

BIOINFORMÀTICA

COMPARACIÓ DE GENOMES

D'EUCARIOTA

Informe Seguiment I

Héctor González Gómez

1r Tutor: Mario Huerta

2n Tutor: Jordi González

Metodologia, passos a seguir i planificació

La metodologia que s'utilitzarà en aquest projecte serà una metodologia àgil, que es basa en un desenvolupament iteratiu i incremental, ja que al treballar amb Big Data es requereix d'una metodologia amb la qual es pugui modificar els diferents objectius amb el temps.

La planificació inicial es basarà en els objectius esmentats anteriorment. No obstant això, poden haver modificacions degut als possibles entrebancs que puguin sorgir en el procés.

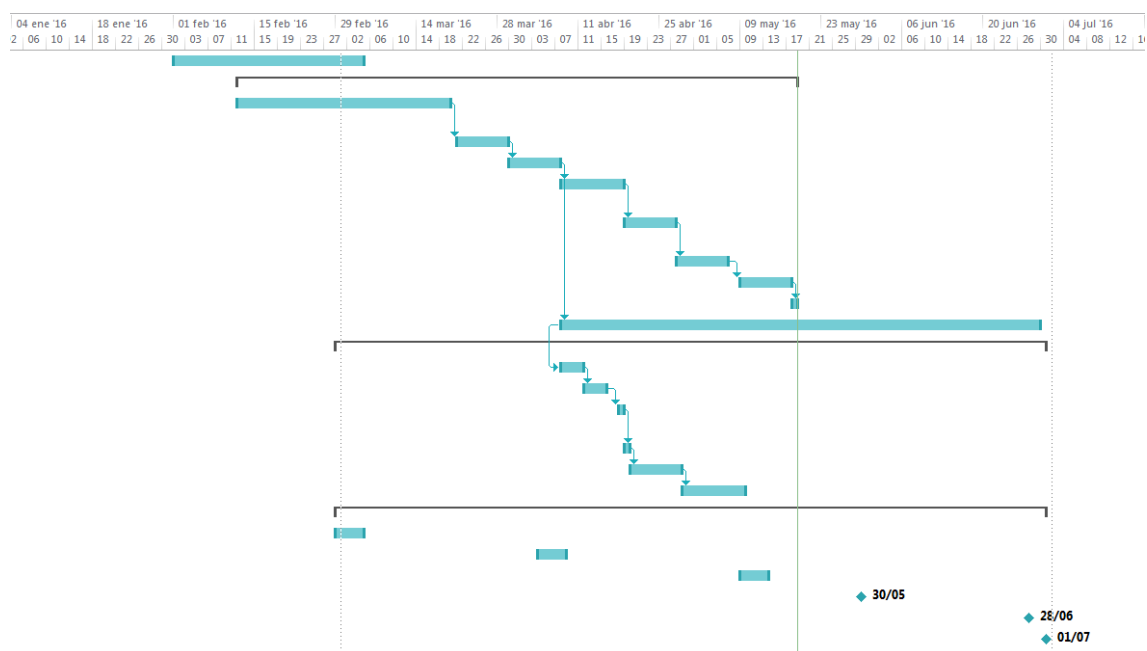


Diagrama de Gantt

		Modo de ▾	Nombre de tarea ▾	Duración ▾	Comienzo ▾	Fin ▾	Predecesoras ▾
1			Estudi treball previ	25 días	lun 01/02/16	vie 04/03/16	
2			◀ Comparació de genomes	69 días	vie 12/02/16	mié 18/05/16	
3			Correcció temps màxims d'execució	27 días	vie 12/02/16	sáb 19/03/16	
4			Tractament RAM	7 días	lun 21/03/16	mar 29/03/16	3
5			Paral·lelisme	7 días	mié 30/03/16	jue 07/04/16	4
6			Post-Tractament Seqüencial de Comparacions penjades	7 días	vie 08/04/16	lun 18/04/16	5
7			Correcció sortides i entrades de dades a la comparació	7 días	mar 19/04/16	mié 27/04/16	6
8			Test i correcció de fitxers E/S	7 días	jue 28/04/16	vie 06/05/16	7
9			Preparació de les dades d'entrada	7 días	lun 09/05/16	mar 17/05/16	8
10			Matriu de Similitud	1 día	mié 18/05/16	mié 18/05/16	9
11			Execució Sistema	59 días	vie 08/04/16	mié 29/06/16	5
12			◀ Correcció d'errors	89 días	lun 29/02/16	vie 01/07/16	
13			Descàrrega nous genomes	2 días	vie 08/04/16	lun 11/04/16	11CC
14			Correcció Offsets	4 días	mar 12/04/16	vie 15/04/16	13
15			Verificació de correcta entrada i sortida de dades a SMUMS	1 día	lun 18/04/16	lun 18/04/16	14
16			Correcció Sortida SMUMS	1 día	mar 19/04/16	mar 19/04/16	15
17			Adequació Sistema a Kron	7 días	mié 20/04/16	jue 28/04/16	16
18			Configuració execucions del Sistema	7 días	vie 29/04/16	lun 09/05/16	17
19			◀ Documentació	89 días	lun 29/02/16	vie 01/07/16	
20			Informe inicial	5 días	lun 29/02/16	vie 04/03/16	
21			Informe de Progrés I	5 días	lun 04/04/16	vie 08/04/16	
22			Informe de Progrés II	5 días	lun 09/05/16	vie 13/05/16	
23			Tancament del Projecte	0 días	lun 30/05/16	lun 30/05/16	
24			Documentació Final	0 días	mar 28/06/16	mar 28/06/16	
25			Presentació	0 días	vie 01/07/16	vie 01/07/16	

[Passos a Seguir](#)

Nivell de seguiment de la planificació

Ha hagut un retràs important degut a que s'ha hagut de corregir errors de programació en les modificacions que no s'han detectat a temps i que feien que no es generessin MUMS, la qual cosa impossibilitava l'avenç del projecte. Això ha obligat a parar totes les execucions llençades i a refer les comparacions que ja s'havien llençat anteriorment una vegada corregit l'error, ja que aquestes s'havien generat malament.

Fins ara s'ha aconseguit afegir un conjunt de millores o correccions a la comparació de genomes i s'està corregint els errors dels Offsets, encarregats del desplaçament i ordenació dels MUMS parcials entre cromosomes en un mateix fitxer de MUMS de tots els cromosomes, alhora que s'està verificant la sortida dels SMUMS.

Amb això queda finalitzada la fase de comparació de genomes, s'ha tornat a relançar les execucions y actualment s'està finalitzant la part de Correcció d'errors.

Exposició dels resultats

No són encara els resultats finals de la matriu de similitud ja que el llançament de les comparacions de tots contra tots ha estat detingut degut als errors mencionats anteriorment. Això és degut al fet de que si es generaven malament els MUMS implicava resultats erronis de tot el procés posterior.

De tota manera, s'ha aconseguit finalitzar:

- La correcció dels temps màxims d'execució que anteriorment eren molt grans en comparació a com havien de ser, amb això també es va afegir la reprogramació automàtica del rellotge per tal d'aconseguir que el programa considerés cada cert temps si s'havia de matar processos de comparacions.
- El tractament de la RAM, que avanç no es calculava bé la memòria RAM disponible (o lliure).
- S'ha verificat el correcte funcionament del paral·lisme a la funció de comparació dels genomes.
- S'ha afegit un post-tractament de comparacions seqüencials d'aquelles comparacions a nivell de cromosomes que es maten en el procés de comparació en paral·lel.
- S'ha corregit dades d'entrada i sortida al procés de comparació de genomes que feien que el procés de les comparacions fos més costós.
- S'ha afegit una nova funcionalitat, ara es genera de forma dinàmica una matriu de similitud cada vegada que finalitza una comparació entre genomes.
- S'ha modificat la generació de resultats per tal de que es generin després de cada comparació entre dos genomes i no quan finalitzi tot el procés de totes les comparacions.
- S'ha verificat la correctesa de les dades d'entrada a la comparació dels genomes.

Per altre banda, de forma menys significativa, s'ha netejat el codi per tal de reduir el temps d'execució eliminant sleeps innecessaris, eliminant la generació d'arxius necessaris per certs processos de la comparació modificant els processos que llegien d'aquests arxius, s'ha modificat els requisits necessaris per deixar pas a una nova comparació en el procés de comparació de genomes en paral·lel i s'ha generat un conjunt de scripts per tal de verificar certs arxius d'entrada i de sortida del procés de comparació.

Conclusions provisionals

Hem aconseguit corregir certes parts del programa que no funcionaven exactament com haurien de funcionar teòricament i s'ha optimitzat una mica el procés per tal de reduir el temps d'execució del programa. Tot i això, encara queda feina pendent com per exemple, arreglar errors d'offsets, adequar el sistema per a la utilització Kron, per tal de programar execucions cada cert temps, etc.

Bibliografia i 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] **Entrez Programming Utilities Help**, ajuda on-line per l'us de les consultes amb e-utils
<http://www.ncbi.nlm.nih.gov/books/NBK25501/>
- [4] <http://platypus.uab.es/>, 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).
- [5] **Efficient Space and Time multicomparison of Genomes**, Mario Huerta, Xavier Messeguer, Technical report LSI-02-64-R. Llenguatges i Sistemes Informatics, Universitat Politècnica de Catalunya (2002).
- [6] **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).
- [7] 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)
- [8] **Suffix Tree Construction with slide nodes**, Mario Huerta. technical report LSI-02-63-R Dep. Llenguatge i Sistemes Informàtics, Universitat Politècnica de Catalunya (2002).
- [9] <http://ibb.uab.cat/ibb/>, Web de l'Institut de Biotecnologia i de Biomedicina.
- [10] http://cic.puj.edu.co/wiki/doku.php?id=materias:laboratorio_de_lenguajes_ii:lableng2:reg_exp, Web amb informació de expressions regulars.
- [11] <http://geekalt42.net/el-porque-no-usar-un-task-killer-en-android-optimizando-la-ram-3813>, Web amb informació referent a la optimització de memòria RAM.
- [12] <http://www.chuidiang.com/linux/ipcs/semaforo.php>, informació de programació amb semàfors.
- [13] <http://linux.die.net/man/2/wait>, <http://osr507doc.sco.com/en/man/html.C/wait.C.html>, manage i informació de les funcions fork i wait.

- [14] <http://www.eclipse.org/> , pàgina oficial del projecte eclipse.
- [15] <http://www.enterprisedt.com/products/edtftpj/> , llibreries per la connexió FTP via java.
- [16] <http://www.cplusplus.com/>, Recurs imprescindible per la programació en C++.
- [17] <http://download.oracle.com/javase/6/docs/api/>, Recurs imprescindible per la programació en Java.