(6): 1939–1947.

Sun L, Chen M, Yang H, Wang T, Liu B, Shu C, Gardiner DM (2011). Large scale gene expression profiling during intestine and body wall regeneration in the sea cucumber *Apostichopus japonicus*. *Comparative Biochemistry and Physiology Part D Genomics and Proteomics,* 6: 195–205.

Sun R, Li ZY, He HJ, Wei J, Wang J, Zhang QX, Zhao J, Zhan XM, Wu ZD (2012). Molecular cloning and characterization of a matrix metalloproteinase, from *Caenorhabditis elegans*: employed to identify homologous protein from *Angiostrongylus cantonensis*. *Parasitology Research, 110:* 2001–2012.

Sørensen HP, Sperling-Petersen HU, Mortensen KK (2003). Production of recombinant thermostable proteins expressed in *Escherichia coli*: completion of protein synthesis is the bottleneck. *Journal of Chromatography B Analytical Technologies Biomedical Life Science*, 786(1-2): 207-214.

T

Tallant C, García-Castellanos R, Baumann U, Gomis-Rüth FX (2010a). On the relevance of the Met-turn methionine in metzincins. Journal of Biological Chemistry 285: 13951–13957.

Tallant C, Marrero A, Gomis-Rüth FX (2010b). Matrix metalloproteinases: fold and function of their catalytic domain. *Biochimica et Biophysica Acta - Molecular Cell Research*, 1803: 20–28.

Tassew NG, Charish J, Seidah NG and Monnier PP (2012). SKI-1 and Furin generate multiple RGMa fragments that regulate axonal growth. *Developmental Cell, 22*: 391–402.

Terashita T, Oda K, Kono M, Murao S (1985). Purification and some properties of metal proteinases from *Lentinus edodes*. *Agricultural and Biological Chemistry,* 49: 2293–2300.

Terauchi M, Yamagishi T, Hanyuda T, Kawai H (2017). Genome-wide computational analysis of the secretome of brown algae (Phaeophyceae). *Marine Genomics, 32:* 49–59.

Tochowicz A, Goettig P, Evans R, Visse R, Shitomi Y, Palmisano R, Ito N, Richter K, Maskos K, Franke D, Svergun D, Nagase H, Bode W, Itoh Y (2011). The dimer interface of the membrane type 1 matrix metalloproteinase hemopexin domain: crystal structure and biological functions. *Journal of Biological Chemistry, 286:* 7587–7600.

Touré BB, Munzer JS, Basak A, Benjannet S, Rochemont J, Lazure C, Chrétien M and Seidah NG (2000). Biosynthesis and enzymatic characterization of human SKI-1/S1P and the processing of its inhibitory prosegment. *The Journal of Biological Chemistry, 275*: 2349–2358.

Travis J, Salvesen GS (1983). Human plasma proteinase inhibitors. *Annual Review of Biochemistry 52*: 655–709.

Tunstall AM, Merriman JML, Milne I, James K (1975). Normal and pathological serum levels of alpha-2-macroglobulins in men and mice. *Journal of Clinical Pathology, 28*: 133–139.

Turk B (2006). Targeting proteases: successes, failures and prospects. *Nature Reviews Drug Discovery 5*(9): 785-799.

Turk B, Turk V (2009). Lysosomes as "suicide bags" in cell death: myth or reality? *Journal of Biological Chemistry, 284*: 21783-2787.

Turk B, Turk D, Turk V (2000). Lysosomal cysteine proteases: more than scavengers. *Biochimica et Biophysica Acta* 1477: 98–111.

Turk B, Turk D, Salvesen GS (2002). Regulating cysteine protease activity: essential role of protease inhibitors as guardians and regulators. *Current Pharmaceutical Design, 8*: 1623–1637.

U

Uchida Y, Tsukada Y, Sugimori T (1979). Enzymatic properties of neuraminidases from *Arthrobacter ureafaciens*. *Journal of Biochemistry;* 86:1573–1585.

Ueno T, Nakaoka T, Takeuchi H, Kubo T (2009). Differential gene expression in the hypopharyngeal glands of worker honeybees (Apis mellifera L.) associated with an age-dependent role change. *Zoological Science, 26:* 557–563.

Uparanukraw P, Morakote N, Harnnoi T, Dantrakool A (2001). Molecular cloning of a gene encoding matrix metalloproteinase-like protein from *Gnathostoma spinigerum*. *Parasitology Research*, 87: 751–757.

Urban S (2009). Making the cut: central roles of intramembrane proteolysis in pathogenic microorganisms. *Nature Reviews Microbiology, 7*(6): 411-423.

Urlaub G, Käs E, Carothers AM, Chasin LA (1983). Deletion of the diploid dihydrofolate reductase locus from cultured mammalian cells. *Cell, 33*(2): 405–412.

Urquhart L (2019). Top drugs and companies by sales in 2018. *Nature Reviews Drug Discovery* doi: 10.1038/d41573-019-00049-0.

V

van den Ent F, Löwe J (2006). RF cloning: a restriction-free method for inserting target genes into plasmids. *Journal of Biochemical and Biophysical Methods*, 67(1): 67-74.

van der Hoorn RAL, Colby T, Nickel S, Richau KH, Schmidt J, Kaiser M (2011). Mining the active proteome of *Arabidopsis thaliana*. *Frontiers in Plant Science,* 2: 89.

Van den Nieuwenhof I, Koistinen H, Easton RL, Koistinen R, Kamarainen M, Morris HR, van Die I, Seppala IM, Dell A, Van den Eijnden DH (2000). Recombinant glycodelin carrying the same type of glycan structures as contraceptive glycodelin-A can be produced in human kidney 293 cells but not in Chinese hamster ovary cells. *European Journal of Biochemistry, 267*: 4753–4762.

Van Leuven F (1982). Human α2-macroglobulin: structure and function. *Trends in Biochemical Sciences* volume 7, issue 5, P185-187.

Van Oers MM (2011). Opportunities and challenges for the baculovirus expression system. *Journal of Invertebrate Pathology, 107*: S3-15.

Van Wart HE, Birkedal-Hansen H (1990). The cysteine switch: a principle of regulation of metalloproteinase activity with potential applicability to the entire matrix metalloproteinase gene family. *Proceedings of the National Academy of Sciences* U. S. A. 87: 5578–5582.

Varshavsky A (2001). Proteolysis. *Encyclopedia of Genetics, 1573–1575.*

Vaughn JL, Goodwin RH, Tompkins GJ, McCawley P (1977). The establishment of two cell lines from the insect *Spodoptera frugiperda* (Lepidoptera; Noctuidae). *In Vitro, 13*: 213-217.

Vincent MJ, Sanchez AJ, Erickson BR, Basak A, Chretien M, Seidah NG and Nichol ST (2003). Crimean-Congo hemorrhagic fever virus glycoprotein proteolytic processing by subtilase SKI-1. *Journal of Virology, 77:* 8640–8649.

Vogel CW, Finnegan PW, and Fritzinger DC (2014). Humanized cobra venom factor: Structure, activity, and therapeutic efficacy in preclinical disease models. *Molecular Immunology, 61*: 191-203.

Vishnuvardhan S, Ahsan R, Jackson K, Iwanicki R, Boe J, Haring J, Greenlee KJ (2013). Identification of a novel metalloproteinase and its role in juvenile development of the tobacco hornworm, *Manduca sexta* (Linnaeus). *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution, 320:* 105–117.

W

Wada K, Sato H, Kinoh H, Kajita M, Yamamoto H, Seiki M (1998). Cloning of three *Caenorhabditis elegans* genes potentially encoding novel matrix metalloproteinases. *Gene, 211:* 57–62.

Wakefield LM, Smith DM, Broz S, Jackson M, Levinson AD, Sporn MB (1989). Recombinant TGF-beta 1 is synthesized as a two-component latent complex that shares some structural features with the native platelet latent TGF-beta 1 complex. *Growth Factors 1*(3):203-218.

Wakefield L.M, Smith DM, Flanders KC and Sporn MB(1988). Latent transforming growth factor-8 from human platelets: a high molecular weight complex containing precursor sequences. *Journal of Biological Chemistry 263*,7646-7654.

Walsh G (2010). Biopharmaceutical benchmarks 2010. *Nature Biotechnology, 28*(9): 917-924.

Wang X, Fischer G and Hyvonen M (2016). Structure and activation of pro-activin A. *Nature Communications,* 7: 12052.

Wang J, Zuo X, Yu P, Byeon IJ, Jung J, Wang X, Dyba M, Seifert S, Schwieters CD, Qin J, Gronenborn AM, Wang YX (2009). Determination of multicomponent protein structures in solution using global orientation and shape restraints. *Journal of the American Chemical Society, 131*(30): 10507–10515.

Webb DJ, Atkins TL, Crookston KP, Burmester JK, Qian SW, Gonias SL (1994). Transforming growth factor beta isoform 2-specific high affinity binding to native alpha 2-macroglobulin. Chimeras identify a sequence that determines affinity for native but not activated alpha 2-macroglobulin. *Journal of Biological Chemistry, 269*(48):30402-30406.

Webb DJ, Wen J, Karns LR, Kurilla MG & Gonias SL (1998). Localisation of the binding site for the transforming growth factorbeta in human α2-macroglobulin to a 20 kDa peptide that also contains the bait region. *Journal of Biological Chemistry, 273*: 13339–13346.

Westwood M, Aplin JD, Collinge IA, Gill A, White A and Gibson JM (2001). Alpha 2-macroglobulin: a new component in the insulin-like growth factor/insulin-like growth factor binding protein-1 axis. *The Journal of Biological Chemistry*, vol. 276, no. 45, pp. 41668–41674.

Whiten DR, Cox D, Horrocks MH, Taylor CG, De S, Flagmeier P, Tosatto L, Kumita JR, Ecroyd H, Dobson CM, Klenerman D, Wilson MR. (2018). Single Molecule Characterization of the Interactions between Extracellular Chaperones and Toxic a-Synuclein Oligomers. *Cell Reports*, vol. 23, no. 12, pp. 3492–3500.

Wide L, Eriksson K, Sluss PM, Hall JE (2010). The common genetic variant of luteinizing hormone has a longer serum half-life than the wild type in heterozygous women. *Journal of Clinical Endocrinology and Metabolism, 95*: 383–389.

Williams SE, Kounnas MZ, Argraves KM, Argraves WS, Strickland DK (1994). The α2-Macroglobulin Receptor/Low Density Lipoprotein Receptor-Related Protein and the Receptor-Associated Protein. *Annals of the New York Academy of Sciences*, 737(1):0077-8923.

Wilson C, Bellen HJ, Gehring WJ. 1990. Position effects on eukaryotic gene expression. *Annual Review of Cell and Developmental Biology,* 6: 679-714.

Winslow GM, Hayashi S, Krasnow M, Hogness DS, Scott MP (1989). Transcriptional activation by the Antennapedia and fushi tarazu proteins in cultures *Drosophila* cells. *Cell, 57*: 1017-1030.

Wolf BB and Gonias SL (1994). Neurotrophin binding to human alpha 2-macroglobulin under apparent equilibrium conditions. *Biochemistry*, vol. 33, no. 37, pp. 11270–11277.

Wong SG, Dessen A (2014). Structure of a bacterial alpha-2-macroglobulin reveals mimicry of eukaryotic innate immunity. *Nature Communications, 5*:4917

Wright CS, Alden RA and Kraut J (1969). Structure of subtilisin BPN' at 2.5 angstrom resolution. *Nature*, *221*: 235–242.

Wurm FM (2004). Production of recombinant protein therapeutics in cultivated mammalian cells. *Nature Biotechnology, 22*(11): 1393-1398.

Wyatt AR, Constantinescu P, Ecroyd H, Dobson CM, Wilson MR, Kumita JR, Yerbury JJ (2013a). Protease activated alpha-2-macroglobulin can inhibit amyloid formation via two distinct mechanisms. *FEBS Letters*, vol. 587, no. 5, pp. 398–403.

Wyatt AR, Kumita JR, Mifsud RW, Gooden CA, Wilson MR and Dobson CM (2014). Hypochlorite-induced structural modifications enhance the chaperone activity of human 2-macroglobulin *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 20, pp. E2081–E2090.

Wyatt AR, Yerbury JJ, Dabbs RA, Wilson MR (2012). Roles of extracellular chaperones in amyloidosis. *Journal of Molecular Biology,* 421:499–516.

Wyatt AR, Zammit NW and Wilson MR (2013b). Acute phase proteins are major clients for the chaperone action of a2-macroglobulin in human plasma. *Cell Stress & Chaperones*, vol. 18, no. 2, pp. 161–170.

Y

Yadav L, Puri N, Rastogi V, Satpute P, Ahmad R, Kaur G (2014). Matrix metalloproteinases and cancer - roles in threat and therapy. Asian Pacific Journal of Cancer Prevention*, 15*(3): 1085-1091.

Yang H, Li J, Du G, Liu L (2017). Chapter 6 - Microbial Production and Molecular Engineering of Industrial Enzymes: Challenges and Strategies. *Biotechnology of Microbial Enzymes,* 151-165.

Yates JL, Warren N, Sugden B (1985). Stable replication of plasmids derived from Epstein-Barr virus in various mammalian cells. *Nature, 313*(6005): 812–815.

Ye J1, Rawson RB, Komuro R, Chen X, Davé UP, Prywes R, Brown MS, Goldstein JL. (2000). ER stress induces cleavage of membrane-bound ATF6 by the same proteases that process SREBPs. *Molecular Cell*, 6(6):1355-64.

Yerbury JJ, Kumita JR, Meehan S, Dobson CM and Wilson MR (2009). A2-macroglobulin and haptoglobin suppress amyloid formation by interacting with prefibrillar protein species. *The Journal of Biological Chemistry*, vol. 284, no. 7, pp. 4246–4254.

Z

Zakharova E, Horvath MP, Goldenberg DP (2009). Structure of a serine protease poised to resynthesize a peptide bond. *Proceedings of the National Academy of Sciences USA, 106*: 11034-11039.

Zhao J, Liu X, Malhotra A, Li Q, Zhang F, Linhardt RJ (2017). Novel method for measurement of heparin anticoagulant activity using SPR. *Analytical Biochemistry*, 526: 39-42.

Zhao B, Xu S, Dong X, Lu C and Springer TA (2018). Prodomain-growth factor swapping in the structure of pro-TGF-β1. *Journal of Biological Chemistry,* 293: 1579–1589.

Zou Z, Sun PD (2006). An improved recombinant mammalian cell expression system for human transforming growth factor-beta2 and -beta3 preparations. *Protein Expression and Purification, 50(*1): 9-17.