

I422, un cos										
	rz	ry	rz	tx	ty	tz	CCF	RF	CCI	Dist
cos1	38.5	-0.9	-18.0	0.0023	1.0011	0.1906	54.6	44.6	59.6	0.3
cos1	57.6	-0.8	-7.1	0.0012	0.0016	0.1909	53.7	44.9	59.8	0.4

C2, tres cossos										
cos1	89.7	89.7	194.8	0.1934	-0.0004	0.3487	37.4	46.1	42.7	98.6
cos2	90.4	89.8	295.1	0.0277	0.5041	0.6819	53.4	40.9	62.5	59.7
cos3	89.7	90.2	65.1	0.3592	0.5034	0.0149	69.2	35.0	75.8	59.4

Taula 3.10: Reemplaçament Molecular. Solucions del reemplaçament molecular del cristall del tercer article. (I422, plana 91), i d'un dels cristalls amb simetria C2. La posició de la molècula model està definida per tres angles d'Euler (rz,ry,rz) i tres translacions (tx,ty,tz) respecte al model situat a l'origen de coordenades i amb l'eix principal d'inèrcia perpendicular al pla a-b. CCF és el coeficient de correlació entre les Fo i les Fc del model, RF és l'Rfactor entre ells i CCI el coeficient de correlació calculat amb les intensitats. Dist és la distància (en Å) entre els centres de massa de molècules simètriques. A la taula I422 es mostren les dues millors solucions fent servir un model del connector sencer [48]. Com que cada connector ocupava 4 unitats asimètriques (3 proteïnes p10 per unitat asimètrica i rotades 30°), es van trobar dues solucions equivalents rotades 30°. La tercera solució no es va trobar, segurament a causa de que aquest model no és totalment simètric. La recerca es va fer a una resolució de 4 Å i fixant el centre al voltant de l'origen. Les distàncies entre centres de massa són nul·les a causa de que les solucions es sobreposen. A la taula C2 es mostra una única solució amb tres connectors per unitat asimètrica (una en cada línia). En aquesta cel·la hi havia espai per a tres connectors per unitat asimètrica. La recerca es va fer a 4.5 Å amb el model I422 afinat. A part d'aquesta solució es van trobar moltes solucions equivalents, cadascuna corresponent a diferents rotacions de múltiples de 30° del connector. En aquesta segona taula es pot veure com la correlació i l'Rfactor milloren, augmentant i disminuint, respectivament en afegir més cossos.

El resultat de les rotacions i translacions és un conjunt de solucions possibles. Moltes es poden descartar per un mal empaquetament, però la funció de correlació és molt sensible a petites diferències en la posició, i cal fer un afinament final de les 6 dimensions, i de cada molècula, al voltant de les millors posicions i orientacions (afinament del sòlid rígid).

La correlació està calculada com:

$$C = \frac{\sum_h (|Fo_h| - \langle |Fo| \rangle) (|Fc_h| - \langle |Fc| \rangle)}{[\sum_h (|Fo_h| - \langle |Fo| \rangle)^2 \sum_h (|Fc_h| - \langle |Fc| \rangle)^2]^{\frac{1}{2}}}$$

on F_o representa el factor d'estructura observat, i F_c , el calculat pel Reemplaçament Molecular.

Un altre valor, l'*Rfactor* es farà servir també en l'afinament posicional. Representa la desviació dels valors d'amplitud calculats per la proteïna que s'està modelant respecte als observats:

$$R = \frac{\sum_h ||Fo_h| - |Fc_h||}{\sum_h |Fo_h|}$$

Per tant, és un valor que ha de disminuir amb la qualitat del model. Els valors típics finals d'*Rfactor* en estructures proteïques afinades oscil·len entre el 20% i el 30%.

A la taula 3.10 es mostren les solucions de reemplaçament molecular de dos cristalls diferents del connector. El primer (I422) correspon al presentat en aquesta tesi (tercer article, plana 91). El segon (C2), a un dels cristalls obtinguts al principi (primer article, plana 77), que s'ha pogut resoldre *a posteriori* a partir de la solució afinada. Aquest segon model no s'ha afinat, ja que la qualitat i resolució de les dades no aportaria detalls nous al model actual (taula 3.4, plana 34).

3.5.3 NCS

Quan una proteïna està composta de diverses parts idèntiques, és molt freqüent que aquestes estiguin relacionades per una operació de simetria local dins de la unitat asimètrica. Això passa quan l'operació de simetria no pertany a un grup espacial, o bé quan l'empaquetament està relaxat, i li falta poc per constituir grup espacial (veure la figura 3.1 de la plana 22).

En aquest cas es diu que hi ha *Simetria No Cristal·logràfica* (NCS). La NCS és d'una gran ajuda en totes les etapes de la resolució d'una estructura:

- A la difracció provoca relacions entre reflexions més enllà de les del grup espacial. Això permet trobar les direccions dels eixos de simetria de la molècula amb la funció d'autorotació. La funció d'autorotació és la correlació del Patterson de les reflexions amb el mateix Patterson, amb una rotació diferent. A les planes 86 i 87 es poden veure dos exemples de mapes d'aquest tipus.

- Per solucionar una estructura, parts que no s'han resolt bé poden ser completades gràcies a la informació de parts simètriques. Aquesta tècnica forma part de l'extensió de fases, que es comentarà més endavant (secció 3.5.5, plana 51).
- En l'afinament de la estructura, la NCS relaciona cadenes peptídiques, disminuint el nombre de paràmetres lliures per afinar.

Hi ha dos tipus d'NCS, l'estricta o rígida, definit per les matrius que relacionen les parts repetides, i el restringit, on només s'imposa que la “forma” de la cadena d'aminoàcids ha de ser el més igual possible. En aquest cas, enlloc de fer servir, per exemple, les coordenades (i temperatures, torsions, etc) per a cada aminoàcid, es fan servir les d'un dels relacionats, més les diferències amb els altres. Aquestes diferències es deixen créixer segons un pes, que pel cas de simetria NCS estricta, seria infinit i per no simetria seria de zero.

3.5.4 Màscares

Una màscara és equivalent a un mapa de densitat, però composta només d'uns i zeros. Multiplicant els punts equivalents dels mapes de densitat i de les màscares s'obtenen mapes on només hi ha les zones on la màscara val 1.

Una màscara és necessària quan cal aplicar certes operacions a una part d'una estructura. Per modular l'efecte del solvent en les amplituds (plana 60), es pot fer servir una màscara que separi la part que és “cristal·logràfica” de la que està composta només de solvent desordenat. Es tracta d'una operació iterativa que forma part del cicle de l'afinament, ja que aquesta màscara variarà lleugerament a cada cicle. La idea consisteix a calcular quina seria l'aportació de la part del solvent en la difracció, calculant uns factors d'estructura (Fbs) a partir d'un mapa on només hi ha un valor constant de la densitat ($0.36 e^-/\text{\AA}^3$). Aquesta aportació es calcula només per a les reflexions de baixa resolució, fins uns 6 o 8 Å, ja que és fins aquest límit quan es pot observar l'efecte del solvent desordenat. Aquest límit és el que fa que el *Wilson Plot* (fig. 3.5, plana 35) perdi el comportament rectilini.

La màscara es construeix perfilant la proteïna, a una distància d'uns 3 Å, i a la màxima resolució. Per això cal que sigui iteratiu, ja que al llarg de l'afinament és la part d'alta resolució la que variarà més. A l'hora de calcular els *Rfactors*, es sumen les Fbs a les Fc per aproximar millor el valor a Fo .

En la extensió de fases amb simetria no cristal·logràfica (NCS, plana 51) és necessari promitjar les densitats en zones relacionades en una unitat asimètrica. Aplicant la màscara es pot aïllar cadascuna d'aquestes zones, normalment un monòmer, i llavors, aplicar les transformacions necessàries

per sobreposar-les i promitjar-les. Una altra aplicació en NCS és la de generar la unitat asimètrica a partir d'una zona que es consideri millor definida. Es tracta de fer el mateix, però enlloc de promitjar, es copien els valors d'un dels mapes emmascarats.

Per construir les màscares se sol partir d'un model d'àtoms, però en el nostre cas, es partia d'un model de microscòpia electrònica. Les màscares inicials es van construir d'una manera similar al model de microscòpia (plana 69), però tallant un sector cilíndric de 30° , seguint la forma d'un monòmer. Més endavant, en estendre les fases, la resolució del model de microscòpia no era suficient. Es va fer servir una màscara editada manualment a partir dels mapes a aquella resolució o bé una màscara construïda a partir d'un model de pseudoàtoms que cobria una unitat asimètrica. Aquest mètode està explicat a la secció de materials i mètodes del tercer article (plana 101).

3.5.5 L'extensió de fases

L'extensió de fases és un procediment iteratiu que intenta obtenir fases a una resolució més alta de la que es disposa amb només un model o amb un traçat fet a partir de fases experimentals a menor resolució.

El típic cas és amb cristalls de derivats. El fet d'introduir àtoms pesants en un cristall, moltes vegades provoca un empitjorament de la qualitat del cristall, i, per tant de la difracció.

Un altre cas, el fet servir per aquest treball, és el d'estendre fases a partir d'un model de molt baixa resolució. Aquest model pot ser una estructura de microscòpia electrònica, o, fins i tot un model geomètric, com un icosaèdre en el cas de càpsides de virus, si es parteix de molt baixa resolució (menys que 10 o 20 Å).

La tècnica consisteix en construir un model que contingui la informació esperada a baixa resolució. Per exemple, un model de microscòpia tindrà el perfil de la proteïna, i una certa distribució de densitat a l'interior corresponent a zones més denses de la molècula. Amb aquest model es construeix una cel·la igual que l'experimental, aprofitant totes les dades que es puguin obtenir a partir de la síntesi de Patterson[6], l'empaquetament, la funció d'autorotació, o de micrografies electròniques de la mateixa cel·la. S'ha d'intentar ajustar aquest model perquè la seva correlació amb les reflexions de baixa resolució sigui la millor possible, el que és un problema no gaire ben resolt, ja que la informació del solvent desordenat s'obté, com hem vist abans, a partir de l'alta resolució (plana 50).

Un cop s'ha obtingut un bon ajust del model i de les reflexions, s'intenten extrapolar les fases a una resolució lleugerament més alta, per transformació de Fourier del model. Si pot ser, que inclogui només una reflexió nova, o

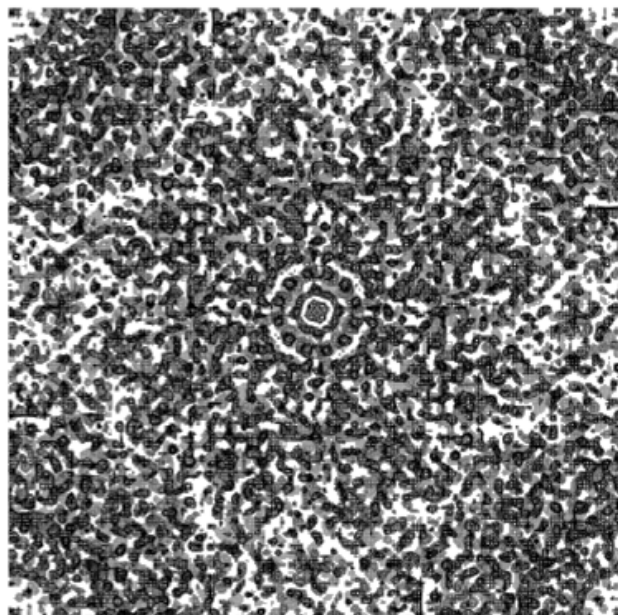


Figura 3.8: Extensió de fases I. Projectió al llarg de l'eix c del mapa resultant de l'extensió de fases explicada a la plana 91 [42].

almenys només una capa de resolució ben petita. L'extrapolació es fa per un procés iteratiu, on, en cada increment de resolució, es fa una modificació de la densitat del model per que tingui una distribució adequada a aquella resolució, comparant amb distribucions d'altres proteïnes (*histogram matching*).

Una altra modificació molt efectiva és la de l'aplanament de solvent. La densitat corresponent a la zona de solvent desordenat es posa a un valor constant per augmentar el contrast entre proteïna i solvent.

La més efectiva de les modificacions a fer en cada increment de resolució és la del promitjat de la simetria no cristal·logràfica. Per això cal, és clar, que la molècula tingui aquesta simetria, però en grans molècules sol ser el cas. En casos com aquests, la baixa resolució inicial permet usar màscares tal com abans s'ha explicat, ja que les diferències entre els monòmers no se solen apreciar fins mitja o alta resolució.

La densitat es pot modificar també per una millora de les màscares en cada iteració. També per una esqueletonització, que consisteix en connectar parts que tenen una densitat més alta i intentar seguir el traç de la proteïna. I també per mètodes de màxima versemblança a l'espai real, semblants als fets servir per l'afinament de l'estructura (plana 60), però substituint el paper de les reflexions "Rfree" per petites esferes de densitat [2].

Un cop s'han aplicat totes les modificacions de densitat adequades, es calculen les F_c del model i es comparen amb les experimentals. Si la tendència

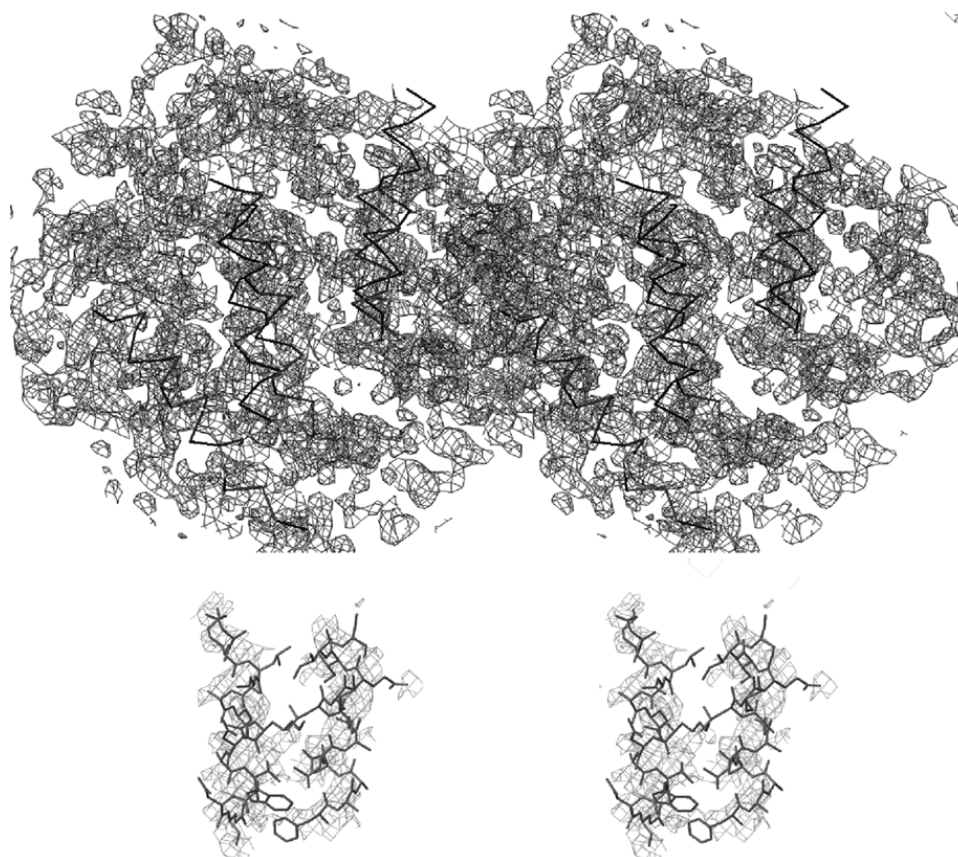


Figura 3.9: Extensió de fases II. Detall del mapa resultant de l'extensió de fases, corresponent a les hèlixs del tronc central del connector. A la imatge superior, mapa estèreo d'aquesta zona amb la cadena principal de les hèlixs de l'estructura final traçada amb barres. A la inferior, detall d'una d'aquestes hèlixs amb cadenes laterals.

no és bona, es poden afegir més ajustos, si és bona, es calculen les amplituds i les fases a una resolució una mica més alta, i es tornen a comparar amb les experimentals. En el cas que siguin bones, es prendran aquestes fases noves per a un proper cicle.

La mesura de la bondat de les reflexions noves es fa a partir de la versemblança tal com s'explica a l'afinament (secció 3.6.2, plana 60). En aquest cas, però, en lloc de predir les reflexions "*R_{free}*", es tracten de predir les amplituds de les reflexions a la propera capa de resolució. La precisió d'aquesta predicció queda reflectida en la *figura de mèrit* (FOM), que és una constant que multiplicarà a cada reflexió a l'hora de construir els mapes. Una FOM

baixa, fa que la reflexió tingui poc pes en la síntesi de Fourier, i a més informa al programa que és una reflexió (i una fase) que cal millorar en el proper cicle.

L'aplicació pràctica d'aquesta extensió de fases està explicada a materials i mètodes del tercer article (plana 91).

3.5.6 El traçat de l'estructura

Es considera que s'ha trobat la *solució* d'una estructura quan s'obté un mapa interpretable en termes de cadena d'aminoàcids (o dels àtoms que la formen). A partir d'aquest moment comença l'etapa de traçat i d'afinament.

El traçat consisteix a inspeccionar el mapa i anar situant la cadena peptídica al llarg d'ell. Depenent de la qualitat de les dades s'ha de fer gràficament o es pot fer automàticament si la resolució és molt alta. Generalment es comença introduint elements d'estructura secundària en els mapes inicials. Quan es té un traçat inicial, encara que només sigui aproximat o de només certes parts de les cadenes, es calculen mapes de diferències $2Fo - Fc, \phi c$ i $Fo - Fc, \phi c$. Aquests mapes s'obtenen fent la síntesi de Fourier amb les diferències entre les amplituds observades (Fo) i les calculades amb el material traçat (Fc), i com a fases (ϕ) es fan servir les disponibles, és a dir, les provinents del càlcul de Fc o bé una combinació amb altres fonts que també donin informació de fases, com la provinent d'àtoms pesants, o de l'extensió de fases.

Els mapes $2Fo - Fc, \phi c$, tenen un pes doble en la amplitud observada respecte de la calculada. El motiu és aconseguir mapes que estaran més desviats cap a l'estructura correcta, que no pas síntesis del tipus $Fo, \phi c$, ja que el pes negatiu d' Fc ajudarà a eliminar el biaix dels mapes. En una estructura afinada Fo s'aproxima molt a Fc i el resultant és un mapa que en la pràctica representa un mapa $Fo, \phi c$. És a dir, un mapa on les amplituds són les experimentals i les fases són les del model afinat.

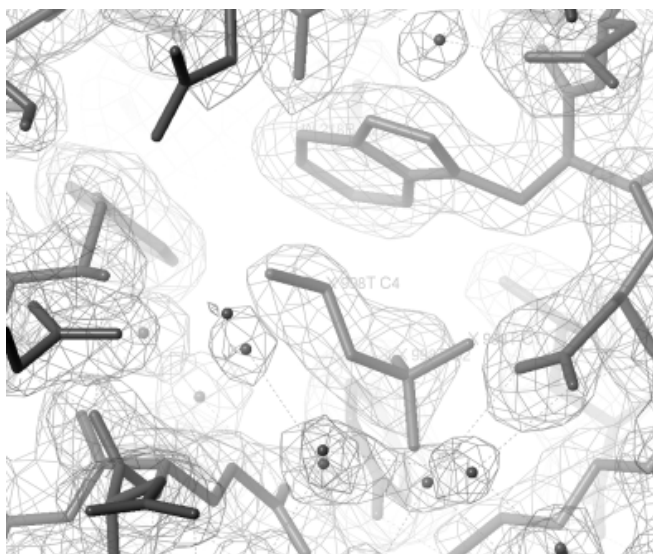
Els mapes $Fo - Fc, \phi c$ ajuden a la interpretació de parts conflictives dels mapes. Generalment són més intensos en les zones on hi ha errors en l'estructura, o aquesta no ha estat traçada encara. Les zones a on falten àtoms tindran una densitat positiva ($Fo > Fc$), i les zones on n'hi ha de mal traçats tindran densitat negativa ($Fc > Fo$). En l'estructura final, aquests mapes han de desaparèixer ja que s'igualen els dos components. De totes maneres, com que s'està treballant amb quoficients propers a zero (un cop feta la resta), els mapes $Fo - Fc, \phi c$ són molt més sorollosos que els $2Fo - Fc, \phi c$, per la qual cosa és necessari inspeccionar-los a un nivell o σ més elevat. Generalment es treballa amb densitats d' 1σ pel cim del valor mitjà dels mapes $2Fo - Fc, \phi c$ i de ± 2 o 2.5σ en els mapes $Fo - Fc, \phi c$.

També són possibles altres mapes. Els programes automàtics, per exemple, treballen amb mapes $3Fo - 2Fc$, ϕc , la *suma* dels dos mapes anteriors, encara que per al treball manual és més pràctic separar els mapes.

Tot aquest treball és iteratiu. Entre cada cycle de traçat, manual o automàtic, s'intercalen cycles d'afinament fins que totes les cadenes de la proteïna resultin traçades el millor possible.

Figura 3.10: Molècula del criotampó.

De vegades el criotampó es difon dins del cristall. En el centre, una de les 4 molècules d'MPD (2-methyl-2,4-pentandiol) que es va localitzar per unitat asimètrica (16 per connector). Els carbonis de la part superior de la molècula d'MPD estan encarats cap a un triptòfan, mentre que l'oxigen inferior està lligat per ponts d'hidrogen a molècules d'aigua. El mapa correspon a l'estructura afinada del cristall I422 del tercer article (plana 91).



3.6 Estructura final de la proteïna

3.6.1 L'afinament de l'estructura

Quan una estructura està traçada en la seva major part, els mapes resultants ja haurien de ser de prou qualitat com per poder afinar les posicions dels àtoms individualment. Tant a l'afinament com en el traçat, s'han de combinar cicles automatitzats amb cicles manuals, fent servir un programa gràfic que permeti modelar manualment l'estructura en les zones on un programa automàtic no sigui capaç d'obtenir bons resultats.

L'afinament automàtic intenta alhora situar els àtoms del model en el lloc correcte i mantenir l'estereoquímica de la proteïna. Per això construeix una funció d'energia formada per termes cristal·logràfics (S_x) i estructurals (S_g):

$$S_x = \sum_i w_i (Fo_i - Fc_i)^2$$

$$S_g = \sum_i w_i (d_i - d_c)^2$$

L'afinament consistirà a minimitzar aquesta funció, generalment pel mètode del gradient conjugat. Tots aquestes termes tenen forma d'oscil·lador, de manera que l'energia augmenta si els valors s'allunyen dels valors promig que s'han descrit per als aminoàcids. Els termes cristal·logràfics representen la desviació dels F_c respecte els F_o d'una manera similar als mapes. Els estructurals representen distàncies d'enllaç, angles de torsió, angles dièdrics, temperatura o agitació dels àtoms, fases experimentals, etc. Depenent de la qualitat i la resolució de la difracció i del tipus d'afinament es dona més pes (w_i) a uns termes o a altres. Els termes estructurals s'afinen contra valors (d_i) calculats o tabulats.

Cada estructura és diferent i no hi ha una recepta exacta del que cal fer per afinar-la. Se sol començar amb un rescalfament simulat per eliminar de l'estructura les rigideses imposades en el traçat inicial o en la solució de reemplaç molecular.

El rescalfament simulat consisteix a simular els efectes d'escalfar la proteïna al voltant d'uns 5000K, per deixar-la refredar a poc a poc després. La intenció és que els enllaços i torsions de la proteïna guanyaran mobilitat amb el calor, i podran ajustar-se més fàcilment als mapes, saltant mínims locals de potencial que no es podrien saltar a baixa temperatura.

A continuació s'intercalen cicles automàtics i manuals d'afinament individual de les posicions dels àtoms.

Altres paràmetres que s'afinen són la *temperatura* i l'ocupació de cada àtom. La *temperatura* és un paràmetre relacionat amb la fluctuació al voltant de la posició mitjana de cada àtom. Marca el grau de precisió amb què s'ha vist l'àtom en aquella coordenada del cristall, ja que d'una cel·la a una altra aquesta posició pot estar lleugerament modificada. Els valors més alts de la temperatura es donen en zones exposades al solvent o en zones *actives* de la proteïna, que presenten diferents conformacions per poder exercir la seva funció. L'ocupació és un paràmetre que representa la proporció en què aquell àtom es troba a totes les cel·les d'un cristall. A les resolucions típiques de la cristal·lografia de proteïnes l'ocupació no sol ser afinable, ja que es podrien emascarar errors de l'afinament general a base de modificar les ocupacions. De vegades, en llocs mòbils de l'estructura que presenten unes poques conformacions estables, es possible traçar una cadena diferent per a cada cas, i llavors assignar a cadascuna una ocupació proporcional.

L'ocupació és més útil quan hi ha ions pesants dintre d'un cristall. Com que generalment no s'hauran difos igual per tot el cristall, llavors afinar l'ocupació té sentit, ja que les seves coordenades espacials es poden fixar.

L'objectiu de l'afinament és aconseguir que el model construït predigui el

millor possible les reflexions observades, alhora que conservi una geometria correcta en termes estructurals. S'ha observat que les conformacions dels aminoàcids, dels enllaços entre ells i les interaccions entre cadenes properes estan restringits a una sèrie de valors que estan tabulats. Una molècula ha de complir aquest requisits o ha de tenir algun motiu especial per no complir-los. Per exemple es poden trobar desviacions en centres actius o per la influència d'algun metall.

En un enllaç peptídic, es poden trobar restriccions mitjançant *ponts d'hidrogen* en una determinada estructura se

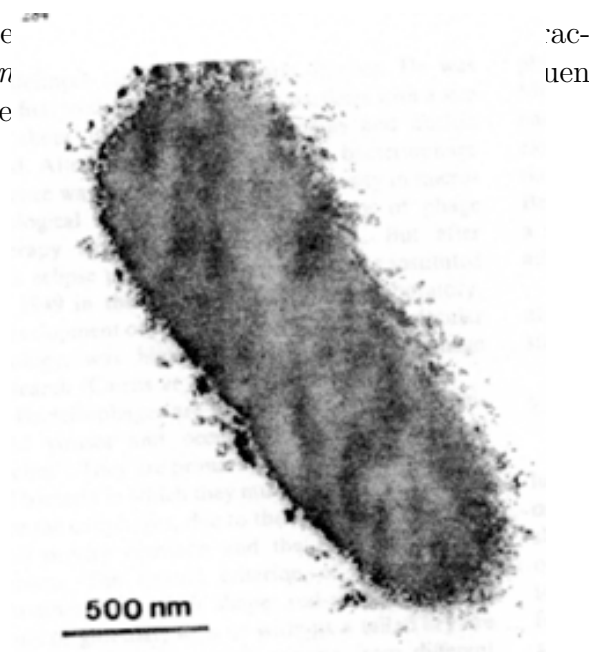


Figura 3.11: Angles peptídics. Els angles Ψ i Φ de la cadena principal estan limitats per l'estereoquímica de l'enllaç peptídic.

Totes aquestes restriccions geomètriques s'han de complir en una proteïna i es poden anar comprovant al llarg de l'afinament. Són molt útils els diagrames de Ramachandran (fig. 3.13).

Per mesurar la qualitat de la predicció de la difracció amb el model, es fa servir un concepte molt similar a l'escalat, l'*Rfactor*. Es tracta simplement de la desviació entre les amplituds observades i de les calculades, amb un pes per compensar la variació de la intensitat de la difracció a diferent resolució:

$$R = \frac{\sum_h |Fo_h - kFc_h|}{\sum_h Fo_h}$$

on k és el factor d'escala entre els valors mitjans d' Fo i d' Fc en cada capa de resolució.

Encara que és difícil per les restriccions geomètriques, és possible obtenir una estructura almenys parcialment incorrecta, deguda a sobreestimacions en l'afinament. Per evitar-ho, es defineix un altre residual similar a l'*Rfactor*,

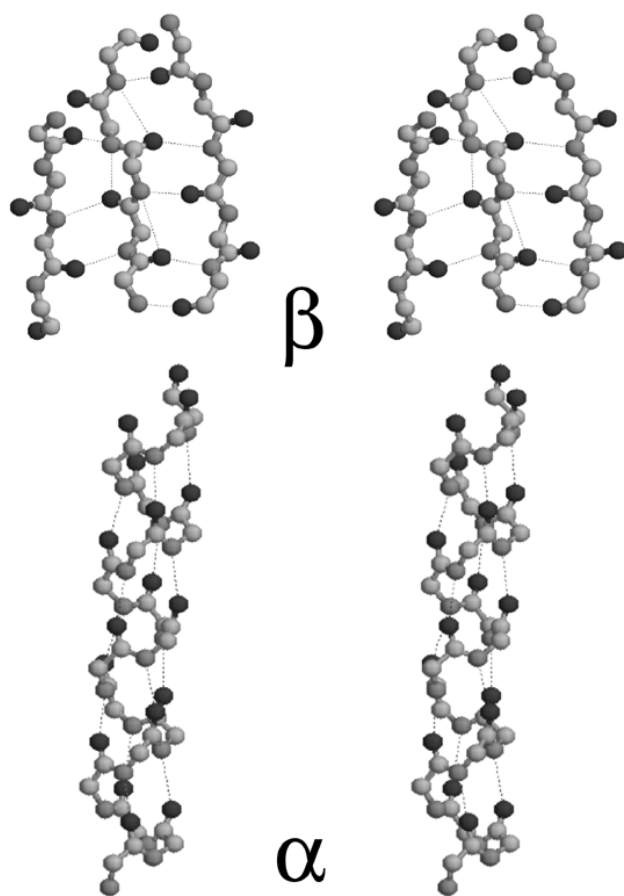


Figura 3.12: Hèlix α i fulla β . Cadena principal de la fulla β d'una part del domini SH3 del connector (plana 91) i de la hèlix $\alpha 3$. Els ponts d'hidrogen (línies de punts) segueixen un ordre característic en cada cas. En les fulles β aquests s'estableixen entre els carbonis α i els oxígens d'aminoàcids paral·lels, formant una làmina. En les hèlixs α , entre els mateixos àtoms, però d'aminoàcids separats per altres dos en la seqüència. Els angles formats s'especifiquen en la figura 3.13.

l'*Rfree*. Es construeix amb un grup de reflexions que des d'un principi se separa de les que es fan servir per resoldre, traçar i afinar l'estructura. Generalment entre un 5 i un 10% de les reflexions [5, 35]. Es pot monitoritzar la qualitat de l'afinament calculant aquest paràmetre, ja que en no servir pels càlculs, la seva correcta predicció indicarà un bon funcionament de l'afinament.

En una estructura l'aigua hi és present de dues maneres: localitzada, al voltant de la proteïna, i formant part d'ella, ja que l'estructura tridimensional i el funcionament correcte d'una proteïna depenen també de la presència d'aigua (excepte en proteïnes liposolubles, com les de la membrana cel·lular) i també deslocalitzada, formant part del gruix de solvent que forma el cristall, que normalment és de l'ordre d'un 50% de tota la matèria que conté.

L'aigua localitzada (*aigua cristal·logràfica*) es construeix i afina amb la resta de l'estructura. A les resolucions típiques, no són més que una sèrie

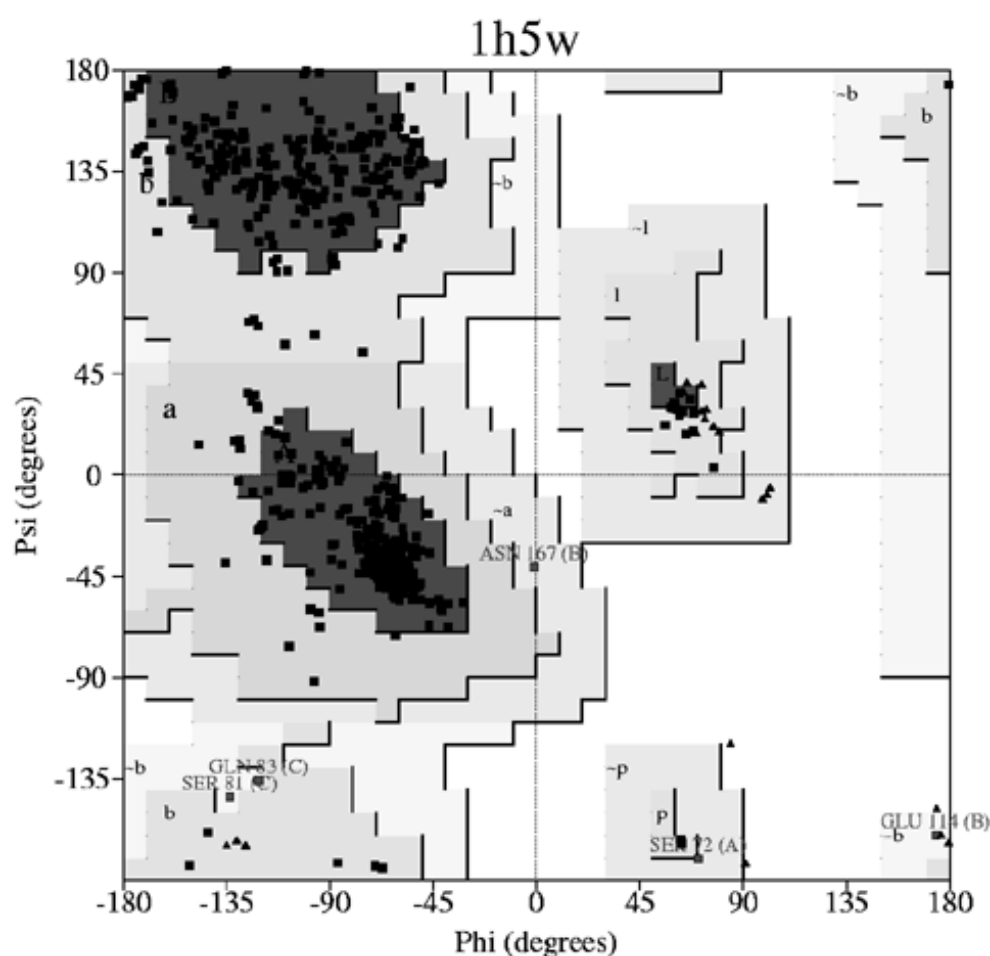


Figura 3.13: Ramachandran del connector. Diagrama de Ramachandran: Les zones fosques representen els valors més típics dels angles peptídics. La zona superior representa conformacions en fulla β . Les centrals representen hèlixs α dextrògires (esquerra) i levògires (dreta). Les zones grises representen valors menys habituals d'aquests angles. Aquest valors s'han calculat a partir d'estructures conegudes de proteïnes. Els quadres negres corresponen als angles en l'estructura afinada del connector, i els triangles a glicines, que en no tenir cadena lateral, poden abraçar més angles.

d'esferes, ja que els hidrògens no s'arriben a veure. Normalment s'espera a les últimes etapes de l'afinament per afegir-les, ja que podrien ser confoses amb altres àtoms, o bé sobreafinar l'estructura. És força fàcil aconseguir bons valors de l'*Rfactor* (però no de l'*Rfree*) posant aigua en tots els llocs dels mapes on hi aparegui densitat.

L'aigua deslocalitzada (solvent desordenat) no es pot afinar, ja que no representa àrees repetides al llarg del cristall. Només es repeteix el seu perfil a baixa resolució. Les reflexions de baixa resolució contenen aquesta informació i s'han de modular per tenir-la en compte. Una manera consisteix en construir una màscara al voltant de la proteïna, i calcular la contribució del que queda fora d'aquesta màscara en les reflexions (plana 50).

Les estadístiques de l'afinament estan resumides a la taula 1 del tercer article (plana 101).

3.6.2 Màxim de versemblança i Mínims quadrats

L'objectiu de l'afinament és obtenir una sèrie de coordenades que representin la situació i els seus errors dels àtoms de l'estructura en la cel·la unitat del cristall, a partir d'un model ja prèviament traçat. Aquestes coordenades han de complir una sèrie de restriccions geomètriques donats per les conformacions dels aminoàcids i les relacions que defineixen la seva estructura secundària, terciària i quaternària (hèlixs α , fulles β , girs, ponts d'hidrògen, altres interaccions amb molècules veïnes, etc). Alhora han de representar l'estructura que ha donat lloc a la difracció, per la qual cosa han de complir altres restriccions amb les amplituds obtingudes en la difracció.

La quantitat de variables és molt gran i les restriccions geomètriques no permeten (avui dia) generar una estructura mitjana *de novo* a partir de només càlculs. La posició avantatjosa de la cristal·lografia recau en el fet que que es poden fer servir resultats experimentals per resoldre aquest problema.

De totes maneres, la quantitat de variables (3 coordenades + 1 *b-factor* per a cada àtom) obliga a obtenir una quantitat mínima de reflexions igual. Per exemple, per resoldre l'estructura d'aquest treball, cal conèixer la posició de 12 monòmers amb 290 aminoàcids aproximadament cadascun, i unes 3500 molècules d'aigua, considerant una de promig per a cada aminoàcid. Això obliga a treballar amb unes 120000 reflexions i per tant amb una equació com:

$$Fo_{\vec{h}} = \sum_j f_j e^{2\pi i \vec{h} \cdot \vec{x}_j} e^{2\pi i \vec{h} \cdot \vec{b}}$$

per a cadascuna de les reflexions $Fo_{\vec{h}}$. La simetria del cristall i la NCS (plana 3.5.3) permeten reduir la quantitat d'equacions necessàries.

El sistema conté les incògnites en exponencials i per tant no es pot resoldre directament. En realitat, la situació és pitjor, ja que cal resoldre els mòduls d'aquesta expressió.

L'única manera de resoldre'l és disposar d'un model de partida prou bo (el traçat) i minimitzar l'equació de l'energia (plana 55) per mantenir les

restriccions. Aquesta equació es minimitza pel mètode del gradient conjugat.

L'aplicació d'aquest mètode es fa generalment minimitzant l'equació per *mínims quadrats*, però la minimització pot conduir molt fàcilment a un mínim local que no permeti millorar més enllà l'estructura. Per evitar això, s'ha d'avançar a poc a poc, començant en la resolució en què s'ha resolt l'estructura i incrementant-la només quan està correcte en aquella resolució.

Una manera molt efectiva d'evitar el perill dels mínims locals és fer servir tota la resolució de cop. D'aquesta manera es treballa amb el conjunt sencer de les dades, i estadísticament és més difícil caure en un mínim local. Cal, però un mètode per discernir què hi ha de correcte en el model, ja que ara no hi ha el criteri d'avançar cap a l'alta resolució.

Un model correcte suposa unes fases correctes, i per tant es pot tractar de jutjar *com* de correctes són unes fases. Com que no es té cap evidència experimental d'aquestes fases, cal jutjar-les a partir de l'únic que es té experimentalment: les amplituds [44]. Un model correcte predirà unes amplituds correctes, però l'afinament tendeix a situar els àtoms de manera que es prediguin amplituds correctes. Per evitar aquest cercle tancat es fan servir les reflexions deixades de banda (les de l'*R_{free}*). Si aquestes es prediuen bé, se suposarà que les fases també estan ben predites, en la seva corresponent capa de resolució, i s'assignarà un pes (figura de mèrit, FOM) alt a les reflexions corresponents. En cas contrari, el pes serà baix. En treballar amb les reflexions *lliures*, s'elimina en gran part el problema del biaix [2].

La funció a minimitzar és (funció de màxima versemblança):

$$\ln L \sim \sum_i \ln[p(x_i)]$$

on $p(x_i)$ és la probabilitat (anomenada versemblança) de que el paràmetre x_i del model s'ajusti a les observacions [44].

Quan tots els errors tenen forma gaussiana, el màxim de versemblança és equivalent al mètode dels mínims quadrats:

$$p(x) \sim \frac{1}{\sqrt{2\pi}\sigma(x)} e^{-\frac{(x-\langle x \rangle)^2}{2\sigma^2(x)}}$$

$$-\ln L \sim \sum_i \ln(\sigma_i) + \frac{e_i^2}{\sigma_i^2}$$

on e_i i σ_i són els errors i les seves desviacions associats a cadascun dels paràmetres que es minimitzen en la funció d'energia.

3.7 Macles

Un cristall no sempre creix ben ordenat. Apart de petites diferències a causa de la mosaïcitat (plana 33), el cristall pot contenir dominis macroscòpics orientats de maneres diferents. Cada domini és, en la pràctica, un cristall diferent, i difracta independentment dels altres. En un lupa de laboratori es poden veure els dominis amb un polaritzador. Si són prou grans es pot veure directament per la forma del cristall: branques, estrelles, barres. . .

De vegades és possible separar els cristalls abans de muntar-los, trencant-ne una part. Pot ser amb una agulla molt fina prement al punt d'unió, pot ser canviant-lo a una altra solució a diferent temperatura (plana 33). Les interfícies entre els cristalls són menys fortes que entre cel·les d'un mateix cristall i qualsevol acció sobre elles pot ajudar a separar-les.

Quan no és possible separar els components d'una macla, però es poden diferenciar visualment les parts, es pot ajustar el raig posant les obertures de la sortida dels raigs X a una mida molt petita i només difractar una part del cristall (fig. 3.14).

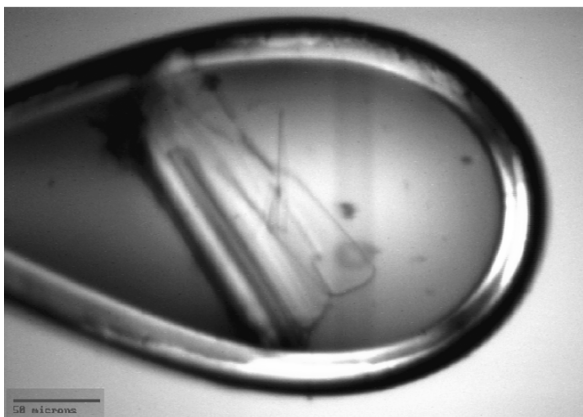
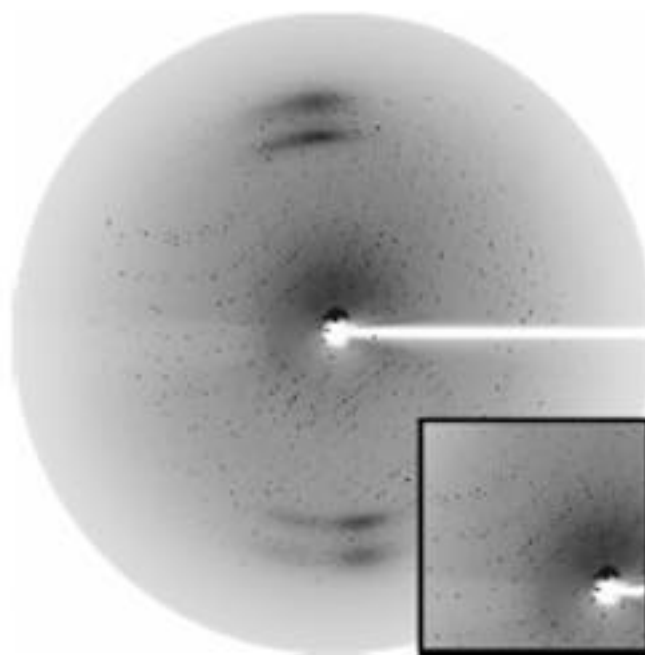


Figura 3.14: Difracció amb microfocus. Si un cristall maclat no es pot separar en parts, és convenient utilitzar un feix de secció molt petita i enfocar-lo només a una zona del cristall. Aquesta imatge correspon a la proteïna PhoB [4]. El feix està centrat en una làmina lateral del cristall. Es pot veure el punt d'incidència del feix en el cristall (la taca fosca) i el senyal que ha fet el feix de raigs X en el criotampó (la banda vertical). La secció del feix és d'uns 20 μm .

Un problema que generalment passa en aquest casos es que en anar girant el cristall durant la recollida, és molt possible que s'arribi a una orientació en la qual es difracten altres parts de la macla. El resultat són patrons de difracció que contenen diferents conjunts de reflexions (fig. 3.15). Si les orientacions dels dominis són prou diferents, aquestes imatges es poden processar tenint cura de no associar les reflexions al cristall equivocat. Algunes de les reflexions se sobreposaran i es molt difícil en aquests casos determinar a quina part del cristall pertanyen o bé quina proporció de cada cristall conté cada reflexió.

Figura 3.15: Difracció d'una macla. Difracció d'un cristall maclat. La difracció correspon a uns cristalls del connector amb TaBr difractats a la línia BM30 de l'ESRF. Les reflexions es van separar processant-les per separat. A la part inferior, detall de la imatge, o es veuen els patrons de punts en direccions casi perpendiculars.



Molt sovint els cristalls que formen la macla tenen orientacions casi similars. Aquest és el pitjor dels casos, ja que les reflexions estan gairebé allà mateix i no es poden separar. La forma de la *taca* és allargada, amb varis màxims al llarg d'ella. No solen ser útils per a resoldre una estructura. Com a molt per determinar els paràmetres de la cel·la i tractar d'hipotetitzar què es pot fer per millorar aquests cristalls.

3.7.1 Macles hemièdriques

Determinats grups espacials, per exemple els tetragonals o els hexagonals contenen una simetria que permet que les cel·les creixin en diferents orientacions. Això és a causa de, per exemple, que certs cantons de la cel·la són iguals i el grup espacial deixa que es puguin anar formant cel·les alternades en diferents orientacions.

Si aquesta alternança és estricta (sempre es repeteix), la cel·la resultant serà una de més gran que contindrà les dues. En la difracció, però, la cel·la aparent serà més petita i contindrà una barreja de les dues. La solució a aquest problema, si es detecta a posteriori, és la de reprocessar les dades forçant una cel·la més gran.

Si aquesta alternança és per petits dominis en dues orientacions diferents, la macla s'en diu hemièdrica [21]. La difracció proporcionarà una cel·la apa-

rent que és la superposició dels dos dominis i no serà possible trobar una altra indexació que les separi [18]. A la figura 2 del primer article (plana 80) hi ha una imatge de difracció d'un cristall d'aquest tipus.

En una macla d'aquest tipus les reflexions es sobreposen segons la simetria d'aquesta macla. La intensitat de les reflexions serà la suma ponderada de les dues parts:

$$I(h, k, l) = \alpha I_1(h, k, l) + (1 - \alpha) I_2(h, k, l)$$

$$\alpha = V_1 / (V_2 - V_1); V_2 > V_1$$

on I_1 i I_2 són les intensitats individuals de cadascun dels components de la macla i α és la proporció en volum de cadascun dels dos components.

Per determinar si unes dades de difracció corresponen a una macla d'aquest tipus es poden comprovar varies estadístiques. En primer lloc, la distribució acumulada de les intensitats de les reflexions hauria de presentar un perfil en forma de corba de saturació [50, 45, 21] (fig. 3.16). Un cristall maclat d'aquesta manera té un perfil sigmoide.

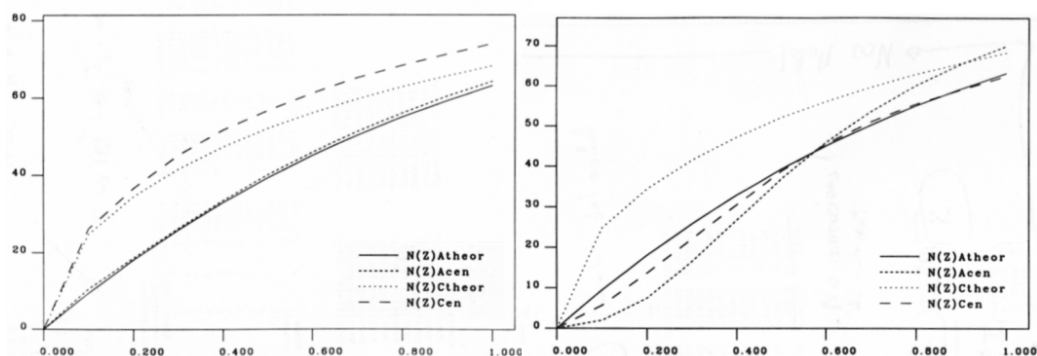


Figura 3.16: Distribució acumulada. Distribucions acumulades d'un cristall no maclat i un de maclat. La corba amb perfil sigmoide indica que el cristall està maclat. Esquerra: proteïna TrwB [22]. Dreta: un dels cristalls del connector amb simetria $P4_22_12$. A l'eix horitzontal es representa el valor de $I/\langle I \rangle$ amb el promig fet per capes de resolució i a l'eix vertical el tant per cent de reflexions.

Altres característiques que es poden comprovar són les distribucions de Wilson [59] i d'Stanley [50]. S'ha comprovat que aquestes distribucions presenten uns valors mitjos determinats que diferencien si la difracció prové d'una cel·la maclada o no:

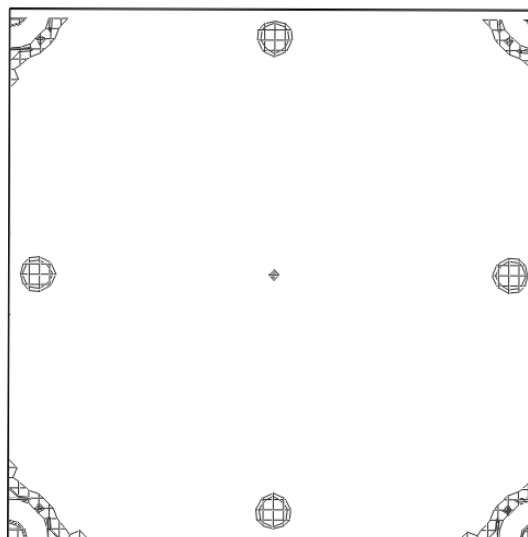
$$\text{Stanley : } \frac{\langle I^2 \rangle}{\langle I \rangle^2} = \begin{cases} > 2 & \text{desordre} \\ 2 & \text{normal} \\ 1.5 & \text{macla} \end{cases}$$

$$\text{Wilson : } \frac{\langle |F| \rangle^2}{\langle I \rangle} = \begin{cases} > 0.785 & \text{macla} \\ = 0.785 & \text{normal} \\ < 0.785 & \text{desordre} \end{cases}$$

Tots aquest mètodes estan basats en l'estadística de les reflexions per una distribució aleatòria d'àtoms dins d'una cel·la unitat. Pot passar que per motius geomètrics la distribució real dels àtoms no sigui prou aleatòria. Per exemple, si hi ha simetria no cristal·logràfica o bé si hi ha grans arees de solvent. Llavors les reflexions no segueixen tant fidelment aquestes estadístiques, i és d'esperar desviacions al comportament estàndard.

Per això és necessari assegurar-se també amb altres mètodes *a posteriori* si la cel·la pot estar maclada. El més evident, és un aturament de la qualitat de l'afinament, a valors alts d'R i d'R_{free}, però a l'hora mostrant mapes millors que els esperats a aquests valors. Una altra possibilitat és l'obtenció de conjunts de solucions sobreposades en el Reemplaçament Molecular, o bé solucions molt clares a partir de la síntesi de Patterson, però que impliquen empaquetaments impossibles (fig. 3.17).

Figura 3.17: Mapa de Patterson. Projectió al llarg de l'eix c del mapa de Patterson d'un dels cristalls amb simetria $P4_22_12$. L'empaquetament que implica un connector aproximadament centrat en cada unitat asimètrica no es pot donar, ja que les dimensions del connector ho impedeixen. El mapa està representat a 4σ del valor mig i els quatre pics estan situats a (0.5, 0.05, 0.5) i simètrics.



Gràcies a un mapa de Patterson d'aquest tipus, es pot preveure quines són les cel·les individuals que han format la macla. Han de ser cel·les idèntiques, relacionades per una rotació simple, per exemple de 90 graus.

Resoldre aquest tipus de macla comporta dificultats, que degraden la qualitat del resultat i actualment no s'han resolt gaires estructures d'aquest tipus. Sol ser molt més efectiu tractar d'aconseguir cristalls no maclats.

Si no es troba la manera d'aconseguir cristalls simples, la difracció és de prou bona qualitat, (i l'esforç val la pena), és possible separar la macla (*desmaclar*) a partir de certes suposicions sobre l'estructura:

Si prenem dues reflexions (h i h') relacionades per la simetria d'una macla hemièdrica [21]

$$Im(h) = \alpha I(h) + (1 - \alpha)I(h')$$

$$Im(h') = (1 - \alpha)I(h) + \alpha I(h')$$

on Im representa les reflexions observades (maclades) i I les del grup espacial no maclat, l'equació només té solució quan α és diferent de 0.5:

$$\begin{cases} I(h) = \frac{(1-\alpha)Im(h') - \alpha Im(h)}{1-2\alpha} \\ I(h') = \frac{(1-\alpha)Im(h) - \alpha Im(h')}{1-2\alpha} \end{cases}$$

Per resoldre macles amb $\alpha=0.5$ es pot construir un model de partida similar al de reemplaçament molecular, bé traçant un mapa obtingut per derivats i separant les cadenes, o bé a partir d'un model de criomicroscòpia.

A partir d'aquest model, i amb ajuda del mapa de Patterson, o alguna altra manera que ens permeti situar la molècula al seu lloc en la unitat asimètrica, es construeix la macla afegint les contribucions de cada orientació. D'aquesta manera es construeix una "cel·la maclada" tal com s'ha observat. L'afinament es farà només per una molècula individual, però en cada pas en el que calgui comprovar la fiabilitat del afinat caldrà reconstruir la macla i calcular les estadístiques a partir de les reflexions calculades d'aquesta manera.

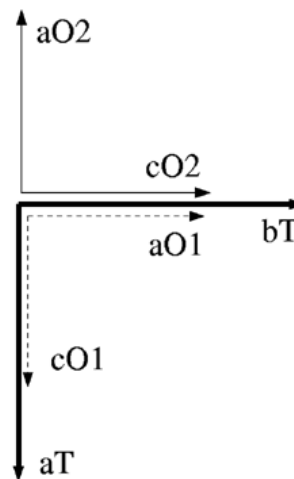
3.7.2 Procediment pràctic

Per a desmaclar les reflexions dels cristalls amb grup espacial aparent $P4_22_12$ es va haver d'afrontar un altre problema: com que el model era de molt baixa resolució, el nombre de reflexions a ajustar contra un model era proporcionalment molt baix, pel que calia alhora tractar d'extendre les fases a més resolució per comprovar la fiabilitat del procediment.

El grup espacial aparent era el $P4_22_12$, amb $a=b=170.2 \text{ \AA}$ i $c=156.9 \text{ \AA}$ (tetragonal o T, d'ara endavant). Per consideracions d'empaquetament i de la síntesi de Patterson (fig. 3.17) es va veure que es podia descomposar en dues cel·les amb simetria $P2_12_12$ $a=c=170.2 \text{ \AA}$ i $b=156.9 \text{ \AA}$ (ortoròmbic o

O, d'ara endavant). Una rotada 90° al voltant de l'eix b respecte l'altra. La cel·la tetragonal s'aconseguia sumant aquestes dues i rotant 90° al voltant del nou eix eix c (fig. 3.18).

Figura 3.18: Eixos de la macla. Relació entre els eixos de les cel·les que formen la macla i la cel·la aparent. O1 i O2 són les dues cel·les ortoròmbiques i T és la tetragonal. O1 i O2 ($a=c$) estan entre elles rotades 90° al voltant de l'eix b. La cel·la T ($a=b$) es construeix permutant els eixos de manera que l'eix b de les ortoròmbiques coincideixi amb l'eix c de la tetragonal.



Per tant una reflexió (h, k, l) en el grup espacial maclat seria la suma

$$I_T(h, k, l) = \alpha I_O(k, l, h) + (1 - \alpha) I_O(-h, l, k)$$

de les reflexions del grup no maclat.

Aquesta idea la vàrem implementar escrivint un programa que combinava la búsqueda de la posició del connector i la formació de la macla. Les coordenades del centre del connector estaven ben definides en el pla a-b gràcies al mapa de Patterson, però calia trobar també la posició al llarg de l'eix c i també la orientació dels lòbuls.

El programa feia una cerca de diferents posicions i orientacions del connector en una cel·la $P2_12_12$. Primer es construïa un mapa de densitat electrònica a baixa resolució a partir del model de microscòpia (de 20 a $12\text{\AA}$). Llavors es calculaven les reflexions exteses al grup espacial $P1$, per tenir totes les simètriques. Amb elles es calculaven les reflexions maclades en $P4_22_12$ a partir de la fórmula anterior. La correlació entre les reflexions es calculava amb una fórmula idèntica a la de la correlació en el Reemplaçament Molecular (plana 3.5.2), però fent servir intensitats. Les correlacions es calculaven globalment per totes les reflexions, per capes de resolució i també per les reflexions $(h, k, 0)$ i $(h, 0, l)$.

Les reflexions $(h, k, 0)$ representen, en el grup espacial $P4_22_12$, la projecció sobre el pla a-b, que era convenient ja que en aquesta direcció es coneixia bé el centre de la molècula i a més coincidia amb la màxima resolució del model

de microscòpia. Les reflexions $(h,0,l)$ representen la projecció en el pla a - c i es feien servir per buscar la màxima correlació en l'eix c . No es varen poder fer servir reflexions $(0,k,l)$ a causa de que se'n van mesurar molt poques. A més es podia aplicar un decaïment a cada reflexió proporcional a la resolució i dependent de la direcció per a simular l'anisotropicitat de les reflexions.

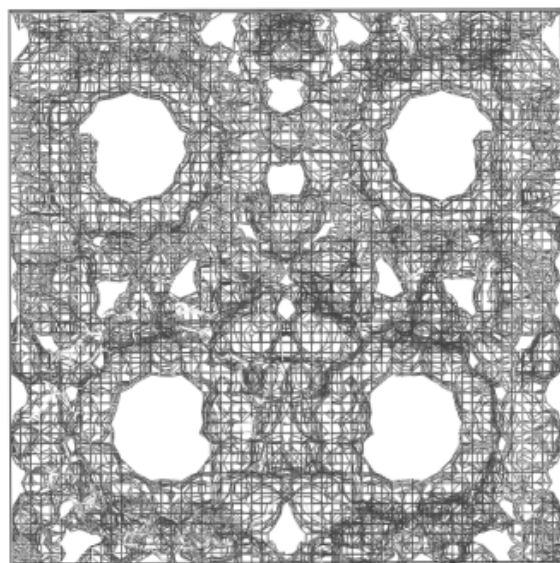


Figura 3.19: Macla reconstruïda. Cel·la tetragonal maclada reconstruïda a partir d'una de les possibles orientacions. El connector està aproximadament centrat en la unitat asimètrica, però la mida d'aquest no permet que hi càpiga. En les cel·les ortoròmbiques, al tenir menor simetria hi ha menys connectors en la cel·la i és possible l'empaquetament.

Per a poder comprovar la fiabilitat de la reconstrucció, les orientacions de major correlació se'ls aplicaven cicles d'extensió de fases amb $NCS=12$, alternats amb desmaclat de les reflexions a partir de les noves fases.

Per assegurar la fiabilitat del mètode es va construir una macla fictícia a partir d'un model fet modificant l'estructura d'una proteïna de dimensions similars i donant-li una forma semblant a la del connector. Amb aquest model es va simular un de microscòpia, calculant un mapa de baixa resolució (a $10\text{\AA}$), i també les reflexions d'una cel·la maclada. Els tests realitzats localitzaven correctament la proteïna test, fins i tot si la forma es modificava una mica, o treballant a menys resolució. Les correlacions eren de l'ordre del 99% tant en les búsquedes en dues dimensions com en les de tres. En canvi, quan es van fer servir les dades reals, les correlacions van caure a menys del 78% en dues dimensions i a un 46% en tres, no mostrant cap solució prominent. Les búsquedes al llarg de l'eix c no van donar cap posició especial. A la secció 3.8.2 de la plana 70 hi ha una argumentació de per què aquestes búsquedes tenien tant poca precisió.

3.8 Microscòpia electrònica

Els electrons poden ser conduïts per mitjà de lents electromagnètiques. Així, en la microscòpia electrònica, al contrari que a la cristal·lografia de raigs X, és possible enfocar els electrons, i obtenir imatges directes de la mostra en estudi. Aquestes imatges representen projeccions de l'estructura perpendiculars al pla focal de microscopi. Cal, per tant, un procés de reconstrucció de la imatge a partir de diverses projeccions.

Si hi ha cristalls bidimensionals disponibles, aquests són adients per experiments de microscòpia electrònica (difracció d'electrons) (figs. 1.4, plana 6 i 1.9, plana 9). Els electrons tenen poca penetració en les mostres, i cal evitar alhora interaccions extres dels electrons amb el cristall. Es poden recollir també dades de difracció d'electrons d'aquests cristalls. D'aquesta manera, s'obtenen també factors d'estructura que es poden combinar amb les imatges directes per a obtenir major resolució.

3.8.1 Reconstrucció de la M.E.

Els models de microscòpia electrònica ens eren presentats en forma de seccions perpendiculars al llarg de l'eix del connector. Calia, doncs, construir un programa per obtenir un mapa tridimensional compatible amb els mapes usats en cristal·lografia.

Les seccions venien ordenades en forma de fitxers numerats seqüencialment. Cadascun contenia les dades de densitat electrònica en forma seqüencial en valors enters d'1 bit o flotants de 4 bits, representant una matriu de 128×128 punts. No contenien cap informació sobre la mida real. El volum tridimensional s'aconseguia sobreposant aquestes seccions en una matriu tridimensional de $128 \times 128 \times (\text{nombre de seccions})$. El nombre de seccions era variable segons la resolució del model. En la criomicroscòpia la resolució obtinguda en el pla perpendicular del connector era més alta que al llarg de la seva alçada, pel que el nombre de seccions era només de 30 a 60.

Depenent de quin model, calia primer fer una *neteja* per eliminar soroll de fons i imatges residuals d'altres connectors veïns. Aquests filtres es van incorporar dins dels programes: Es podia determinar un valor de tall i eliminar tota la densitat menor a aquest, o bé fer servir una màscara per a cada imatge. La màscara era un fitxer binari del mateix format que les seccions. Es construïa la secció modificada multiplicant prèviament punt a punt per la màscara.

Es va decidir de construir mapes en format "newEZD" ja que resultaven més fàcils d'editar a posteriori al ser composts només de caràcters ASCII. El pas a altres formats s'aconseguia fàcilment amb el programa `mapman`[36]. La

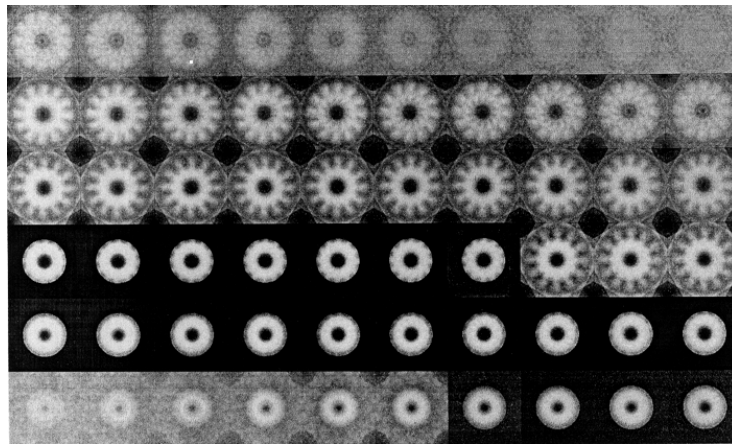


Figura 3.20: Seccions M.E.. Seccions del model de microscòpia electrònica, que es van fer servir per a construir els models inicials per l'extensió de fases a alta resolució.

mida del model s'obtenia escalant la cel·la unitat de la reconstrucció.

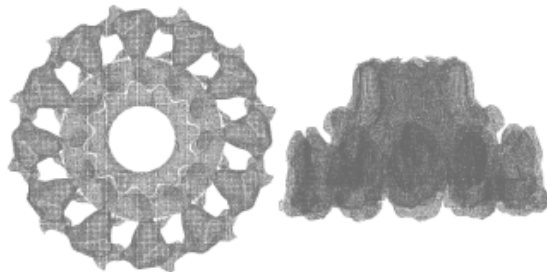


Figura 3.21: Reconstrucció de la M.E.. Vista frontal i lateral de la reconstrucció del model de microscòpia electrònica, construït a partir de les seccions de la figura 3.20

Per conveniència amb els programes estàndard, a partir de la reconstrucció es van construir altres programes que transformaven la densitat a pseudoàtoms, a pseudoaminoàcids o a màscares per aplicar NCS.

3.8.2 Ajust de la microscòpia i del model afinat

Amb les coordenades del model final afinat es va fer un ajust amb el model de microscòpia per a veure les diferències entre aquests dos mètodes. A la figura 3.22 es mostra aquest ajust, junt amb un mapa del connector afinat a una resolució equivalent. Com es pot veure, la forma general d'ambdós mapes és similar, tal com era d'esperar, hi ha diferències prou importants

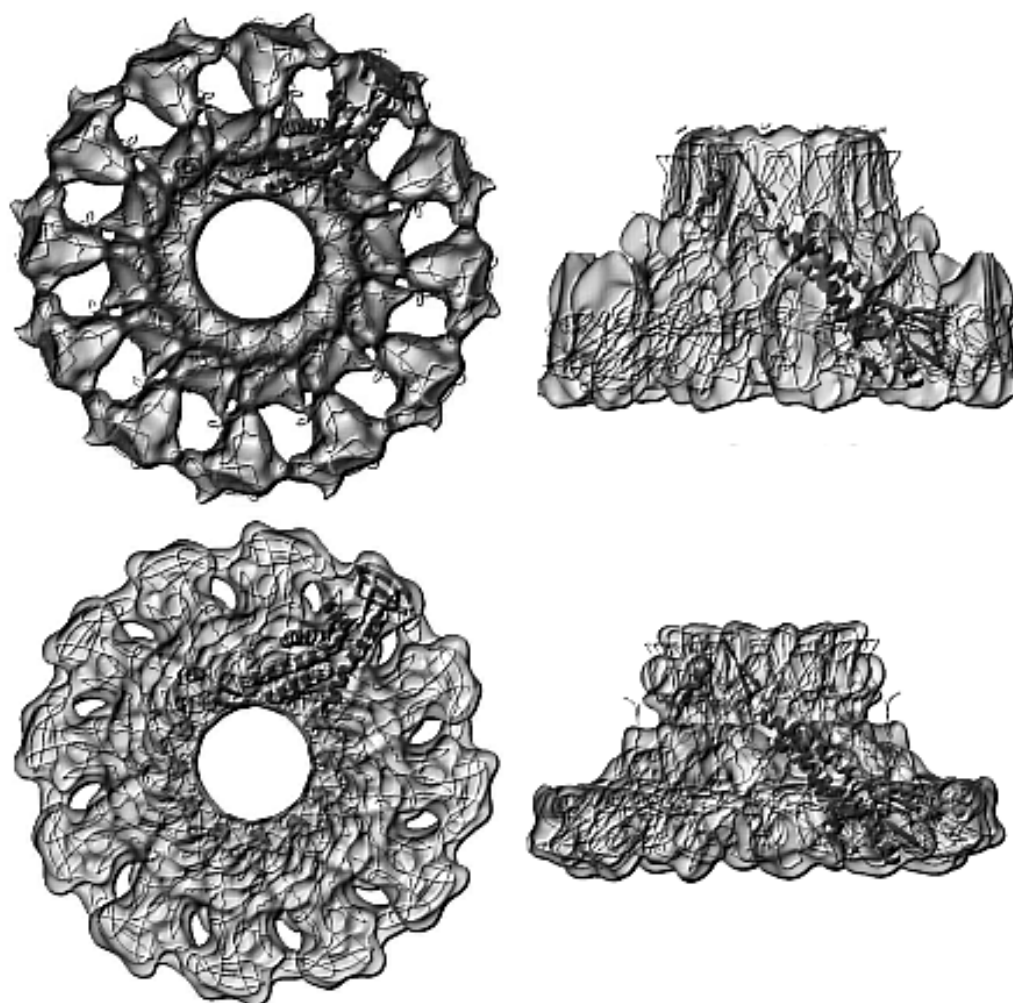


Figura 3.22: Ajust de la microscòpia i del model afinat. Vistes axials i laterals del mapa de microscòpia (a dalt) i del mapa final (a baix), a baixa resolució (10 Å), amb la cadena principal superposada. La forma general és similar, però els mapes difereixen força en la zona dels lòbuls, essent menys alts i més amples en el model final.

com per a impedir que la cerca per correlacions fos prou acurada a baixa resolució.

El model final és més cilíndric en la zona dels lòbuls, impedit un bon contrast entre solvent i proteïna, pel que la cerca per correlació a baixa resolució no podia discernir clarament l'orientació exacta del connector en la projecció al llarg de l'eix c. També és força diferent en l'alçada dels lòbuls.

Els del model afinat són la meitat d'alts. Aquest detall impedia també una bona localització al llarg de l'eix *c* fent servir el model de microscòpia. Per això les búsquedes per correlació donaven poc contrast entre les orientacions possibles.

Altres detalls que difereixen estan en el tronc central. En aquest cas, els lòbuls del model final no són tant prominents ja que corresponen a zones poc ordenades de la densitat. En el dibuix inferior es veuen les cadenes traçades fora del mapa, ja que aquest no era prou dens pel contrast de la figura.

3.9 El PDB

El Protein Data Bank (PDB, www.pdb.org) és la base de dades centralitzada on es publiquen les estructures de proteïnes i d'àcids nucleics fetes per difracció de raigs X, per ressonància magnètica nuclear, difracció de neutrons, models teòrics, etc.

Un "pdb" és un arxiu ASCII on cada línia conté una entrada amb les coordenades dels àtoms de la molècula, de quin àtom es tracta, a quin aminoàcid, nucleòtid o compost pertany, quina ocupació i quin *B factor* tenen. Es tracta d'un format ordenat per columnes per permetre búsquedes i modificacions ràpides, fàcils d'automatitzar i d'incorporar en programes. Altres entrades del pdb contenen informació sobre l'estructura secundària, el cristall, com s'ha aconseguit la molècula, com s'ha fet l'experiment, on s'ha publicat, i quines estadístiques ha donat. Es tracta de que amb un sol fitxer es trobi tota la informació per regenerar l'estructura i per calibrar la seva qualitat, així com la informació genètica necessària, generalment posada en forma de referències a bases de dades de gens. A l'hora d'escriure aquest text (30 de novembre de 2001) el nombre de molècules dipositades és de 16651.

El connector s'ha dipositat en el PDB amb el codi **1h5w**.