Eines i Recursos Bioinformatics

Miquel Perello Nieto

6 d'abril de 2012

Resum

En aquesta sessió veurem i provarem algunes eines microinformàtiques per a l'anàlisis de seqüències així com les bases de dades de referencia per a informació genòmica.

1 Introducció

Durant aquest laboratori podrem veure tots els passos seguits per estudiar una seqüència bacteriana i familiaritzar-nos amb les eines disponibles al nostre abast.

Primer de tot en la secció 2 veurem les eines que tenim per identificar quin es el bacteri amb el que estem tractant, comptant únicament amb una part del seu genoma.

Tot seguit en la secció 3 procurarem veure la estructura de la feature del genoma gracies a una eina especialitzada en aquesta tasca.

Després en la secció 4 seleccionarem alguns bacteris més i veurem amb quines eines podem comparar-los amb el nostre bacteri, i quines conclusions en podem treure.

2 Identificació de la seqüència

En aquest cas, com que sabem que les nostres seqüències provenen de genomes bacterians, utilitzarem una versió adaptada a la cerca en genomes microbians.

2.1 BLAST

El programa BLAST [1] ens ha fet una cerca entre tots els genomes bacterians que conté la seva base de dades per mostrar-nos quins són els més coincidents amb la seqüència que nosaltres li enviem.

Els resultats de la cerca han indicat que aquest genoma pertany a un *Acidiphilium cryptum* amb una probabilitat d'error de 1e-71. Tot seguit es presenten els tres resultats més similars a la cerca, i les dades de similitud amb l'original.

- Acidiphilium cryptum JF-5 chromosome
  - Max Score : 278
  - Total Score : 795
  - Query Coverage : 100%
  - E value : 1,00E-071
  - Max ident : 100%
- Acidiphilium multivorum AIU301
  - Max Score : 272
  - Total Score : 770
  - Query Coverage : 99%
  - E value : 5,00E-070
  - Max ident : 100%
- Acidiphilium sp. PM Ctg_00011
  - Max Score : 267
  - Total Score : 760
  - Query Coverage : 99%
  - E value : 2,00E-068
  - Max ident : 100%

La resposta de BLAST també ve acompanyada per una gràfica lineal en la que es pot veure el nivell de coincidències trobades en cada un dels genomes resultants (figura 1).

A demés el mateix BLAST ens proporciona la visualització d'un arbre genealògic dels diferents bacteris que ens ha trobat, en el que es pot veure les diferents punts de bifurcació en la seva línia evolutiva (figura 2).

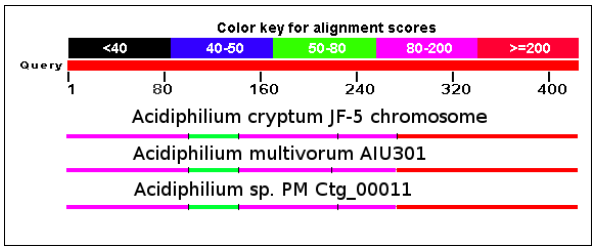

Figura 1: Gràfica de coincidències de BLAST

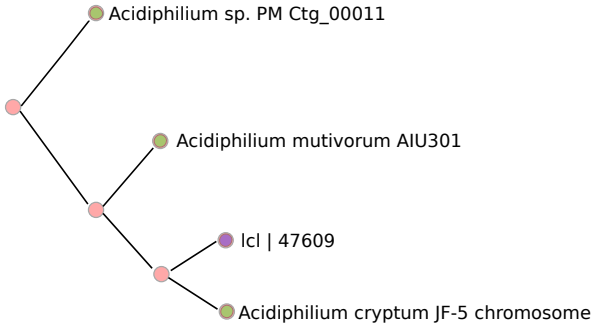

Figura 2: Arbre genealògic creat per BLAST

3 Estructura de la feature

En aquesta secció volem veure si a la feature del nostre bacteri podem apreciar una estructura repetida, o una diferenciació de zones, per descobrir això ens ajudarem de la eina Dotlet.

Amb un cop d'ull general a la matriu generada (figura 3)es pot apreciar que el genoma de la feature no es aleatori, i que segueix algun tipus de patró, ja que d'una seqüència aleatòria s'esperaria observar un núvol gris sense patrons, però podem observar que es generen uns requadres que indiquen la concentració d'algunes bases.

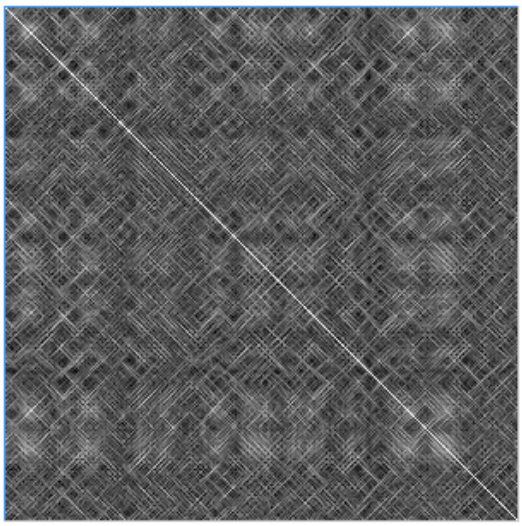

Figura 3: Matriu de punts generada amb Dotplot

3.1 Dotlet

Gracies a la eina Dotlet [4] podem generar una matriu de punts amb una escala de grisos modificable, per tal de comparar seqüències curtes i veure si existeixen zones de similitud. Una altra opció es comparar una seqüència amb ella mateixa, per poder comprovar l'existència de zones repetides o diferenciades.

4 Comparació de seqüències

El que volem veure és si podem trobar les semblances entre la seqüència genòmica del nostre bacteri i alguns altres, per tant hem seleccionat el genoma complert del nostre bacteri *Acidiphilium cryptum* i quatre més, entre els quals hi han dos que comencen amb la nomenclatura *Acid* i podrien contenir alguna similitud amb el nostre. Aquests genomes els hem pogut obtenir de *National Center for Biotechnology Information* [1].

- **Acidiphilium cryptum seq4**
- Acidaminococcus fermentans DSM 20731
- Acidithiobacillus caldus SM 1
- Bifidobacterium animalis lactis AD011
- Legionella pneumophila Philadelphia 1

Existeixen un parell de eines que ens faciliten aquesta tasca i ens mostren els resultat de manera visual, un es el programa M-GCAT [2]i l'altre es el Mummer.

M-GCAT és una eina per visualitzar i alinear les regions més grans en múltiples seqüències genòmiques.

La eina Mummer [3] ens permet comparar un parell de seqüències llargues i mostrar els resultats en una matriu de punts (entre d'altres representacions).

Primer de tot s'ha intentat alinear els cinc genomes indicats anteriorment, per comprovar si existia alguna similitud, però el resultat no ha tingut cap èxit. Un cop comprovat això, s'ha agafat el genoma del *Acidiphilium cryptum* i s'ha intentat aparellar amb tots els altres.

4.1 Acidiphilium i Acidaminococcus

Aquests dos genomes han tingut un alineament que sembla poc significatiu però que indica que tenen algunes zones de semblança. Es veu que el segon genoma es més petit, però sembla coincidir amb la primera part del primer en alguns casos (figura 4).

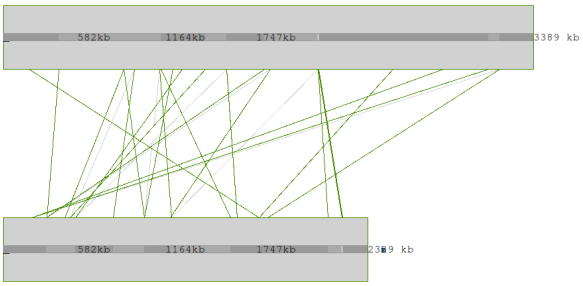

Figura 4: Alineament de M-GCAT per Acidiphilium i Acidaminococcus

Generant una gràfica de punts també podem veure la semblança entre els dos genomes, els diferents colors ens mostren la semblança de les cadenes comprovades i les seves complementaries (figura 5).

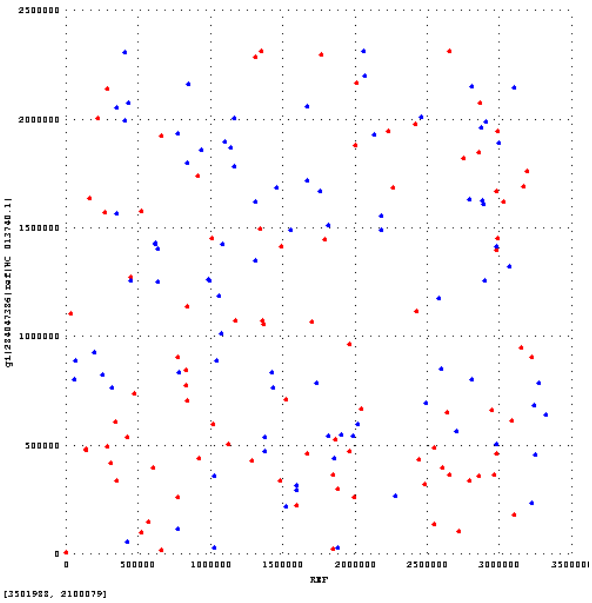

Figura 5: Gràfica de punts de Mummer per Acidiphilium contra Acidaminococcus

4.2 Acidiphilium i Acidithiobacillus

Aquests dos bacteris ja tenen bastantes més semblances tant en la mida del seu genoma com amb el nombre de coincidències en la seqüència. Podem veure que l'alineament es creua bastant en comptes de estar alineat de forma lineal(figura 6).

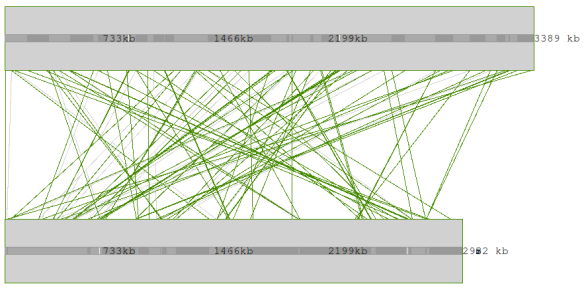

Figura 6: Alineament de M-GCAT per Acidiphilium i Acidithiobacillus

En el cas del gràfica de punts també podem veure que tal com deia el M-GCAT, existeix molta abundància de coincidències, el que ens pot indicar que aquest dos genomes poden tenir alguna relació (figura 7)

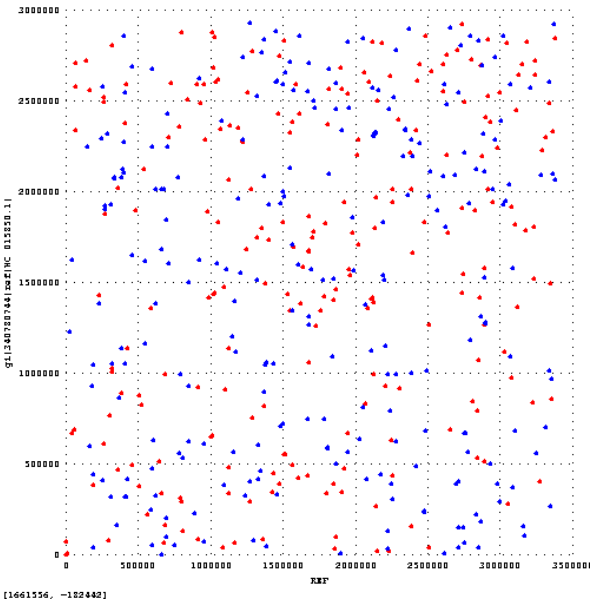

Figura 7: Gràfica de punts de Mummer per Acidiphilium contra Acidithiobacillus

4.3 Acidiphilium i Bifidobacterium

El bacteri Bifidobacterium fa de llargada dues terceres parts del primer, i com en la parella anterior també existeixen varies coincidències en les seves seqüències. Així com la majoria dels alineaments semblen estar creuats en la seqüència, existeixen varis alineaments que mantenen el seu ordre (figura 8). La gràfica generada ens demostra que existeixen dos patrons que comparteixen els tres genomes, que en el cas del primer i l'ultim es presenten en el mateix ordre, però que en canvi en el segon l'alineament està invertit.

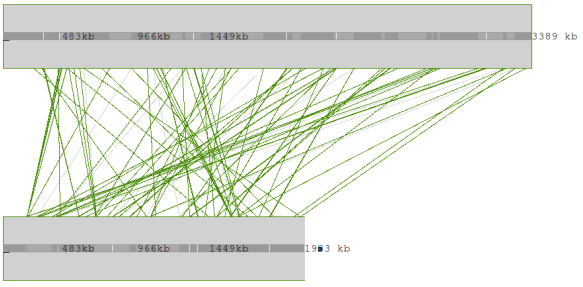

Figura 8: Alineament de M-GCAT per Acidiphilium i Bifidobacterium

En aquest cas també es veu una gran quantitat de coincidències entre els dos bacteris, així que amb la gràfica de punts (figura 9) es pot apreciar que el núvol de punts està distribuït per tota la seqüència i per les seves complementaries.

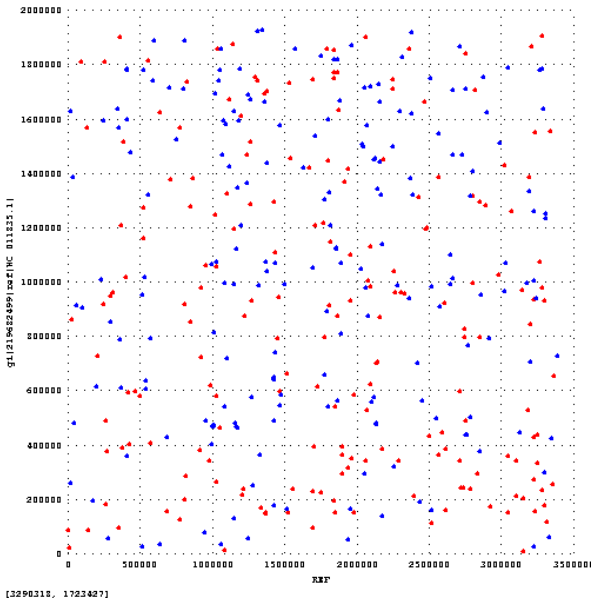

Figura 9: Gràfica de punts de Mummer per Acidiphilium contra Bifidobacterium

4.4 Acidiphilium i Legionella

En el cas dels alineaments amb la *Legionella* ens ha estat impossible trobar-los amb el programa M-GCAT, que ens ha dit que no hi havia coincidencies significatives. El programa Mummer en canvi ens ha dibuixat la gràfica, encara que es pot veure que no es gens significativa la relació que podria existir entre els dos bacteris.

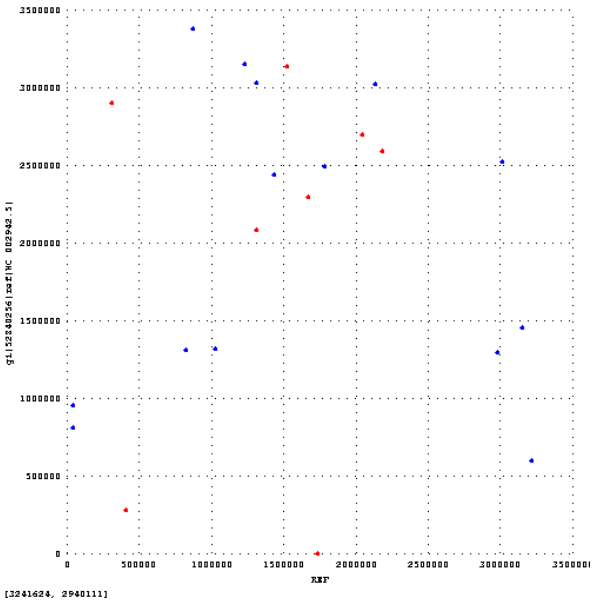

Figura 10: Gràfica de punts de Mummer per Acidiphilium contra Legionella

4.5 Acidiphilium , Acidithiobacillus i Bifidobacterium

Un cop vist que tant el *Acidithiobacillus* i el *Bifidobacterium* tenien algunes semblances significatives amb el nostre bacteri *Acidiphilium*, s'ha fet una comparació especifica per aquests tres bacteris. Voliem comprovar si entre tots tenien algunes similituds, així que els resultats de l'execució amb M-GCAT han estat les següents (figura 11).

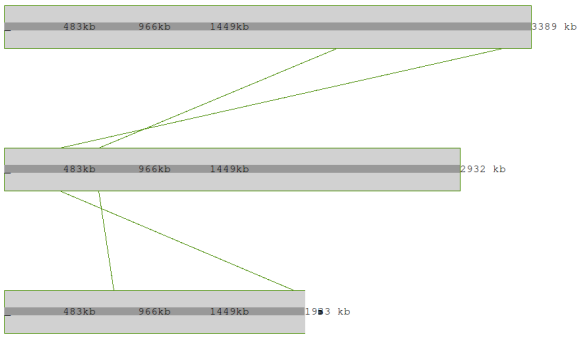

Figura 11: Alineament de M-GCAT per Acidiphilium , Acidithiobacillus i Bifidobacterium

5 Conclusions

Finalment amb l'elaboració d'aquesta práctica hem pogut veure moltes eines i recursos disponibles per l'analisis de dades biológiques. En aquest cas concret hem estudiat la sequencia d'un bacteri, que en un principi era desconegut, pero que primerament aplicant una cerca a la base de dades de BLAST hem determinat que es tractaba del bacteri *Acidiphilium cryptum*.

Tot seguit hem comprovat la feature d'aquest bacteri, per veure mitjançant el programa DOTLET si existía alguna estructura destacable en la seva sequencia, que hem vist que presentava certes zones diferenciables.

Per finalitzar l'anàlisis del bacteri hem comprovat si tenia alguna similitud amb altres bacteris agafats al atzar de la base de dades del *National Center for Biotechnology Information* [1], i hem intentat fer l'alineament entre el nostre bacteri i els escollits al atzar. Per fer aquests alineaments ens hem ajudat de dues eines informàtiques, la primera ha estat M-GCAT [2] la qual ens ha indicat els diferents alineaments entre les seqüències. I per altra banda el programa Mummer [3] que ens ha mostrat les coincidències en una gràfica de punts.

Referències

[1] "National Center for Biotechnology Information", `http://www.ncbi.nlm.nih.gov/`, U.S. National Library of Medicine

[2] "M-GCAT", `http://alggen.lsi.upc.es/recerca/align/mgcat/`, UPC

[3] "MUMmer", `http://mummer.sourceforge.net/`,

[4] "Algorithmics and Genetics Group", `http://alggen.lsi.upc.edu//`, UPC