Comparació de Genomes de Cèl·lules Eucariota

Héctor González Gómez

Les similituds que es poden trobar entre els genomes de dues espècies ens permet obtenir molta informació de l'evolució d'aquests. Aquesta informació afavoreix el descobriment de gens que conserven les mateixes funcionalitats entre diferents espècies. Aquest camp té importants aplicacions médiques que ajuden a la investigació de diferent tipus d'enfermetats que s'hereden i alhora ens permet entendre els processos evolutius que han propiciat la diversitat d'espècies a l'actualitat.

El treball que s'ha desenvolupat té com a objectiu donar un arbre de l'evolució de les espècies dels genomes d'eucariota de forma que cada vegada que s'afegeixi un nou genoma a les bases de dades de genomes d'eucariota, hi hagi una actualització de l'arbre per tal de permetre l'estudi d'aquest a l'arbre evolutiu i mostrar d'on prové aquest genoma a més d'actualitzar els genomes que ja hi siguin a l'arbre evolutiu quan s'actualitzin en aquestes bases de dades. A més d'això, es verificarà les dades de tot el procés fins a la generació d'aquest arbre evolutiu d'espècies de genomes d'eucarita per tal d'assegurar la correctesa de l'arbre i que no hi ha errors en l'evolució dels genomes que conté.

Espècies, genomes, evolució, gens, cromosomes, enfermetats, mums, smums, arbre filogenètic, servidor, alarma, comparació, genètica, biologia.

The similities that can be found between the genomes of two species allow us to obtain a lot of information of the evolution of these. This information favours the finding of genes that preserve the same functionalities between different species. This field has important medical applications that help to the research of different type of diseases that heritate and at the same time allow us understand the evolutionary processes that have done the diversity of species to the actuality.

The work that has been developed has as a objective give an evolution tree of eucaryote genomes so that every time that it adds a new genome from the eukaryote genomes databases, there will be an update of the evolution tree in order to allow the study of this to the evolutionary tree and show from where it comes besides updating the genomes that already exists on the evolutionary tree when updated on these databases. In addition, it will verified the data of all processes until the generation of this evolutionary tree.

The work that has been developed has as objectives improve the control of parallel processes, improve the error control on the comparison process between genomes, improve the error repairment occurred on these comparisons, optimize the genome comparison process and, configure the server in order to automate the updating and addition of new genomes process constantly and finally, the constant update of the result tree that shows the evolution of every gene.

Species, genome evolution , genes , chromosomes , diseases , Mums , smums , phylogenetic tree , server , alarm , comparison , genetics , biology

1 INTRODUCCIÓ

1.1 Motivació

EL fet de poder aportar el meu granet de sorra a un camp com és la bioinformàtica ja va ser suficient motivació per agafar aquest projecte, però quan vaig assaventarme que la part técnica era principalment C++ i ShellScript a backend vaig decidir-me sense dubtar ja que podia posar a prova els conceptes de la carrera i reforçarlos i, alhora, millorar el meu rendiment com a programador de backend, que és allò en que m'estic desenvolupant professionalment.

1.2 Organització del document

Al document hi trobarem primerament un apartat d'introducció a l'estat de l'art amb un conjunt de coneceptes teòrics necessaris per entendre el perquè del programa i totes les fases que té aquest.

Darrerament, s'especificarà els objectius d'aquest treball i els diferents resultats als quals s'ha arribat per tal d'aconseguir assolir els objectius proposats, explicant en detall el motiu d'aquestes decisions i després dels quals, es farà un informe tècnic de l'apartat de comparació de genomes.

Finalment, hi haurà un apartat de conclusions i un petit agraïment a tots els implicats amb el projecte.

1.3 Estat de l'art

Abans que tot, hem de definir un conjunt de fonaments teòrics necessaris per entendre bé el projecte i les seves parts més a fons, els quals són definits seguidament.

Una *cèl·lula* és l'element més petit que es pot conside-

rar viu. Depenent de les característiques de la cèl·lula d'un organisme, els podem classificar. Per posar un exemple, si la cèl·lula disposa d'un nucli al medi intracel·lular, podem determinar que la cèl·lula pertany a un organisme de eucariota, com per exemple, les cèl·lules del ésser humà.

Les principals molècules que podem trobar a una cèl·lula són els glúcids, lípids, proteïnes i àcids nucleics:

- Les *proteïnes* són una successió d'aminoàcids constituïda sobre un alfabet de 20 símbols on cada símbol és un dels aminoàcids possibles. Les proteïnes regulen el metabolisme de la cèl·lula i tenen funcions estructurals, motrius, catalítiques, o fins i tot sensorials. Depenent del conjunt de proteïnes que regulin el metabolisme de la cèl·lula, es determinarà l'especialització de la cèl·lula i el seu funcionament.
- Els *àcids nucleics*, més coneguts per ADN, àcid desoxiribonucleic, i ARN, àcid ribonucleic, composen una successió de nucleòtids que es poden distingir per la base nitrogenada que els constitueix. Els quatre tipus de bases nitrogenades a l'ADN són A (adenina), T (timina) , G (guanina) i C (citosina). Al ARN podem trobar les anteriors, però substituint la T (timina) per U (uracil).

La funcionalitat del ADN és transmetre la seqüència de nucleòtids que la composa i la del ARN comporta funcions variades com catalítiques o de regulació de síntesi proteica.

Tota la informació que es codifica a la seqüència de les bases nitrogenades conté les possibles proteïnes que es generaran i els ARN que regularan el funcionament de la cèl·lula, entre d'altres. Aquestes subseqüències del ADN les anomenem gens.

L'ARNm dins de la molècula de ARN fa de missatger, transportant la informació d'un gen al Ribosoma (orgànul de la cèl·lula encarregat de generar les noves proteïnes).

Resumint, el codi original del ADN està format per quatre símbols (A,T,G,C) i es tradueix a una proteïna formada per una seqüència codificada amb 20 símbols. Així, les agrupacions de tres nucleòtids ens permeten codificar un dels vint aminoàcids que sintetitzarà amb certa redundància. Aquesta redundància es pot interpretar com rastres evolutius o tolerància a errors.

El fet que un gen s'expressi o no, és a dir, que generi la proteïna o ARN que codifica, depèn de diversos factors, com pot ser l'existència de molècules al medi cel·lular que loguegin la seva lectura. La aparició d'aquestes molècules ve donada per l'expressió d'altres gens, així que ens podem fer una idea de les complexes dinàmiques que regulen el control de una cèl·lula.

Un *genoma* és la totalitat d'informació genètica que conté un organisme codificat a les seves molècules d'ADN. Podem adonar-nos que l'estudi dels gens del genoma d'una espècie pot aportar-nos multitud d'informació referent al metabolisme de la cèl·lula i per tant informació del organisme en qüestió. Fins i tot, les parts no codificats del genoma, com els gens que comentàvem abans que no s'expressen, es creu que participen activament a la dinàmica cel·lular.

La *genètica* és l'estudi de la natura, organització, funció, expressió, transmissió i evolució de la informació genètica codificada als organismes. Podem distingir les següents àrees:

- La *genètica clàssica*, o estudi de la transmissió i localització de gens als cromosomes.
- La *genètica molecular*, o estudi de l'estructura i control d'expressions genètiques.
- La *genètica evolutiva*, o estudi de processos evolutius d'espècies.
- La *genòmica*, anàlisis i interpretació dels genomes.

El treball actual s'emmarca dins de la genòmica evolutiva, una de les àrees més avantguardistes de la biologia, i estudia les relacions entre els genomes de diferents espècies amb l'objectiu d'entendre el funcionament dels organismes i obtenir la informació que proporciona les diferències evolutives a partir de la selecció natural.

L'*evolució biològica* és el conjunt de transformacions o canvis a través del temps que origina la diversitat de formes de vida que poblen avui la terra a partir de un avantpassat comú. Els diferents processos que afecten a aquestes transformacions són:

- *Selecció natural*: Adaptació al medi i al entorn.
- *Deriva genètica*: Dinàmiques dins d'una població.
- *Mutació*: Variacions segueixen mecanismes i regulacions que són desconegudes dins del codi genètic.
- *Flux genètic*: Migració entre poblacions.

Considerant aquests conceptes, la comparació genòmica de les espècies ens servirà per veure la proximitat evolutiva, la detecció de característiques que han perdurat a les espècies i veure els gens que han perdurat.

Aquesta cerca de similituds entre regions de diferents genomes aporta molta informació sobre les relacions de les espècies i com les propietats d'un gen són assignades a un altre gen.

Les espècies poden estar classificades en un arbre *filogenètic*, això vol dir que les espècies es classifiquen científicament segons les seves relacions de proximitat amb altres espècies. D'aquesta manera estem reconstruint la història de la diversificació o filogènesis de les espècies des de l'origen de la vida fins avui dia.

Veiem que els tres grans dominis a la classificació de cèl·lules (i segons aquestes, la classificació d'espècies també) són els *arqueobacteris*, *bacteris* i *eucariotes*. Els arqueobacteris i bacteris pertanyen al grup de les cèl·lules sense nucli diferenciat, les procariotes, que tenen l'ADN dispers al citoplasma

Les *eucariotes* són més complexes i les seves cèl·lules contenen un nucli cel·lular que emmagatzema l'ADN. A

més, tenen orgànuls com els mitocondris, que regulen la síntesis del combustible del metabolisme cel·lular (l'ATP), que contenen ADN propi i es creu que provenen de simbiosis anteriors amb altres procariotes.

La *bioinformàtica* és un camp d'estudi multi-disciplinari que necessita de l'aplicació de la informàtica, les matemàtiques, la estadística, la intel·ligència artificial, la química i la bioquímica. Els volums d'informació que pot tenir un genoma eucariota, implica la necessitat d'automatitzar els processos cerca de estructures i patrons als genomes i de comparacions de seqüències. Al nostre treball les estructures de dades que traiem de les comparacions de genomes i que haurem de tractar són els MUMs.

Un *MUM* o maximal unique matching [4], és la subseqüència única màxima que coincideix a la comparació de dos genomes. Dit d'altre forma, es un fragment de codi genètic on les bases nitrogenades coincideixen correlativament amb les d'un altre genoma, aquesta coincidència es la més llarga i no es repeteix.

El conjunt de MUMs formen l'esquelet (*skeleton*) de la comparació de genomes, la seva estructura bàsica que té un significat biològic. És lògic dons, que a les comparacions de genomes de espècies evolutivament properes, trobem gran quantitat de MUMs, o que els que hi hagin tinguin una mida gran.

Si trobem MUMs de gran llargària vol dir que els fragments d'ADN han perdurat intactes a les dos espècies i que la funció dels gens que codifiquen haurien de ser les mateixes (tot i que hem explicat abans que la expressió d'un gen depèn de molts altres factors). També hem dit que els MUMs han de ser únics, això és per descartar fragments del codi genètic que es repeteixen i no ens aporten cap informació evolutiva.

Els gens poden estar codificats de esquerra a dreta o de dreta a esquerra respecte la seqüència del genoma. Per tant, podem trobar fragments de codi, que durant el procés evolutiu inverteixin el seu ordre d'una espècie a una altre i mantenint la mateixa funcionalitat.

Els MUMs contemplen aquest fet, i a la cerca de MUMs els classifiquem en MUMs directes si han mantingut l'ordre o en MUMs inversos si l'han invertit.

Per a genomes grans com els de les eucariotes, és imprescindible agrupar aquests MUMs obtinguts si volem poder tractar la informació amb facilitat. La idea és que podríem trobar seqüències coincidents encara més llargues de no ser per poques bases que s'han alterat i que impedeixin la formació de MUMs més grans.

Amb el concepte de *SMUM*, tenim dos avantatges, un és poder fer viable la comparació de genomes enormes com els d'eucariotes, i l'altre és aplicar una mica de tolerància a la cerca de seqüències coincidents, que poden haver variat mínimament. Per crear SMUMs a partir de MUMs s'han de complir una sèrie de condicions:

- 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.

Podem detectar que, per definició, un SMUM pot originar superposicions, mentre que els MUMs mai es superposen

El programa de forma general consta de dos parts clarament diferenciables, les quals són el front-end i el back-end.

El *back-end* s'encarrega primer que tot, de la descàrrega dels diferents genomes i mapgenes del servidor NCBI [1] (el qual ens proporciona dades dels diferents genomes), un pretractament de les dades per a preparar aquests genomes per al procés de comparació. Posteriorment, es fa la comparació dels genomes, procés que es subdivideix en comparacions més petites, que són les comparacions del diferents cromosomes d'aquests genomes (en aquest procés es on es generen els MUMS dels diferents cromosomes). Finalment, es fa un post-tractament que unifica tots els MUMS, genera els SMUMS i amb això, genera la matriu de similitud.

La *matriu de similitud* és l'encarregada de guardar les diferents similituds entre cada parella de genomes per tal de generar l'arbre filogenètic.

El *front-end* és un conjunt d'interfícies web la qual s'encarrega de mostrar les dades processades pel back-end mostrant arbres filogenètics de arquaeas, bactèries i eucariotes comparant els genomes entre sí.

La comparació dels genomes és un gran problema degut al fet que els genomes de cèl·lula eucariota són molts, però sobretot són molt grans i, per tant, el procés requerit per a fer les comparacions entre els diferents genomes és lent i costós, amb la qual cosa, el tractament d'aquests genomes soposa una dificultat afegida i una metodologia de treball diferent.

El treball realitzat pretén generar un arbre capaç d'explicar l'evolució entre els diferents genomes de cèl·lules d'eucariota. Per dur a terme aquesta tasca, primer que tot es necessita d'un procés de descàrrega dels diferents genomes que hi ha a les bases de dades; seguidament, s'ha d'acomodar aquests genomes per tal de facilitar els següents processos de comparació de genomes; han de ser comparats tots els genomes entre ells per tal de trobar les similituds entre cada parella de genomes; i finalment, generar l'arbre a partir de les similituds trobades.

En aquest treball, la part d'obtenció de les similituds no és instantània, sinó que estem parlant de processos de comparació de genomes que triguen a generarse dies. Per altra banda, aquests resultats requereixen d'una gran quantitat de memòria no volàtil per tal d'emmagatzemar els genomes i els resultats de les comparacions i, alhora, una gran quantitat de memòria volàtil per poder paral·lelitzar les comparacions entre genomes.

Tot això fa que siguin necessaris algoritmes robustos i

ràpids que tinguin en compte tots els cassos possibles.

1.4 Objectius

Després de veure els problemes citats anteriorment, volem que el procés de comparació entre genomes sigui un procés automàtic, més ràpid i més estable. Per altra banda, es pretén aconseguir veracitat a les dades, tant d'entrada com de sortida. Per tant, els objectius a assolir són:

- Procés de comparació de genomes en paral·lel
- Alarma de control de procés de detecció de problemes a les comparacions de genomes
- Fòrmula de detecció de memòria disponible per al control del llançament de comparacions de genomes en paral·lel
- Reparació d'errors captats al procés de comparació de genomes en paral·lel
- Optimització de l'obtenció del procés d'obtenció de la Matriu de Similitut de les comparacions entre genomes
- MUM, SMUM i Matriu de Similitut resultants de les comparacions de genomes a temps real
- Control d'errors de genomes abans de l'inici de la comparació de genomes
- Control d'errors de MUM, SMUM i Matriu de Similitut a la sortida de la comparació de genomes
- Llençament de la fase de creació de l'arbre filogenètic a partir de la matriu de similitut
- Automatització del procés d'adhició i actualització de similituds entre genomes
- Comparacións només dels nous genomes descarregats amb la resta de genomes
- Actualització del post-procés de les descàrregues de nous genomes per a l'acceptació de cromosomes no numèrics
- Priorització del post-procés de les descàrregues de nous genomes per a prioritzar la comparació de genomes nous

2 RESULTATS I DISCUSIONS

2.1 Procés de comparació de genomes en paral·lel

A l'hora de fer la comparació entre dos genomes, el procés que segueix és el de dividir el genoma més petit en els diferents cromosomes que el formen i fer un arbre de cadascun dels cromosomes per tal de comparar aquests amb el genoma no particionat.

Aquest procés està paral·lelitzat per tal d'aconseguir fer tantes comparacions alhora com sigués possible, dins del marge que dona la memòria RAM.

Els problemes que sorgien de la paral·lelització van ser el càlcul de l'espai lliure de la memòria RAM, el qual es va optar per utilitzar la fòrmula i l'alarma que s'encarregaba de mirar que un procés no estigués penjat. Seguidament es parla d'aquests problemes i la solució que es va dur a terme.

2.1.1 Alarma de control de detecció de problemes a les comparacions de genomes

Inicialment, l'alarma que s'utilitzaba al procés de comparació de genomes, estaba programada per saltar una vegada i després de 5 dies, això feia que si havia swap de memòria el primer dia, trigava 4 dies a matar el procés.

Aquesta part ha sigut modificada per una alarma que es reprograma cada 5 hores i s'encarrega de mirar si la memòria té lliure menys d'un llindar. Si aquest llindar és superat, començarà a matar tots els processos de comparació per tal d'alliberar la memòria abans de que fagi swap. En cas contrari, l'alarma es reprogramarà per tornar a mirar al cap d'altres 5 hores.

2.1.2 Fòrmula de detecció de memòria disponible per al control del llançament de comparacions de genomes en paral·lel

A la comparació de genomes, s'ha millorat la fórmula que s'encarregaba de mostrar la quantitat de memòria lliure de RAM. Aquesta millora ha estat incorporada degut al fet de que el càlcul de l'espai lliure a RAM no és tant exacte amb la fórmula:

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

Aquesta fórmula inclou el camp *cached*, però aquest camp conté dues parts de la memòria que administra el sistema Linux, una part que realment és buida, i un altre part que mai es buida, però que Linux controla de forma automàtica i la qual no es pot tenir en compte.

La formula va ser modificaca per la següent

Memòria ocupada_real = (MemTotal-MemFree-Inactive-Buffers)

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

Memòria lliure = MemTotal – Memòria ocupada i no disponible

Aquesta nova fórmula és més restrictiva i només té en compte la part de memòria lliure de *cached*, que és la part *inactive*.

Això és necessari per tal de poder exprimir al màxim els recursos de la màquina per obtenir l'arbre filogenètic abans, però sense arribar al swap de memòria.

2.1.3 Reparació d'errors captats al procés de comparació de genomes en paral·lel

Degut a que es mataven processos de comparació en paral·lel, es requeria d'un mecanisme per tal de recuperar aquestes comparacions no fetes i les executés un altre cop, però com no es pot saber si la raó de per qué es para una

execució, no podem tornar a executar en paral·lel aquelles comparacions tallades, sinó que les hem d'executar de forma seqüencial per tal de certificar que es fan.

Per donar més estabilitat al sistema, es va configurar aquesta comparació per tal de fer 3 intents de comparació en seqüencial, ja que aquestes execucions també poden ser tallades.

Aquesta millora és necessària per tal d'assegurar que tots els genomes s'han comparat completament i que no hi ha cap cromosoma de cap genoma sense comparar. En cas contrari, es hi hauria errors a l'arbre filogenètic resultant de totes les comparacions.

2.2 Optimització del procés d'obtenció de la Matriu de Similitut de les comparacions entre genomes

La sortida final del procés de comparació dels genomes és la matriu de similitut, que ens mostra amb un valor quin és el grau de similitut entre dos genomes. Per a fer aquests càlculs, s'utilitzaven un conjunt d'arxius auxiliars (log_inv i log_dir) que endarrerien la comparació de genomes.

Aquest procés ha sigut modificat, ja que aquests arxius eren redundants i es podien utilitzar altres arxius o variables que tenien la mateixa informació que aquests per tal d'estalviar temps de computació perdut en generació d'arxius i lectura d'aquests.

Aquesta part elimina una part del procés de comparació de genomes per tal de reduir el temps de comparació dels genomes.

2.3 MUM, SMUM i Matriu de Similitut resultants de les comparacions de genomes a temps real

Els arxius de sortida(MUMs, SMUMs i Matriu de Similitut) del procés de comparació entre genomes només els donava una vegada finalitzat el procés de comparació de tots els genomes.

Aquesta metòdica ha sigut modificada per tal de poder tenir les dades d'una comparació immediatament després de finalitzar aquesta comparació.

Això és una part molt important, ja que d'aquesta manera podem aconseguir disposar d'un arbre filogenètic abans de finalitzar tot el procés de comparació de genomes, amb la qual cosa aconseguim unes dades més actualitzades en tot moment.

2.4 Control d'errors de genomes abans de l'inici de la comparació de genomes

Els genomes que es comparaven estaven especificats amb un identificador o ID únic i la connexió genoma-ID estaba especificat en un arxiu concret (genes.txt), per tant, abans de llençar el procés de comparació de genomes, s'havia de fer la verificació de que els genomes eren els que s'especificava en aquest arxiu i que els genomes eren correctes. Per tal de fer això, es tenia que verificar que el genoma amb un ID concret fós el mateix que el genoma que corresponia amb aquest ID a la base de dades del ncbi [1] (p.e. el genoma amb ID 4, corresponent al genoma conegut com a canis familiaris fós realment el canis familiaris que està a la base de dades del ncbi [1]).

Aquesta part es va fer de forma manual mirant cadascún dels genomes al nostre servidor local [3] i comparant aquests genomes amb els del servidor online del ncbi [1].

Això era necessari per assegurar la coherència de l'arbre filogenètic, ja que si no eren correctes aquestes correspondències entre ID i genoma i/o no eren els genomes correctes els genomes, l'arbre donaria informació errònia.

2.5 Control d'errors de MUM, SMUM i Matriu de Similitut a la sortida de la comparació de genomes

Es necessitava verificar que les dades d'entrada i sortida eren correctes, per la qual cosa s'ha creat un conjunt d'eines per tal de controlar que aquests arxius són els valors esperats tant abans com després de la comparació.

L'eina que comprova les dades de sortida mira els MUMs i SMUMs de sortida i els compara amb el tamany dels genomes per tal de verificar que estàn comparats els dos genomes sencers. A més, comprova que els SMUMs no estiguin sol·lapats (uns SMUMs dins d'uns altres).

Aquest procés és necessari per tal d'evaluar el correcte funcionament de les diferents dades de sortida, ja que no podem permetre dret a error, ja que la informació resultant (l'arbre evolutiu) ha de mostrar correctament d'on prové cada espècie.

2.6 Llençament de la fase de creació de l'arbre filogenètic a partir de la matriu de similitut

Per a poder llençar la fase de generació de l'arbre filogenètic després de cada comparació era necessari modificar el funcionament de les sortides de la comparació per tal d'adaptarles a aquesta fase de generació de l'arbre, ja que aquesta part genera l'arbre utilitzant la matriu, la qual fins que no estiguin comparats tots els genomes no estarà completa.

Això es va aconseguir aconseguir cambiant la idea de l'arxiu genes.txt fent que aquest sigués un arxiu auxiliar amb els genomes que estàn comparats i són a la matriu de similitud, mentre que un nou arxiu anomenat genes_completo.txt és el que compleix la funció de l'antic genes.txt. Això, juntament amb l'apartat 2.3, aconsegueix que es pugui tenir l'arbre filogenètic actualitzat per tal de donar suport de forma més constant.

2.7 Automatització del procés d'adhició i actualització de similituds entre genomes

S'ha programat l'execució bimensual del programa de descàrrega de genomes nous o genomes modificats i actualització de les comparacions d'aquests nous genomes o genomes modificats.

Aquesta programació s'ha dut a terme utilitzant la funció de Linux Kron, la qual permet la programació cronològica de qualsevol programa.

Això és necessari per tal d'estalbiar que un responsable estigui cada cert temps llençant el programa reiterativament, a més, comporta que hi hagi una actualització constant de l'arbre filogenètic cada cert temps, per tal de tenir una informació actualitzada i fiable.

2.8 Comparacions només dels nous genomes descarregats amb la resta de genomes

Es va dur a terme un conjunt de modificacions a les comparacions en paral·lel i seqüencials per tal de reduir el temps d'execució de les comparacions.

Quan un nou genoma és descarregat, aquest ha de ser comparat amb tots els altres genomes que ja hi són al servidor local [3], però es pot reduir la quantitat de comparacions comparant un genoma amb els que només tenen un ID menor al seu. Per deixar més clar aquesta part, si són descarregats 3 nous genomes i se'ls assigna els identificadors 6, 7 i 8, aquests genomes seran comparats amb els genomes 1-5, però abans, el genoma 6 també es comparava amb els genomes 7 i 8 i, quan arriba el torn de comparar el genoma 7, compara amb els seus anteriors, incloent la comparació amb el genoma 6, que ja va ser comparat anteriorment.

Això és necessari per tal de reduir el temps que triga tot el procés d'adició i actualització de genomes, ja que cada comparació pot trigar 1 dia, amb la qual cosa, es redueix molt el temps d'execució.

2.9 Actualització del post-procés de les descàrregues de nous genomes per a l'acceptació de cromosomes no numèrics

Havia de ser actualitzat el procés de preparació dels nous genomes i dels genomes actualitzats per tal de poder acceptar els cromosomes no numèrics (cromosomes X, Y,etc.).

Es van afegir els casos de cromosomes no numèrics al procés de preparació dels nous genomes per tal de poder comparar els genomes amb aquests cromosomes.

Aquesta part és necessària per tal d'assegurar que quan estem generant el genoma complet unint tots els cromosomes, estiguin inclosos aquests cromosomes especials i amb l'ordre que toca.

2.10 Priorització de les comparacions de nous genomes respecte a genomes modificats

Era necessari modificar el funcionament del procés de preparació dels genomes descarregats per tal de poder llençar els genomes nous avans que els genomes actualitzats.

Es va optar per modificar l'ordre de la llista que conté els cromosomes que s'han de comparar per tal de que prioritzés els genomes nous, posant-los al davant de la llista.

Aquesta part era necessària pel fet de que és prioritari tenir una mesura de les similituds de nous genomes que no pas corretgir la similitut d'un cromosoma ja comparat amb una versió més antiga d'aquest, ja que les actualitzacions, encara que aporten informació, no és tan rellevant com les que aporta un nou genoma a l'arbre filogenètic. Això aporta finalment que l'arbre filogenètic tingui incorporats els nous genomes avans de tenir incorporats els genomes actualitzats.

3 METODOLOGÍA

3.1 Estructura del sistema de carpetes utilitzada per tot el programa de descàrrega i comparació de genomes

Seguidament es mostra el sistema de carpetes de l'actualitzador de les comparacions de genome:

RobotEucariota/

Aquesta carpeta té emmagatzemada tota la part del programa necessària per a tot el procés d'adquisició, preparació, comparació dels genomes i control d'errors de tot el procés. Dins hi podem trobar:

RobotEucariota.cc: Programa d'automatització bimensual.

downloadRobot: Carpeta que conté el programa de descàrrega de genomes eucariota i tot el necessari per a l'execució.

sync_genes: Carpeta que conté el programa de mapatge de gens sobre el genoma i tot el necessari per l'execució.

sinRenombrar: Aquesta carpeta conté diversos programes del procés post-descàrrega.

actualiza_mums: Aquí disposem de tots els programes on es generen les comparacions.

downloadRobot/

Aquesta carpeta té emmagatzemada tota la part del programa necessària per a la descàrrega. Dins hi podem trobar:

downloadRobot.jar: Programa de descàrrega de genomes d'eucariota.

genesdescargados.txt: Fitxer amb una llista de noms dels genomes d'eucariota que ja tenim descarregats al nostre servidor.

no.txt: Fitxer amb noms de carpetes conegudes del FTP que el programa ha d'ignorar per no contenir genomes.

nuevosMetazoa.txt: Fitxer que genera el programa amb una llista dels genomes que ha descarregat durant l'execució.

output.txt: Guardem la sortida del programa per si hem de consultar alguna cosa.

bin: Fitxers de classes.

genome: En aquesta carpeta es descarreguen els nous genomes. Per cada genoma es crea una carpeta amb el nom del genoma. És una carpeta d'ús temporal.

sync_genes/

Aquesta carpeta té emmagatzemada una part de la preparació dels genomes nous i/o actualitzats per poder llençar les comparacions entre genomes (mapatge de

gens). Dins hi podem trobar:

getgenes.jar: Programa de mapatge de gens al genoma.

genes.txt: Aquest és un fitxer necessari ja que conté una llista amb tots els genomes dels que el programa ha de descarregar els gens del mapatge.

assemblies: En aquesta carpeta hem de crear una carpeta per cada genoma que vulguem que descarregui els gens. Aquesta carpeta ha de tenir el nom del genoma i ha de contenir el fitxer accession.txt del genoma.

bin: Fitxers de classes i fitxer de configuració de carpetes: local del programa i la remota del servidor FTP del NCBI [2].

lib: Llibreries que utilitza el programa.

genomes/Eukaryota/mapgenes: Dins d'aquestes carpetes es generen els fitxers de gens <ID>.gen. Després es mouen a la carpeta del genoma per ser renombrats per l'identificador del genoma.

SinRenombrar/

ngenoma: Fitxer que conté un número que representa la quantitat de genomes al servidor. S'incrementa cada cop que descarreguem un genoma i serveix per obtenir l'identificador únic.

procesoDescarga.cc: Programa que gestiona la inserció del genoma al sistema a la carpeta que toca, amb el format i nom que han d'anar els fitxers i carpetes.

renombrar.cc: Programa que s'encarrega de canviar els noms dels genomes descarregats pel número identificador.

unirCromosomas.cc: Programa que genera el fitxer de genoma complet, a base de concatenar en ordre els cromosomes.

procesados: Carpeta que conté tots els genomes processats, es fa una còpia de seguretat per si el procés d'inserció al sistema és erroni.

actualitza_mums/

Aquesta carpeta té emmagatzemada tota la part d'actualització de les similituds i el procés de comparació entre genomes. Dins hi podem trobar:

offsetscromo: Fitxer que conté el desplaçament de cada cromosoma, necessari per fer la concatenació de MUMs de cromosomes correctament. El genera el MUMOL durant la comparació.

logprocessats: Fitxer que guarda els identificadors de les parelles de genomes que s'han comparat correctament.

actualitza_mums.cc: Programa per llançar la comparació dels diferents genomes actualitzats o nous.

lanza_smums.cc: Programa que llença el càlcul de SMUMs de una comparació.

separasmums.sh: Script per eliminar duplicats de SMUMs a l'ultima volta del càlcul de SMUMs.

smumsortV3.sh: Script que transforma el fitxer de SMUMs per poder ser llegit per Mummy.

libera_sem.sh: Script encarregat d'eliminar els semàfors del sistema declarats i que no es fan servir.

mums: Carpeta que allotja tots els MUMs de cromosomes

smums_paso1: Carpeta on es guarden els SMUMs resultants de la primera passada.

smums_paso2: Carpeta on es guarden els SMUMs resultants de la segona passada.

smums_paso3: Carpeta on es guarden els SMUMs resultants de la tercera passada.

smums: Els SMUMs de la tercera passada es guarden en aquesta carpeta fins l'execució dels scripts separasmums i smumsortV3. Els resultats es traslladen a la carpeta del directori superior smumsort/.

smumsort: Carpeta d'ús temporal.

Factors: Carpeta on es guarden els fitxers temporals per la generació del fitxer factors.txt.

long_Mumunix: Aquesta carpeta conté els programes i llibreries per executar el algorisme MUMOL [4] pel càcul de MUMs.

calculoSuperMums: Aquesta carpeta conté els programes i llibreries per executar el programa smum.cc pel càlcul de SMUMs.

concatena: Conté el programa necessari per la concatenació de MUMs de cromosomes

eliminaNoMum: Conté el programa necessari per eliminar els no-MUMs i el programa per eliminar SMUMs duplicats.

animalia/

Aquesta carpeta és on s'emmagatzemen totes les dades descarregades dels genomes i tot el que es requereix a nivell de genomes per poder llençar les comparacions entre genomes. Dins hi podem trobar:

factors.txt: Aquest fitxer conté una línia per cada comparació i apareix el valor de semblança dels dos genomes. Es genera amb el programa uneFactors.cc i s'utilitza per crear l'arbre de distàncies de genomes.

genes.txt: Aquest fitxer conté el nom de tots els genomes que disposem al nostre servidor. L'ordre d'aquest fitxer és importantíssim doncs la posició que ocupa el nom en aquest fitxer és l'identificador únic del genoma. Fa de traductor entre el nom del genoma i la seva ID.

genome: Carpeta que conté tots els genomes dels que disposem al nostre servidor. Per cada genoma conté una carpeta que coma a nom té l'identificador únic del genoma. Dins de la carpeta conté tots els cromosomes, el genoma complet i una carpeta *info* amb altres fitxers del genoma (accession.txt, entre altes).

old: Quan s'actualitza un genoma, guardem el genoma antic per seguretat.

mapedgenes: Aquesta carpeta conté els gens generats pel programa de mapatge de gens. Es guarden com <ID>.gen Aquests fitxers els carrega el front-end Mummy per visualitzar la comparació.

mums: Carpeta que conté els MUMs finals de les comparacions de tots els genomes d'eucariota. Smums: Carpeta que conté els SMUMs finals de les comparacions de tots els genomes d'eucariota.

smumsort: Carpeta que conté els SMUMs de les comparacions amb l'índex d'ordenació. Aquests fitxers són els

que pot carregar l'aplicatiu web Mummy per visualitzar la comparació.

3.2 Modificacions a l'estructura del sistema de carpetes utilitzada per tot el programa de descàrrega i comparació de genomes i als arxius del programa

Els fitxers que han sigut modificats han sigut els llistats seguidament:

RobotEucariota/

Es va modificar l'arxiu *Robot.cc* per tal d'incorporar la millora 2.9 de l'apartat de Resultats i Discussions.

El procediment per a compilar aquest arxiu és utilitzant el compilador g++ de la forma *g++ Robot.cc –o Robot* i, una vegada compilat, per executar aquest arxiu és amb la comanda ./Robot.

actualitza_mums/

D'aquesta carpeta va ser modificat el codi de l'arxiu *actualitza_mums.cc* per tal d'afegir les millores del punts 2.1, 2.2, 2.3, 2.5, 2.7 i 2.10 de l'apartat 2 de Resultats i Discussions.

El procediment per a compilar aquest arxiu és utilitzant la companda *g++ actualitza_mums.cc –o actualitza_mums* i, una vegada compilat, per executar aquest arxiu és *./actualiza_mums* <*carpeta_genomes*> <*ID*>, <*ID*>, ..., <*ID* > on la carpeta dels genomes és l'ubicació de tots els genomes identificats amb el seu ID únic(carpeta animalia/genome actualment) i els IDs són la llista de genomes que han de ser comparats especificats amb el seu ID.

animalia/

En aquesta carpeta es van afegir nous genomes d'eucariota i genomes actualitzats.

També es va modificar el genes.txt per tal que es completés la millora 2.10, canviant el seu significat, ja que ara, en comptes de tenir els genomes dels que disposem al servidor del ncbi[1], només té aquells genomes disponibles que ja han sigut comparats.

Finalment, s'ha afegit el fitxer genes_completo.txt que compleix la funció que tenia anteriorment l'arxiu genes.txt esmentat anteriorment.

SinRenombrar/

Va ser modificat l'arxiu *unirCromosomas.cc* per tal d'introduir al procés de concatenació dels cromosomes en un genoma únic els cromosomes no numèrics.

El procediment per a compilar aquest arxiu és utilitzant el compilador g++ de la forma *g++ unirCromosomas.cc –o unirCromosomas* i, una vegada compilat, per a executar aquest arxiu és *./unirCromosomas* <*carpeta_genoma*> <*ID*> on la carpeta del genoma és l'ubicació del genoma i l'ID és el nou ID únic amb el que serà conegut.

4 Conclusió

Tot i els problemes trobats, hem aconseguit complir els objectius dins el termini plantejat. Hem arribat a optimitzar el procés de comparació de genomes eliminant parts del procés que no feien falta i corretgint els diferents errors que feien que no funcionés del tot bé la comparació en paral·lel. A més, hem aconseguit crear un sistema més robust i hem comprovat els diferents arxius que tenen a veure amb la comparació dels genomes per tal de verificar que tot el procés és correcte.

Amb tot això hem aconseguit actualitzacions més contínues de l'arbre filogenètic per tal de tenir l'evolució de les espècies al día i, hem aconseguit que aquestes dades de l'arbre estiguin verificades per tal de assegurar la correctesa d'aquest arbre.

5 Referències

[1] Web oficial NCBI , (National center for biotechnology information) Repositori mundial de biotecnologia.
http://www.ncbi.nlm.nih.gov/

[2] FTP NCBI , Conté els genomes actualitzats i tota la informació que necessitem per realitzar tot el procés de comparació.
ftp://ftp.ncbi.nih.gov/

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

[4] 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).
http://revolutionresearch.uab.es/downloads/multigenome_comparison.pdf

[5] Web de l'Institut de Biotecnologia i de Biomedicina.
http://ibb.uab.cat/ibb/

[6] Recurs imprescindible per la programació en C++.
http://www.cplusplus.com/

[7] Recurs imprescindible per la programació en Java.
http://download.oracle.com/javase/6/docs/api/

6 Agraïments

A tots aquells que m'han estat recolzant durant tot el projecte, amb especial atenció a Mario, ja que sense ell no hagués estat possible acabar el projecte a temps, als meus pares, ja que ells han sigut una font inesgotable de paciència i al meu amic Jose, el qual ha sigut qui m'ha ajudat

a mantenir una constància i ha sigut una font inesgotable d'idees.