Bioinforma't amb la bioinformàtica

1.INTRODUCCIÓ

A principis dels anys 80, l'avanç en el camp de la biologia molecular i el conseqüent i simultani desenvolupament d'àrees com la genòmica, han convergit en una impressionant muntanya de dades biològiques. Això últim, evidència de la imperant necessitat d'un vehicle o mitjà mitjançant el qual processar aquesta informació de forma ordenada i accessible, va actuar com a precursor per al sorgiment de la bioinformàtica, àrea de recerca interdisciplinària, situada a la frontera entre la biologia i la informàtica.

2.IMPORTÀNCIA

Tradicionalment, la investigació en l'àmbit de la biologia molecular semblava tenir com a marc únic el treball realitzat al laboratori, no obstant això, i com a conseqüència de la gran quantitat de dades generades en les últimes dècades, un dels grans reptes que ha d'assumir aquesta branca és el tractament d'aquestes.

D'aquesta manera, el principal desafiament que enfronta la bioinformàtica és l'emmagatzematge eficient i intel·ligent d'aquesta muntanya d'informació, la seva anàlisi i el desenvolupament de mitjans que ho permetin, sent també de natural importància proveir un fàcil i fiable accés als mateixos.

Així, podríem reduir a tres els centres al voltant dels quals s'ha de desenvolupar aquesta disciplina:

• La seqüència d'ADN determina la seqüència de proteïnes.

• La seqüència de proteïnes determina l'estructura de les proteïnes.

• L'estructura de les proteïnes determina la funció de les proteïnes.

D'altra banda, l'evolució d'eines capaces de facilitar l'enteniment d'aquests processos, suposaria també un pas més en la comprensió de la biologia dels organismes; aplanant el camí a aquesta " Generació de la seqüenciació".

3.APLICACIONS

Un cop emmagatzemats i disposats de manera que siguin fàcilment accessibles per a la comunitat científica, el següent pas per a l'adequada interpretació d'aquestes dades és la creació de mètodes que extreguin la informació continguda en els mateixos.

Aquest pont ve donat per les eines de la bioinformàtica, programari dissenyats per dur a terme aquest pas analític, segons els següents criteris:

• L'usuari final, pot no estar familiaritzat amb el maneig de la informàtica i la seva tecnologia.

• La seva disponibilitat, sent internet al portal idoni per aconseguir una distribució adaptada a la investigació científica.

Aquests instruments es classifiquen en quatre grans categories:

•*Eines per a la recerca de similitud*

Les seqüències homòlogues es defineixen com aquelles les quals la relació ve donada per divergència d'un antecessor comú. Així, en aquest cas, el grau de similitud entre dues seqüències pot ser avaluat en funció de ser vertadera o no la seva homologia.

Les eines que permeten aquesta comparació són adequades per a la identificació d'aquelles seqüències noves, l'estructura i funció són encara desconegudes.

• *Anàlisi de la funció proteica*

Aquest grup de programes permeten la comparació de seqüències de proteïnes amb aquelles bases de dades contenidores d'informació relativa als dominis i altres característiques i propietats d'interès que faciliten la determinació de la funció de la seqüència

d'una proteïna problema

•*Anàlisi de l'estructura*
Aquests programes permeten comparar estructures de proteïnes amb aquelles estructures ja conegudes i contingudes en bases de dades. Com que la funció d'una proteïna és una conseqüència més directa de la seva estructura que de la seva funció la determinació bi i tridimensional d'una proteïna suposa un pas crucial en l'estudi de la seva funció.

• *Anàlisi de la seqüència*
Finalment, els programaris inclosos en aquesta categoria ofereixen una àmplia gamma de dades relatives a la seqüència a analitzar, com ara anàlisi evolutiva, identificació de mutacions... que brinden punts de gran importància a l'hora d'esbrinar la funció específica de la mateixa.

4.“PROTOCOL”: COM IDENTIFICAR UNA SEQÜÈNCIA DESCONEGUDA?

Tal com hem vist, la gran majoria de funcions bàsiques de les eines en bioinformàtica se centren en la comparació i identificació de seqüències, sigui quina sigui la seva naturalesa.

El pas previ a aquest "anàlisi seqüencial" és lògicament l'obtenció d'aquestes seqüències, per a la qual existeixen diverses metodologies, com són el mètode de Sanger i més actualment, la piroseqüenciació o seqüenciació 454. Ara, és possible redactar una sèrie de passos a manera de recepta, útils a l'hora d'interpretar la informació continguda en les mateixes, partint d'un cas hipotètic en què disposem de dues seqüències problema d'ADN.

a) Introduir les nostres seqüències en el BLAST, indicant prèviament la seva naturalesa (Àcids nucleics) i procedència (Bacteri, Archaea, Eukarya). El programari BLAST troba aquelles altres seqüències presents a la base de dades del NCBI que presenten fragments símils a la nostra. D'aquesta manera ens proporciona un marc útil per esbrinar el gènere o espècie a què correspon la seqüència introduïda.

b) Inserir les nostres seqüències en MultAlin, programa d'alineament que indica el grau de similitud entre les dues seqüències introduïdes.
No obstant, en la majoria de les ocasions no és la informació continguda en els gens allò que interessa esbrinar, sinó la seva funció. Entra en escena l'anàlisi proteic.

a)Durant la traducció, els gens són llegits per codons (triplets), de manera que durant la seva lectura, en cas de tractar d'ADN bicatenari, hi haurà 6 pautes o marcs (ORFs).

b)Introduïm les nostres seqüències pel programari Translator. Obtenim 6 pautes de lectura. Seleccionem aquell marc l'inici correspongui a un codó d'inici (ATG).

c)La proteïna obtinguda, la inserim en el BLAST i indiquem la seva naturalesa.

d)Podem també fer una comparació de dues seqüències d'aminoàcids problema mitjançant el programa MultAlin.

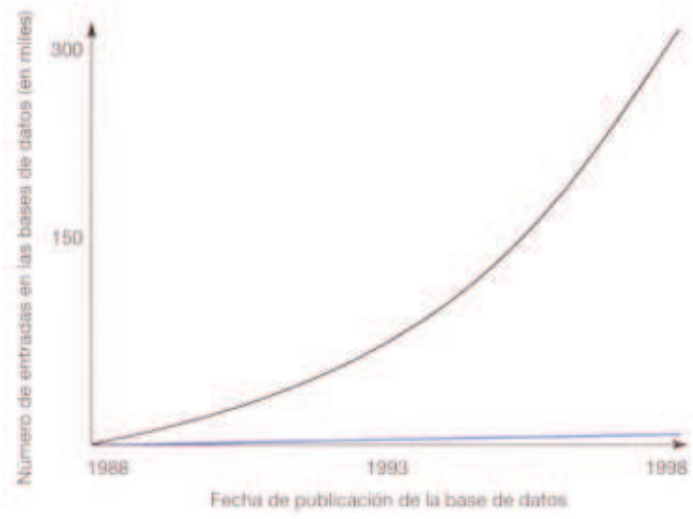
No obstant això, l'estudi de seqüències difereix considerablement en funció de la naturalesa de la seva composició. S'ha de tenir en compte que programes com MultAlin, emprats per esbrinar la similitud existent entre diverses seqüències, fan ús de diferents criteris en funció dels monòmers constituents de les mateixes.
En seqüències de nucleòtids es compara la mateixa seqüència, però en el cas de seqüències d'aminoàcids és compara l'estructura terciària per la qual cosa es consideraran iguals aquells fragments que

tot i no tenir la mateixa seqüència d'aminoàcids, si comparteixin la mateixa funció.
És per tant, de gran importància predir a partir de la seqüència lineal d'aminoàcids, l'estructura tridimensional de les proteïnes. Aquesta dificultat ha estat batejada amb el nom de "el problema del plegament", i representa un dels temes centrals en biologia molecular.
Per a la predicció de l'estructura terciària, els programari fan servir diferents aproximacions:
• Mètodes estadístics empírics que fan ús de paràmetres derivats d'estructures 3D ja conegudes.

• Mètodes basats en criteris fisicoquímics com ara hidrofobicitat, càrrega, impediment estèric, etc.

• Algorismes predictius que utilitzen estructures conegudes de proteïnes homòlogues.

No obstant això, aquesta determinació és encara un pas difícil per al que no existeix una tècnica totalment precisa i adequada. És per això que es parla del dèficit seqüència-estructura.



En negre apareix representat el nombre de seqüències entrades a les bases de dades, i en blau, les que disposen també d'estructura terciària associada.

5. TIPUS DE BASES DE DADES.
Bases de dades primàries.
Són dipòsits centrals en els quals es recull i emmagatzema la quantiosa informació sobre seqüències en literatura científica. Existeixen diversos projectes de bases de dades primàries:

•**ADN**:
-GenBank, la base de dades d'ADN del *National Center for Biotechnology Information* (Centre Nacional per a la Informació Biotecnològica, NCBI). Inclou seqüències de fonts disponibles públicament, i intercanvia dades tant amb la biblioteca de dades de l'EMBL com amb el DDBJ. S'estructura en divisions menors i concretes que faciliten una recerca ràpida i específica. GenBank pot consultar amb seqüències problemes de l'usuari a través de la interfície web del NCBI al conjunt de programes BLAST.

-*The European Molecular Biology Laboratory* (EMBL),la base de dades de l'*European Bioinformatics Institute* (Institut Europeu de Bioinformàtica EBI). Enllaça les principals bases de dades d'ADN i proteïnes amb altres especialitzades en motius, mapatges, estructures, etc . Inclou enllaços amb MEDLINE, i pot ser també consultada amb seqüències problema a través de les interfícies Web del EBI al BLAST.

-*DNA Data Bank of Japan* (Banc de dades de DNA del Japó). És també utilitzat per proporcionar eines de cerca estàndard com el BLAST.

•***Proteïnes***:
-*Protein Information Resource* (PIR). Actualment, i amb l'objectiu d'agilitar la recerca de resultats, es troba dividida en quatre seccions diferenciades per la qualitat de les seves dades i el nivell de dades proporcionat : PIR1 , amb entrades completament clasificadas i anotades, PIR2, que conté entrades preliminars no classificades prèviament amb possibles redundàncies, PIR3 , que inclou entrades no revisades i no verificades i PIR4, amb seqüències pertanyents a les categories traduccions conceptuals de seqüències artefactuals, de seqüències que no són transcrites o traduïdes, seqüències de proteïnes o traduccions conceptuals que han estat modificades genèticament de forma àmplia o seqüències no codificades genèticament no produïdes en ribosomes.

-*Martinsried Institute for Protein Sequences* (MIPS) SWISS - PROT . Proporciona anotacions d'alt nivell, incloent descripcions de la funció de la proteïna, estructura dels seus dominis, variants, modificacions post - traduccionals, etc.

-*TrEMBL* (EMBL traduïda) NRL -3D . Conté traduc-

cions d'aquelles seqüències codificants (CDS) del EMBL.

Bases de dades compostes
Com a conseqüència de la considerable proliferació de bases de dades primàries, va sorgir la necessitat de generar bases de dades en què s'emmagatzemés una varietat de fonts primàries diferents, fent més eficaces les cerques de seqüències en obviar la necessitat de consultar recursos múltiples.

Bases de dades secundàries
Acompanyant els nombrosos recursos primaris i compostos, existeixen les anomenades bases de dades secundàries o de patrons, batejades així pel fet que són el fruit de l'anàlisi de seqüències trobades en les fonts primàries.
Com a conseqüència de la gran varietat de metodologies existents per analitzar seqüències, sobretot proteiques, la informació continguda en aquests recursos secundaris presenta una enorme variabilitat reflectida en els diferents formats que adquireixen.

Alguns exemples de les principals bases de dades secundàries, amb la seva font primària i el tipus de patró emmagatzemat.

Base de datos secundaria	*Fuente primaria*	*Información almacenada*
PROSITE	SWISS-PROT	Expresiones regulares (patrones)
Profiles	SWISS-PROT	Matrices ponderadas (perfiles)
PRINTS	OWL*	Motivos alineados (huellas)
Pfam	SWISS-PROT	Modelos de Markov ocultos (HMM)
BLOCKS	PROSITE/PRINTS	Motivos alineados (bloques)
IDENTIFY	BLOCKS/PRINTS	Expresiones regulares borrosas (patrones)

*SWISS-PROT es la fuente de máxima prioridad de OWL

6. GLOSSARI
a.Alineaments :
Comparació lineal de seqüències aminoacídiques o d'àcids nucleics en la qual s'introdueixen insercions per fer que posicions equivalents en seqüències adjacents se situïn en el registre correcte . Els alineaments són la base dels mètodes d'anàlisi de seqüències i són emprats per ressaltar l'existència de motius conservats.

b.BLAST :
Mètode de recerca de similitud local, a través de l'alineament de seqüències.

c.Fingerprint (Petjada):
Grup de motius extrets d' un alineament de seqüències i emprat per construir una marca característica de pertinença a una família.

d.Homologia :
Estar relacionat pel procés evolutiu de divergència a partir d'un ancestre comú.

e.Modificació post-traduccional:
Alteració, catalitzada per un enzim d'una proteïna després de la seva traducció a partir del RNAm.

f.Motiu:
Una sèrie consecutiva d'aminoàcids en una seqüència proteica on el caràcter general es repeteix o està conservat en totes les seqüències d'un alineament múltiple en una posició concreta. Els motius poden correspondre amb elements estructurals o funcionals dins de les seqüències que caracteritzen.

g.MultAlin :
Eina per a l'alineament de seqüències.

h.Número d'accés:
Número o codi únic que serveix per marcar un registre d'una seqüència o patró en una base de dades primària o secundària.

i.Patrón:
Expressió de consens senzilla derivada d'una regió conservada d'un alineament de seqüències i emprada com a marca característica de pertinença a una família.

j.Pauta de lectura oberta (ORF, open Reading frame):
Sèrie de codons d'ADN, incloent un codó d'inici en 5' i un acabament, que codifiquen un gen conegut o potencial.

k.Secuenciación:
Determinació de l'ordre dels nucleòtids en una molècula d'ADN o ARN, o de l'ordre dels aminoàcids en una proteïna.

l.Secuencia codificant (CDS):
Regió de l'ADN o l'ARN la seqüència determina la seqüència d'aminoàcids d'una proteïna.

m.Sonda :
Seqüència d'ADN o proteïna emprada com a problema en una consulta en una base de dades.

n.Traduccions conceptuals:
Procés computacional d'interpretar la seqüència de nucleòtids de l'ARNm a través del codi genètic fins a una seqüència d'aminoàcids, que poden o no codificar una proteïna.

o.Traducció en sis pautes:
Traducció d'un tros d'ADN que té en compte tres traduccions directes i 3 reverses, que sorgeix de les tres possibles pautes de lectura d'un tros no caracteritzat d'ADN.

7. BIBLIOGRAFIA.

-Achard, F., Vaysseix, G., & Barillot, E. (2001). XML, bioinformatics and data integration. *Bioinformatics*, 17(2), 115-125.

-Baker, P. G., Goble, C. A., Bechhofer, S., Paton, N. W., Stevens, R., & Brass, A. (1999). An ontology for bioinformatics applications. *Bioinformatics*, 15(6), 510-520.

-Barreto Hernández, E. (2008). Bioinformática: Una oportunidad y un desafío. *Revista Colombiana De Biotecnología*, 10(1), 132-138.

-Febles Rodríguez, J. P., & González Pérez, A. (2002). Aplicación de la minería de datos en la bioinformática. *Acimed*, 10(2), 69-76.

-Franco, M. L., Cediel, J. F., & Payán, C. (2008). Breve historia de la bioinformática. *Colomb Med,* 39, 117-120.

-Perezleo Solórzano, L., Arencibia Jorge, R., Conill González, C., Achón Veloz, G., & Araújo Ruiz, J. A. (2003). Impacto de la bioinformática en las ciencias biomédicas. *Acimed*, 11(4), 0-0.

-Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), 2507-2517.

-Teresa, A., & David, P. (2002). *Introducción a la bioinformática [Introduction to Bioinformatics]*. Madrid, España: Prentice Hall.

-*European Nucleotide Archive*. Retrieved [12/16, 2012], from http://www.ebi.ac.uk/embl/

-National Center for Biotechnology Information Retrieved [12/16, 2012], from http://www.ncbi.nlm.nih.gov/