Universitat Autònoma de Barcelona

escola d'enginyeria

ENGINYERIA INFORMÀTICA

MEMÒRIA PREVIA DEL PROJECTE

2682-1
Bioinformática::Comparación de genomas de eucariota

Signatura de l'estudiant	Signatura del director/a o directors/es
Nom: Xavier Martínez Clemente Data: 14/01/2010	Nom/s: Jordi Gonzàlez i Sabaté Mario Huerta Dpt: CVC IBB Data: 14/01/2010

1. Objectiu/s del projecte

Els genomes dels éssers vius contenen la informació genètica de cada espècie. Volem realitzar comparacions de genomes d'eucariota, els més grans, de forma eficient i obtenir resultats que serveixin tant per aplicacions mèdiques com per estudis evolutius d'espècies que es realitzarà mitjançant una aplicació web.

Entre els objectius a assolir, esmentaré els més importants tot i que aquesta seria una llista oberta:

- Automatització completa dels càlculs per generar els fitxers de la comparació dels genomes d'eucariota.
- Donat que fins ara, la comparació de dos genomes d'eucariota triga entre 1 i 2 dies, s'haurà de paral·lelitzar els calcul de la comparació, perquè l'aplicatiu sigui viable, aprofitant l'amplia memòria RAM i l'arquitectura multicore del servidor que disposem al IBB.
- Millorar la selecció final de les dades més rellevants de la comparació anomenades SMUM's (superMUM's, és a dir, agrupacions de MUM's (Maximal Unique Matching)). Actualment es selecciona el milió de SMUM's més gran i obtindríem millors resultats seleccionant els SMUM's més significatius, és a dir, els que ajudin a formar SMUM's encara més grans.
- Automatització del programa (sync_genes) encarregat de descarregar els noms de cada gen i situar-lo al seu lloc dins dels genomes d'eucariota comparats.
- Comprovació automàtica i periòdica de la seqüenciació de nous genomes o canvis significatius en la versió d'un genoma al repositori mundial de genomes NCBI (National Center for Biotechnology Information) [1]. La majoria de genomes d'eucariota no estan completament seqüenciats, pel que utilitzen un sistema de versions per controlar els canvis.
  Si els canvis són importants, haurem de baixar el nou genoma i realitzar la comparació de nou. A més, s'hauran de detectar canvis de posició de gens, així com el descobriment de nous gens fins ara desconeguts per a cada especie en qüestió.

2. Breu introducció a l'estat del art del tema proposat

Un genoma és la totalitat d'informació genètica que conté un organisme viu. [2] Aquest genoma està composat per una llarga seqüència de les quatre bases nitrogenades: A,C,T i G.

Si realitzem la comparació de genomes, podem obtenir una gran informació dels processos evolutius que han portat a l'existència de la gran varietat d'éssers vius que avui poblen la Terra. A més, aquesta comparació de genomes, té una important aplicació mèdica. De fet, l'assignació de funcionalitat als gens o per exemple, la detecció de gens que poden provocar malalties, s'acostuma a fer pel reconeixement de seqüències funcionals que es conserven d'un ésser a un altre.

Els genomes d'eucariota són enormes, a diferència dels genomes de bacteris i arqueobacteris, i això és un problema per la comparació. Per exemple, el genoma humà té tres mil milions de bases mentre que el genoma d'un bacteri té al voltant de tres milions, ens podem fer una idea. Amb aquest gran volum de dades és necessària la adaptació dels càlculs de comparació de genomes per a comparar eucariotes.

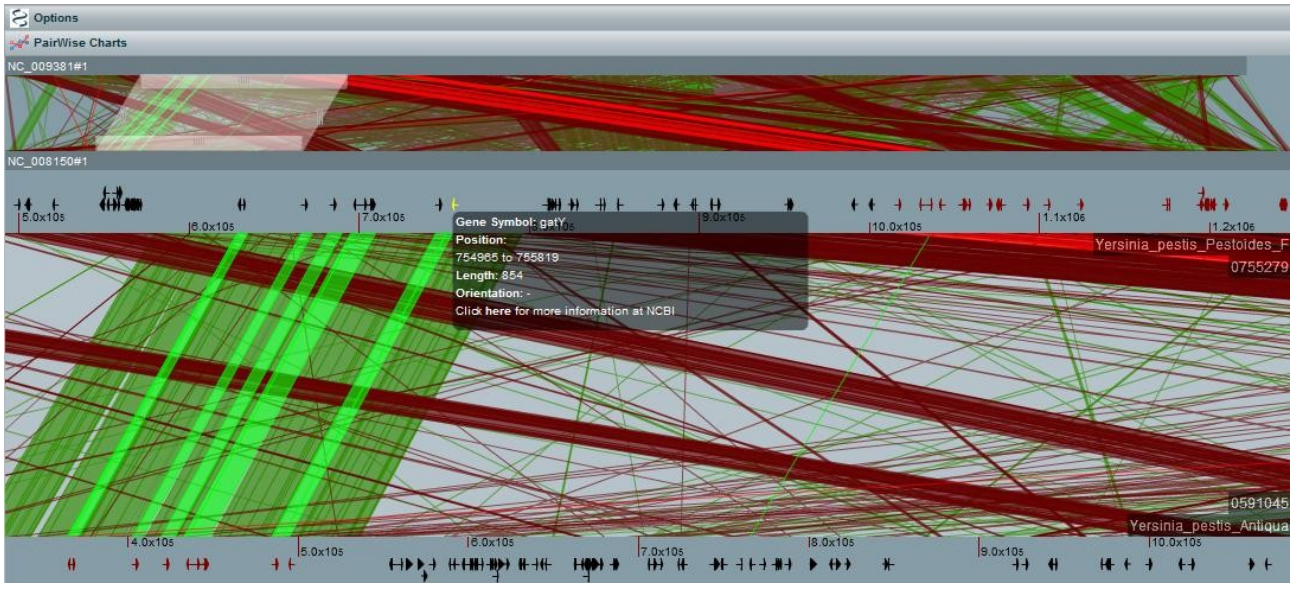
Existeixen diferents formes de comparar genomes. La tècnica que es vol fer servir es basa en trobar seqüencies comuns entre els genomes, calculades prèviament i anomenades MUM's (Maximal Unique Matching) [3][6]. Els MUM's són les sub-seqüencies de mida més gran i úniques de bases nitrogenades comuns als dos genomes que comparem. La totalitat de MUM's trobats formen el que s'anomena "l'eskeleton" sobre el que es fa la comparació global.

Exemple de sub-seqüències de MUM's:
agctcgat**GGGCTTTAGACTCTCGATA**ggcgcagag**GCTCGCTAGAATCGCTAGATC**ac
agacctaa**GGGCTTTAGACTCTCGATA**agtctatcc**GCTCGCTAGAATCGCTAGATC**ta

Un SuperMUM o SMUM es una agrupació de MUMs consecutius que ha de verificar els seguents requisits:

- L'ordre dels MUMs que formen el SMUM al primer genoma, ha de ser el mateix que al segon genoma.
- L'espai (gap) entre un MUM i el següent no pot ser més gran que la longitud d'aquests sumats multiplicats pel multiplicador de gap.
- Un SMUM no pot estar inclòs dins d'un altre SMUM.
- Els MUMS que estan inclosos dins d'un SMUM queden absorbits per aquest.

De vegades pot passar que els MUM's siguin inversos, això vol dir que una base de la subseqüencia comuna a la posició *n* a un genoma es igual a la base de la subsecuencia a l'altre genoma a la posició *n* pero començant pel final d'aquesta.



[Figura 1]: Interficie gràfica via web al servidor per a comparació de genomes del IBB-UAB que mostra MUMs i SMUMs directes i inversos així com els gens dels genomes comparats. A la figura es mostra la comparació entre els Bacteris Yersinia_pestis_Pestoides_F i Yersinia_pestis_Antiqua.

La recerca clàssica de MUM's es realitza mitjançant la construcció de Suffix-tree's, representant tots els sufixos de les seqüències en un arbre que comparteix els seus prefixos comuns (prefixos dels sufixos). El procés de construcció de l'arbre té un alt cost en temps de procés. La variant utilitzada pel algorisme MUMOL permet buscar els MUM's de diverses seqüencies, construint l'arbre per una seqüencia i recorrent-lo després per comparar amb la resta. Aquestes característiques suposen un gran estalvi de temps i espai en comparació amb altres programes existents, ja que el temps es lineal respecte a la longitud de tots els genomes i l'espai utilitzat es lineal respecte a la longitud del genoma més petit [3].

Existeixen aplicacions que fan ús del MUMOL com el MALGEN (2003) [4] eina web amb l'acrònim de Multiple ALignment of GENomes de bacteris. És una eina web per la exploració de relacions entre seqüències de ADN. Una altre aplicació seria el M-GCAT (2006) [5] acrònim de Multiple Genome Comparison and Alignment Tool i és una eina interactiva d'escriptori per la comparació de diversos genomes simultàniament.

Els genomes d'eucariota estan disponibles al servidor públic del NCBI (National Center for Biotechnology Information)[1]. En aquest servidor també es disposa dels gens de cada genoma amb les seves posicions al genoma. La actualització del servidor amb nous genomes seqüenciats, l'ampliació de les regions seqüenciades, i el descobriment de nous gens per a un genoma, ve controlat per un sistema de versions.

3. Estudi de viabilitat del projecte

Existeixen diferents eines desenvolupades al IBB per la comparació de genomes. Aquestes consisteixen un procés tot automatitzat per la comparació de genomes de bacteris i arqueobacteris. Per eucariotes el procés s'ha preparat manualment per provar la comparació de l'homo sapiens amb el macaca mulata amb l'inteficie web mummy. Fent us de shell scripts, programes en C++ i Java es va arribar a aconseguir la comparació en aproximadament, un dia i mig. Això presenta diversos inconvenients, principalment la dificultat a l'hora de llançar una comparació. Suposem que volguéssim comparar 10 genomes d'eucariota tots amb tots, aproximadament trigaríem 3 mesos, pel que es necessita accelerar el procés per obtenir els resultats a temps.

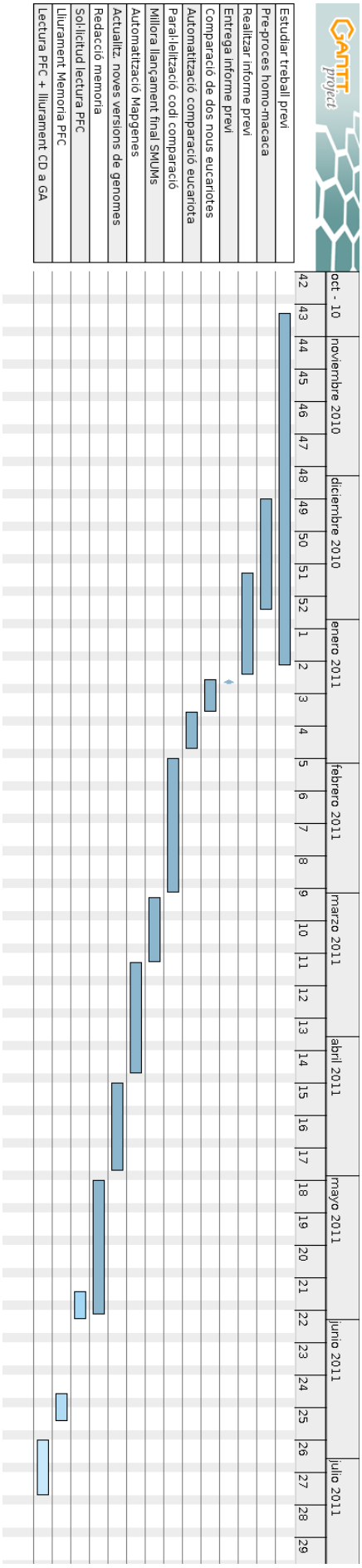
Donat que un dels objectius es comparar tots els genomes d'eucariota disponibles amb tots, es necessita, primer poder disposar dels genomes d'eucariota disponibles al servidor públic del NCBI, automatitzar el proces de comparació el màxim possible, i paral·lelitzar la comparació de genomes aprofitant el servidor del que disposem al IBB.

Aquest nou servidor de l'IBB destinat exclusivament a les comparacions de genomes: http://platypus.uab.es [7], disposa de 24 processadors a 2,67 Ghz (4 x Intel Xeon X5650), és a dir, suporta 24 processos alhora o bé 48 threads al mateix temps. També disposa de 63 Gb de RAM per poder realitzar comparacions amb els genomes més pesants.

Suposant una paral·lelització ideal, la comparació hauria de trigar 24 vegades menys, tot i que de entrada ja es veu que donada la natura del problema de la comparació, no es podrà repartir tant la tasca, bé per falta de memòria, o bé per la complexitat de repartir feina entre els processadors i mantenir el mateix funcionament del programa.

Respecte l'automatització del syncgenes, programa encarregat de descarregar la informació i posició dels gens, el problema rau en la interacció amb l'ftp del NCBI i control de versions dels genomes.

4. Planificació temporal del treball



Nombre	Fecha de ini...	Fecha de fin
Estudiar treball previ	27/10/10	11/01/11
Pre-proces homo-macaca	6/12/10	30/12/10
Realitzar informe previ	22/12/10	13/01/11
Entrega informe previ	14/01/11	15/01/11
Comparació de dos nous eucariotes	14/01/11	21/01/11
Automatització comparació eucariota	21/01/11	29/01/11
Paral·lelització codi comparació	31/01/11	1/03/11
Millora llançament final SMUMs	2/03/11	16/03/11
Automatització Mapgenes	16/03/11	9/04/11
Actualitz. noves versions de genomes	11/04/11	30/04/11
Redacció memoria	2/05/11	31/05/11
Sol·licitud lectura PFC	26/05/11	1/06/11
Lliurament Memoria PFC	17/06/11	23/06/11
Lectura PFC + lliurament CD a GA	27/06/11	9/07/11

5. Altres comentaris

Un cop vist per sobre tot el que avarca el projecte, podem dir que és un treball ambiciós, actual, interessant i el més important, viable.

Personalment, em vaig decantar per aquest projecte pel repte personal que presentava. La genòmica sempre m'ha cridat l'atenció i amb aquest projecte tindré l'oportunitat d'aprofundir aquests coneixements i col·laborar amb l'IBB en investigacions punteres que estan a l'ordre del dia. A més, el problema a solucionar és d'una dificultat considerable, on s'han d'aplicar totes les eines d'un enginyer, si volem obtenir els resultats esperats.

El fet de treballar amb GNU/Linux, shell scripting i c++ em va donar l'empenta final, doncs són sistemes potents, amb un grau de llibertat alt, molt estesos i que m'aniran molt bé per actualitzar coneixements.

A finalitzar el projecte, l'esforç i treball invertits seran d'utilitat per tothom que vulgui estudiar i obtenir resultats de les comparacions de genomes eucariotes.

6. Referències

[1] **Web oficial NCBI**
http://www.ncbi.nlm.nih.gov/

[2] **Definició de genoma a la wikipedia**
http://es.wikipedia.org/wiki/Genoma

[3] **Efficient Space and Time multicomparision of Genomes**, Mario Huerta, Xavier Messeguer, Technical report LSI-02-64-R. Llenguatges i Sistemes Informatics, Universitat Politècnica de Catalunya (2002).

[4] **Identification of patterns in biological sequences at the ALGGEN server: PROMO and MALGEN**, Domènec Farré, Mario Huerta, Romà Roset, José E. Adsuara, Llorenç Roselló, M. Mar Albà, and Xavier Messeguer, Nucleic Acids Research. 2003 31: 3651-3653 (2003).

[5] **M-GCAT: Interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species.** T. Treangen and X. Messeguer. BMC Bioinformatics 2006, 7:433.(2006)

[6] **Suffix Tree Construction with slide nodes,** Mario Huerta. technical report LSI-02-63-R Dep. Llenguatge i Sistemes Informàtics, Universitat Politècnica de Catalunya (2002).

[7] http://platypus.uab.cat/ , Web server for the all-known-genomes comparison by web. Server supported by the Institute of Biotechnology and Biomedicine of the Autonomous University of Barcelona (IBB-UAB).