Universitat Autònoma de Barcelona

escola d'enginyeria

Bioinformàtica::Comparació de genomes de eucariota

Xavier Martínez Clemente

Projecte Fi de Carrera

Enginyeria Informàtica

Directors: Jordi Gonzàlez i Sabaté

Mario Huerta

Index

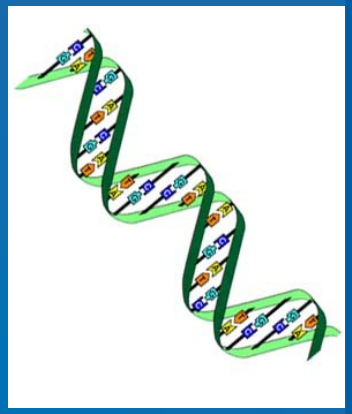
- Introducció
- Estat de l'art
- Objectius
- Especificacions tècniques
- Desenvolupament
- Conclusions
- Bibliografia
- Torn de preguntes

Introducció - IBB

- Institut de Biotecnologia i de Biomedicina
- Diverses línies d'investigació:
  - Biologia cel·lular i d'estructura, genòmica, inmunologia, microbiologia, proteòmica
  - Bioinformàtica: Comparació genòmica

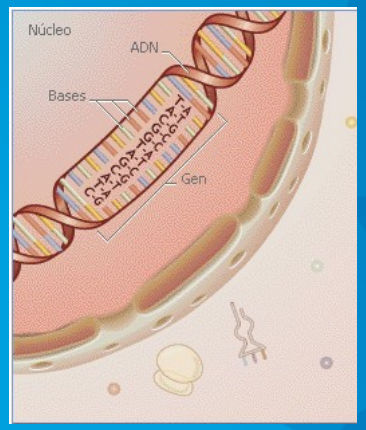
Introducció - Fonaments

- Cèl·lules: Element viu més petit
  - Glúcids i lípids
  - Proteïnes
  - Àcids nucleics → ADN
    - Adenina (A)
    - Timina (T)
    - Guanina (G)
    - Citosina (C)



Introducció - Fonaments

- Què és un genoma?
  - Totalitat d'informació genètica codificada al ADN



ATGCGCTAATGCTACGTAATCG
TAGCATGCGCTAATGCGCTACG

L'estudi d'un genoma i la comparació amb el d'una altra espècie ens permetrà entendre el funcionament dels seus organismes i els gens que perduren dels seus avantpassats

Introducció - Fonaments

- Els tres grans dominis de classificació d'espècies:

Archaea

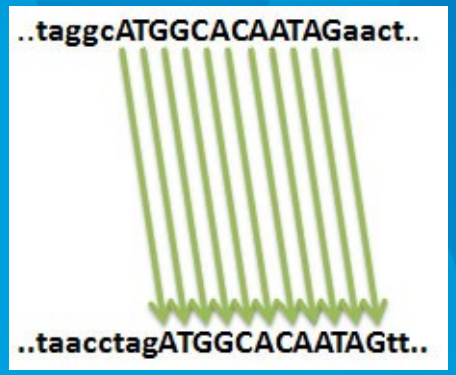
Bacteria

Eucaryota

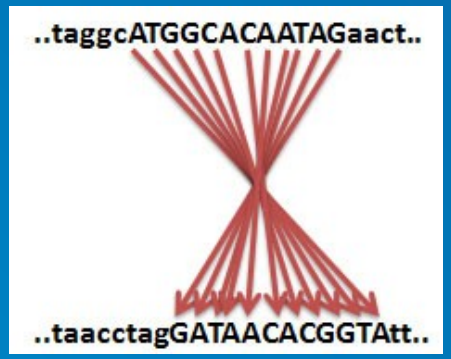
NCBI National center for biotechnology information

Estat de l'art - MUMs

- Algoritme MUMOL
  - Comparació de dos genomes
- Generació de MUMs(Maximal Unique matchings)
  - Seqüència de bases coincidents als 2 genomes
  - Major longitud trobada de la seqüència
  - Seqüència única a tot el genoma



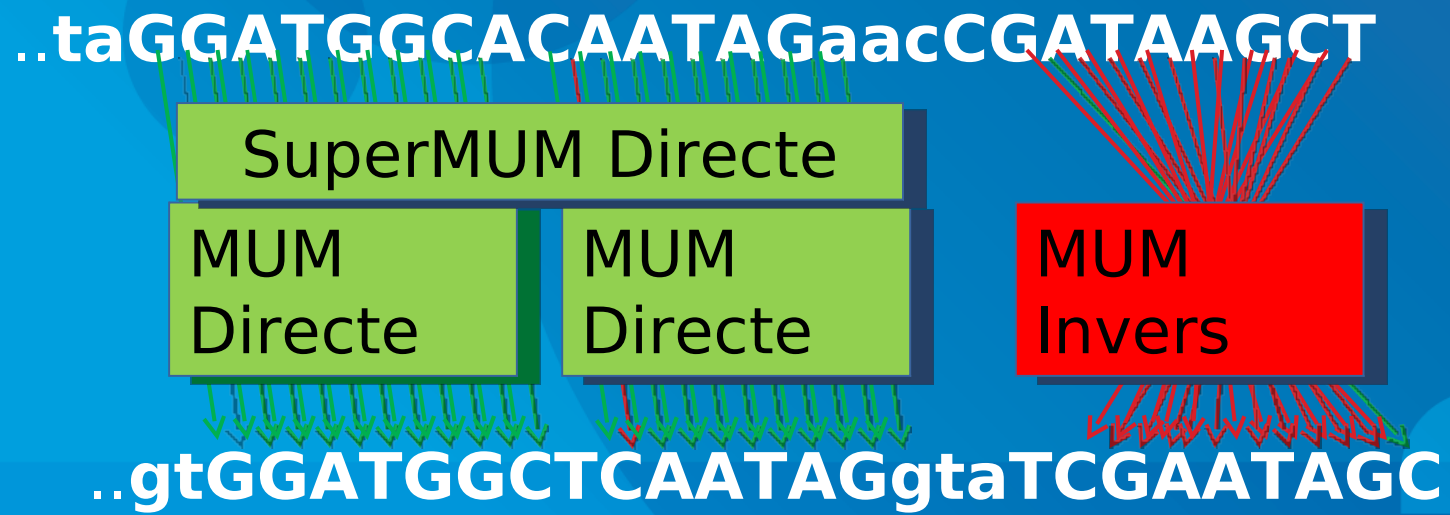
MUM directe



MUM invers

Estat de l'art - SMUMs

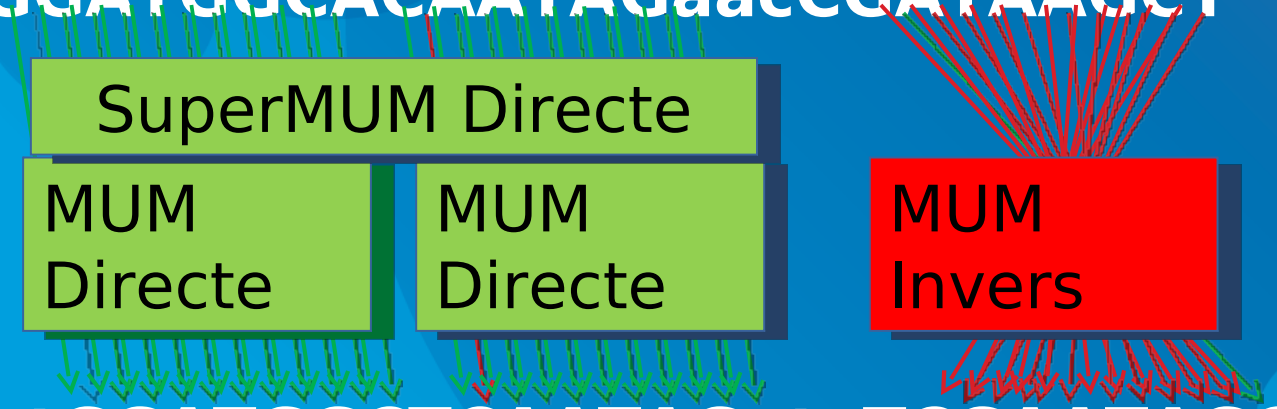
- Algoritme pel càlcul de SMUMs (SuperMUMs)
  - Agrupació de MUMs mitjançant *Approximate String Matching*



Estat de l'art - SMUMs

- Algoritme pel càlcul de SMUMs (SuperMUMs)
  - Mateix ordre de MUMs als 2 genomes
  - Espai entre MUMs < longituds MUMs x multiplicador
  - Un SMUM no pot estar inclós a un altre SMUM
  - MUMs inclosos a un SMUM queden absorbits en aquest

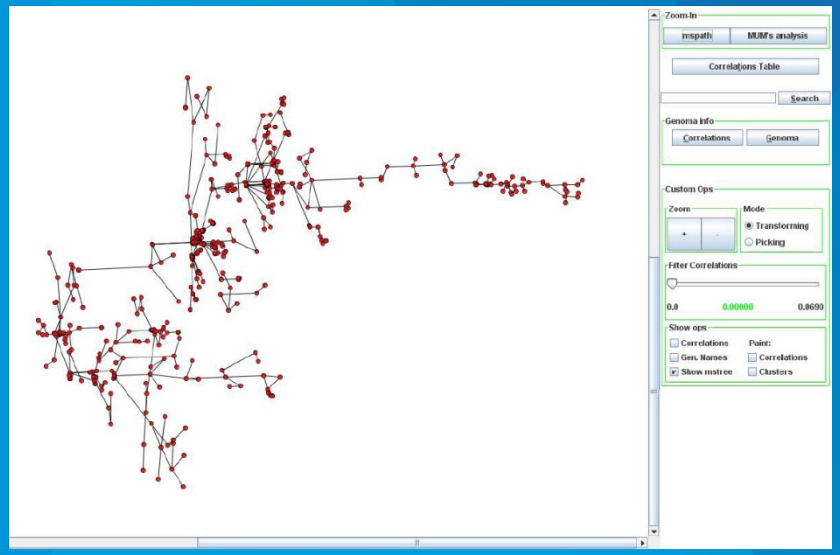
..**taGGATGGCACAATAGaacCGATAAGCT**



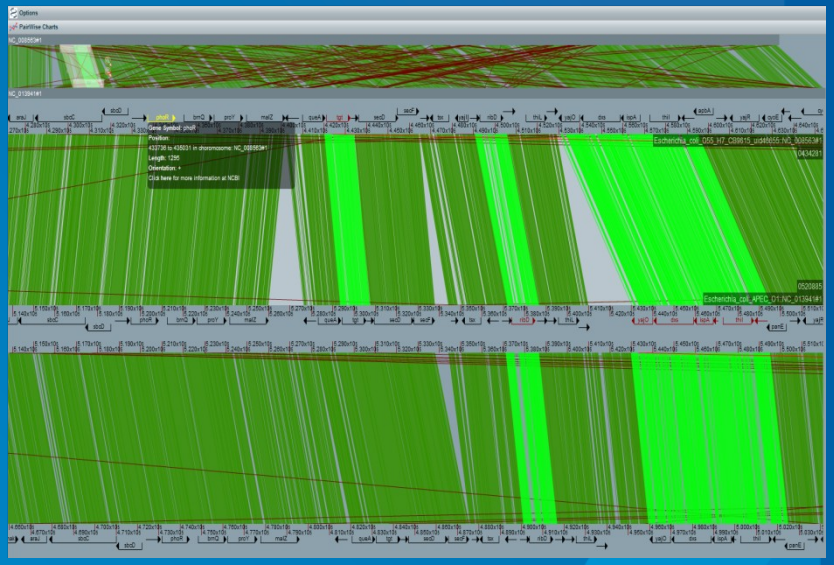
..**gtGGATGGCTCAATAGgtaTCGAATAGC**

Estat de l'art – Aplicacions finals

- Mummy tree
  - Representació del arbre de genomes per similitud



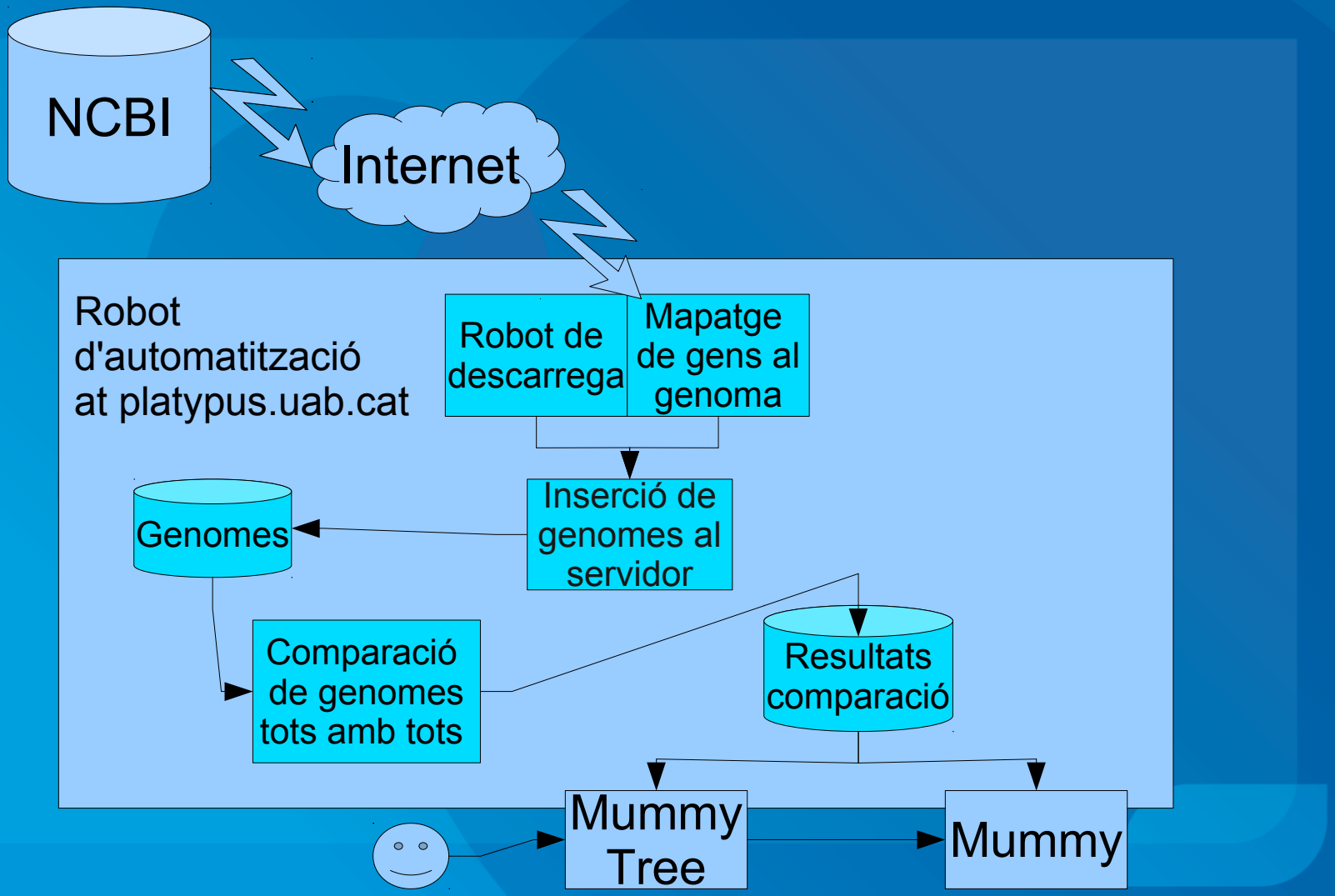
- Mummy
  - Representació de comparació de genomes



Objectius

- Automatització de la comparació de genomes d'eucariota a mesura que aquests són seqüenciats
- Adaptar comparació de genomes per eucariotes
  - Generació de MUMs per cromosomes
  - Generació de SMUMs a partir de SMUMs
- Paral·lelització del càlcul de MUMs
- Automatització i adaptació del programa de mapatge de gens sobre el genoma
- Actualització automàtica bimensual de les comparacions de genomes nous o actualitzats

Objectius

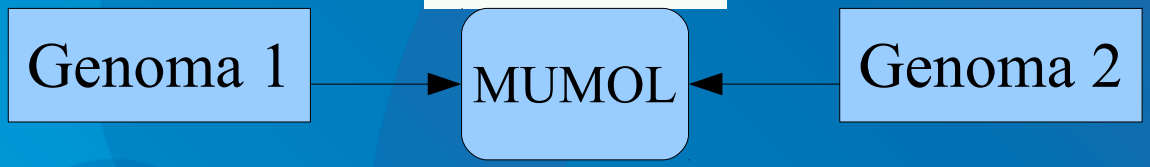

NCBI
Internet
Robot d'automatització at platypus.uab.cat
Robot de descarrega
Mapatge de gens al genoma
Genomes
Inserció de genomes al servidor
Comparació de genomes tots amb tots
Resultats comparació
Mummy Tree
Mummy

Especificacions tècniques

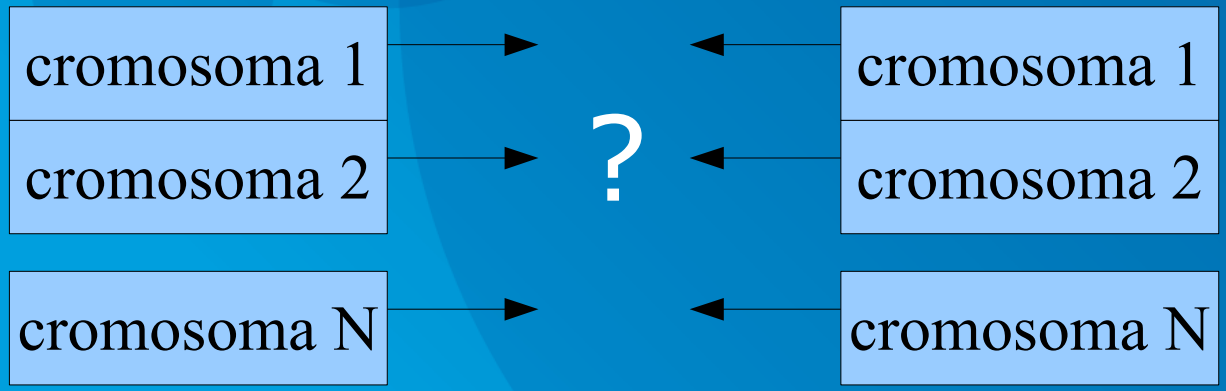
- Entorn de desenvolupament
  - OS: GNU/Linux CentOS 5.5
  - Processador: 4 x Intel Xeon X5650 (24x2,67GHz)
  - RAM: 63Gb RAM
- Eines de desenvolupament:
  - C/C++
  - Scripts Bash
  - Java 1.6 (Eclipse)

Desenvolupament – Adaptació eucariotes

- A bacteris...



- A eucariotes...



Desenvolupament – Adaptació eucariotes

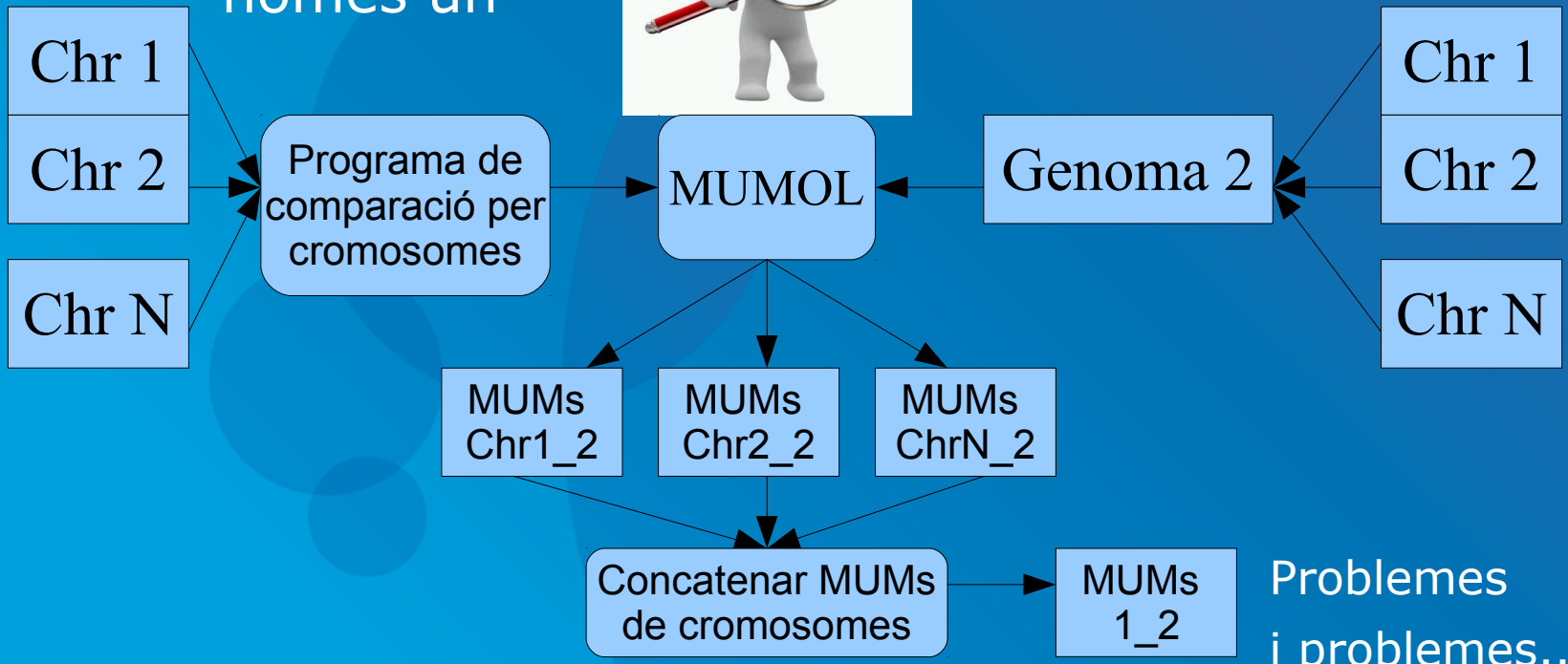
- Concatenar cromosomes

- MUMOL necessita carregar a RAM
  - Genoma 2 sencer (pe. 2 Gb)
  - Genoma 1 en estructura d'arbre
    - Exemple: 2 Gb del genoma 1 * 120 = 240 Gb !

Disposem de 63 Gb...

Desenvolupament – Adaptació eucariotes

- Concatenem només un



Problemes i problemes...

- Important! Per optimitzar la memòria construim l'arbre amb el cromosoma més petit

Desenvolupament – Adaptació eucariotes

- Problema al concatenar, desplaçament de cromosomes

1-chr1_2

Posició inicial genoma 1	Posició inicial genoma 2	Longitud del MUM
3	54	26
30	2	20
...	...	...

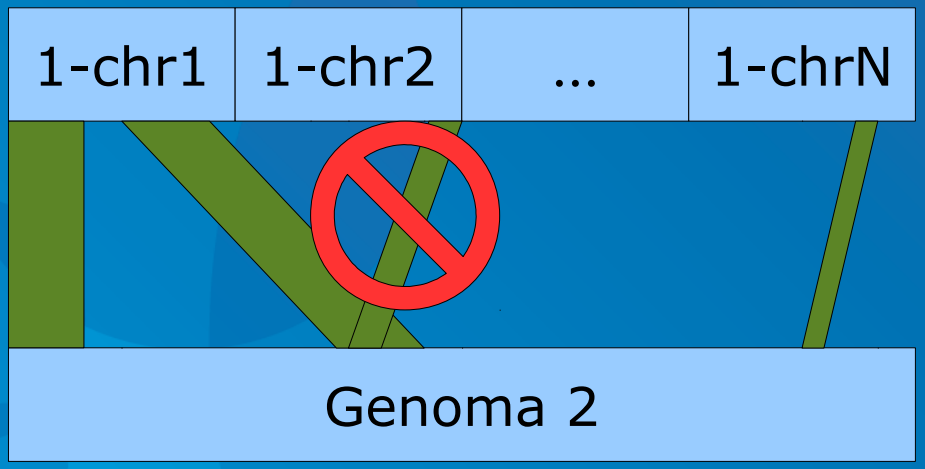
1-chrY_2 (desplaçament de 100)

Posició inicial genoma 1	Posició inicial genoma 2	Longitud del MUM
3	28	26
30	158	20
...	...	...

Posició inicial genoma 1	Posició inicial genoma 2	Longitud del MUM
3	54	26
30	2	20
...	...	...
103	28	26
130	158	20

Desenvolupament – Adaptació MUMs

- Problema al concatenar, eliminar No-MUMs



- Algoritme eficient per eliminar aquests casos

Desenvolupament – Paral·lelització

- A bacteris, la limitació és el número de CPUs
- A eucariotes, la limitació és la memòria RAM
  - Necessitem consultar fiablement la memòria RAM disponible a cada moment:
    - Funció popen a /proc/meminfo
    - Problema: memòria cache s'allibera conforme la necessitem, però no queda reflectida al valor de memòria lliure

Memòria ocupada real = (MemTotal-MemFree-Cached-Buffers)

Memòria ocupada i no disponible = Memòria ocupada real + (MemTotal-52Gb)

Memòria lliure = MemTotal – Memòria ocupada i no disponible

Desenvolupament – Paral·lelització

Solucionar l'inconvenient de la reserva lenta de memòria del MUMOL

Algorisme:

Generem una llista amb la mida que ocuparan els cromosomes a RAM.

Mentre quedin cromosomes a la llista.

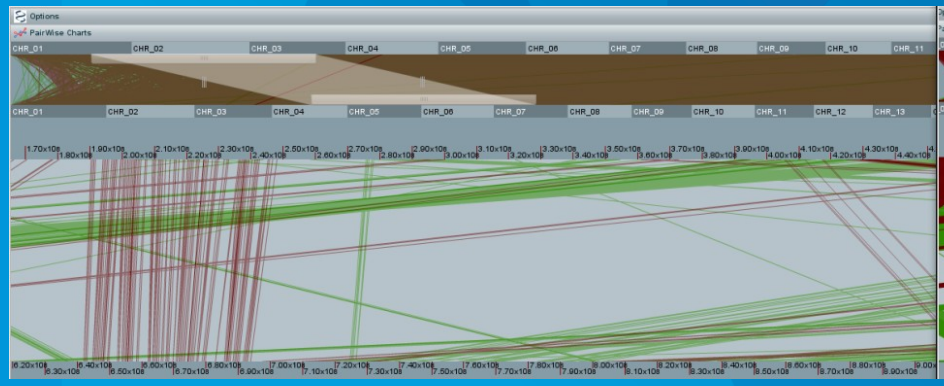
- Comprovem la memòria disponible i la guardem en una variable.
- Recorrem tota la llista de cromosomes en ordre, primer els més grans
- Si queda memòria (a la variable)
- Si el cromosoma actual cap en memòria, l'eliminem de la llista, restem l'espai que ocuparà i creem un nou procés (crida fork) que llança els càlculs de MUMs pel cromosoma.
- Esperem a que acabin els fills.

Desenvolupament – Paral·lelització

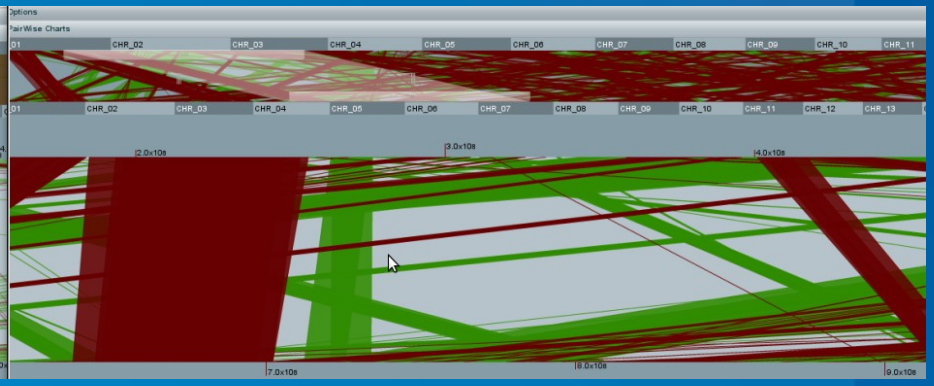
- Ús de Semafors
- Reanomenar fitxers que es reescrivien

Desenvolupament – SMUMs

- Adaptació del càlcul de SMUMs a eucariotes
- Necessitat de habilitar més passades del càlcul
- Adaptar fitxers SMUMs a MUMs i eliminar duplicats
- Realitzar moltes proves... :)



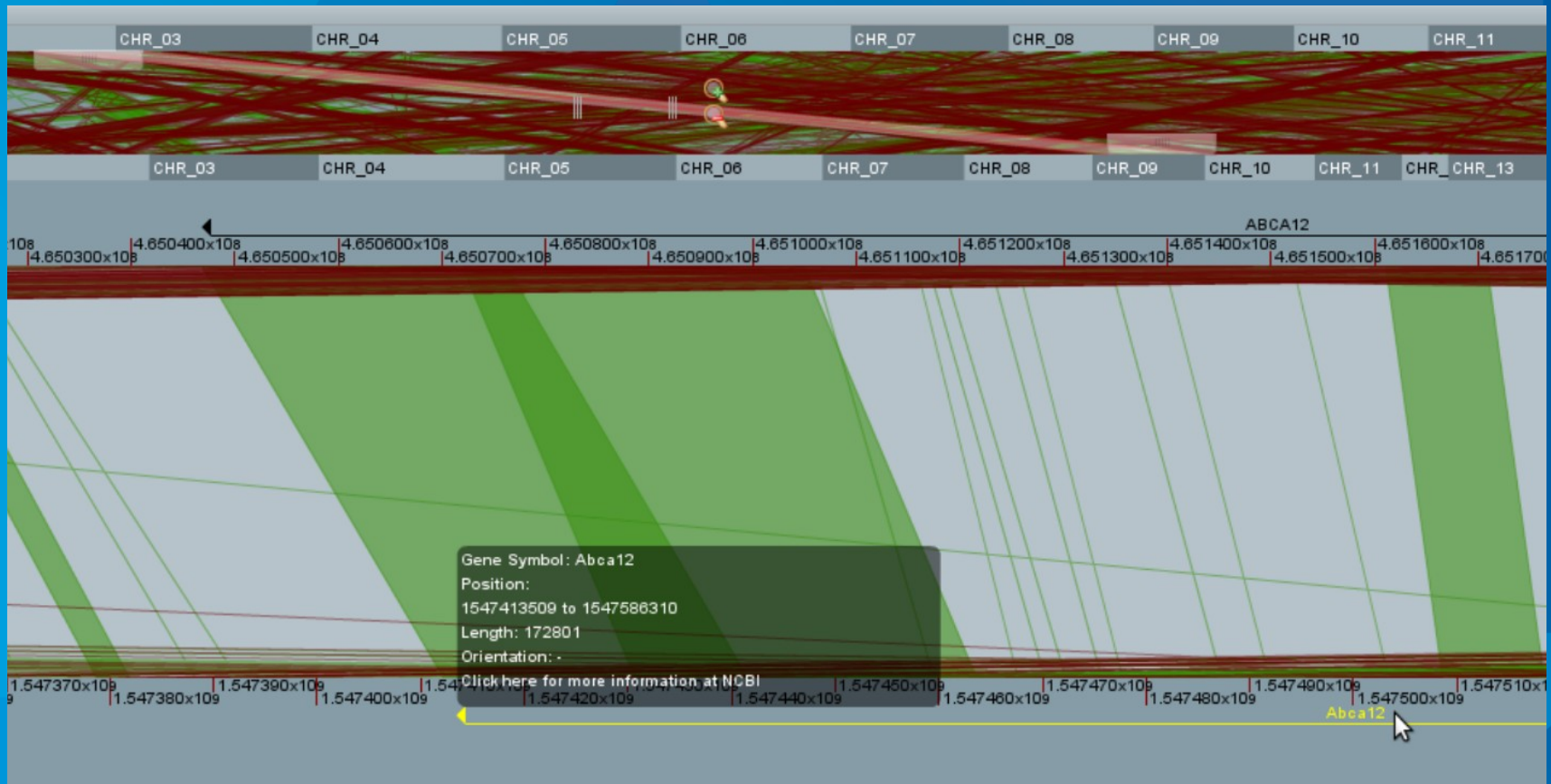
Una passada de 100



Tres passades de 30

Desenvolupament – Mapatge de gens

- Adaptació a eucariotes – Diversos problemes



Desenvolupament – Mapatge de gens

- Adaptació a eucariotes – Diversos problemes
  - Adaptació del fitxer accession. Què és?
  - Adaptació del identificador de cromosoma del fitxer accession amb el nom del fitxer de gens descarregats del NCBI
  - Ignorar fitxers de gens en proves que no apareixen al accession
  - Afegir funcionalitats al parser del mapatge

Desenvolupament – Robot de descarrega

- Automatització total de la descarrega
  - Connexió al FTP del NCBI
  - Ús de funcions Java per tractar FTP
    - Canviar i llistar carpetes, descarrega,...
  - Detectar genomes vàlids, caos!
  - Descarregar genomes que ens falten
  - Descarregar actualitzacions de genomes
  - Classificar genomes segons el seu taxò mitjançant consultes a e-utils
    - http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=taxonomy&id=XX&retmode=xml

Conclusions

- Hem aconseguit automatitzar per complet tot el procés de comparació de genomes eucariotes.
- Hem fet tots els canvis necessaris per adaptar la comparació de genomes eucariotes
- Hem paral·lelitzat el càlcul de MUMs per tenir els resultats sempre actualitzats, conforme es van seqüenciant genomes arreu del món.
- Hem adaptat el programa de mapatge de gens sobre el genoma, necessari per veure els gens al Mummy
- Hem programat una actualització bimensual per comprovar i realitzar comparacions de genomes nous o actualitzats

Conclusions – Treball Futur

- Unió d'interficies gràfiques:
  - Representació del arbre de genomes per similitud
  - Representació de comparacions de genomes
- Optimització d'ús de memòria:
  - MUMOL per cromosomes
- Publicació d'article a revista científica per donar a coneixer el servidor a la comunitat científica mundial

Bibliografia

- Servidor per a la comparació de genomes online: http://platypus.uab.cat
- National Center for biotechnology information (NCBI) http://www.ncbi.nlm.nih.gov/

Bibliografia

- Suffix Tree Construction with slide nodes, Mario Huerta. technical report LSI-02-63-R Dep. Llenguatge i Sistemes Informàtics, Universitat Politècnica de Catalunya (2002).
- Mario Huerta and Xavier Messeguer. Efficient space and time multicomparison of genomes. Research Report LSI-02-64-R Dep. Llenguatge i Sistemes Informàtics, Universitat Politècnica de Catalunya.(2002).
- Domènec Farré, Romà Roset, Mario Huerta, José E. Adsuara, Llorenç Roselló, M. Mar Albà, Xavier Messeguer. Identification of patterns in biological sequences at the ALGGEN server. PROMO and MALGEN.Nucleic Acids Research. 2003 31: 3651-3653 (2003).

Universitat Autònoma de Barcelona

escola d'enginyeria

Bioinformàtica::Comparació de genomes de eucariota

Xavier Martínez Clemente

Projecte Fi de Carrera

Enginyeria Informàtica

Directors: Jordi Gonzàlez i Sabaté

Mario Huerta