DOI: 10.2436/20.1501.02.41

Les biotecnologies
(Pere Puigdomènech i Francesc Gòdia, ed.)

Treballs de la SCB. Vol. 58 (2007) 11-24

BIOINFORMÀTICA

Roderic Guigó

Centre de Regulació Genòmica i Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra.

Adreça per a la correspondència: Roderic Guigó. Centre de Regulació Genòmica i Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra. Dr. Aiguader, 88. 08003 Barcelona. Adreça electrònica: *roderic.guigo@crg.es*.

RESUM

La recerca en biologia no es pot entendre avui sense la computació. A causa, sobretot, del desenvolupament de les tecnologies genòmiques, la biologia ha passat en molt poc temps, de ser una ciència en la qual l'esforç humà s'orientava principalment envers l'obtenció d'unes poques dades, a ser una ciència que genera un volum enorme de dades sense pràcticament intervenció humana. L'esforç de l'investigador s'ha desplaçat, en conseqüència, de la producció a l'anàlisi de les dades. I és en aquest desplaçament en què els mètodes informàtics tenen un paper essencial, tant en la planificació dels experiments com en la seva execució i, sobretot, en l'emmagatzematge i anàlisi dels resultats. Aquests mètodes configuren una nova disciplina científica, que anomenem *bioinformàtica*. En aquest article repassarem, des d'una perspectiva històrica, els fonaments d'aquesta disciplina, que s'articulen al voltant del concepte, entès de manera molt genèrica, de *alineament i similitud entre seqüències*.

Paraules clau: bioinformàtica, computació, seqüències, alineament, similitud.

BIOINFORMATICS

SUMMARY

Nowadays, research in biology can not be understood without computation. Due to the development of the genomic technologies, biology has been transformed in a very short period of time, from being a science in which the human effort was mainly oriented towards data gathering to being a science that generates a huge volume of data with little (or no) human intervention. The effort of researchers has, consequently, moved away from data production towards data analysis. Computational methods play an essential role to cope with this transformation: in the planning of the experiments, as well as in their execution, and, especially, in the storage and analysis of their results. These methods configure

a new scientific discipline named *bioinformatics*. In this article we review from a historical perspective the foundations of this discipline, which articulate around the generic concept of sequence alignment and similarity.

Key words: bioinformatics, computation, sequences, alignments, similarity.

INTRODUCCIÓ

La bioinformàtica és una disciplina en la intersecció entre la biologia i les ciències de la computació que tracta del desenvolupament i aplicació de mètodes computacionals per a l'obtenció, l'emmagatzematge, l'anàlisi i la interpretació de dades biològiques. Tot i que es tracta d'una disciplina recent, (Medline, la base de dades que compila la literatura científica en biologia i medicina, no inclou cap article en el qual aparegui el terme *bioinformatics* abans de l'any 1990), s'ha introduït en un període de temps curtíssim en moltes de les àrees tradicionals de la biologia. A finals de l'any 2007, el nombre d'articles de Medline que inclouen el terme *bioinformatics* s'apropa a vint mil, i això testimonia l'explosió insòlita d'una disciplina científica, possiblement sense precedents en la història de la ciència. Es tracta, en qualsevol cas, d'una disciplina jove i, per tant, encara sense sistematitzar. En un sentit molt ampli qualsevol aplicació de la computació en la biologia (i de la biologia en la computació) pot ser inclosa dins la bioinformàtica. En aquest article, però, entendrem la bioinformàtica, en un sentit més restringit, com la disciplina que tracta de l'anàlisi computacional de seqüències biològiques (DNA o proteïnes). De fet, s'atribueix normalment al desenvolupament de les tecnologies de la genòmica el paper clau que la bioinformàtica té avui dia en la investigació en biologia. En efecte, les tecnologies de la genòmica generen un volum de dades sobre els fenòmens de la vida d'una magnitud sense precedents en biologia. Aquestes dades són majoritàriament, però no únicament, seqüències de DNA dels genomes (cel·lulars, individuals, específics o poblacionals) i seqüències de RNA cel·lulars, l'abundància relativa de les quals, mesurada directament a partir del nombre de molècules seqüenciades o a partir dels senyals d'hibridació obtinguts mitjançant experiments amb matrius de DNA (els microxips de DNA), reflecteix l'activitat del genoma en una determinada condició cel·lular. Els experiments amb microxips són particularment il·lustratius de l'enorme volum de dades que produeix la recerca en biologia. Mitjançant els microxips, el suport físic dels quals tot just supera uns pocs centímetres quadrats, és possible obtenir en poques hores dades sobre el comportament simultani de milers de gens sota unes condicions determinades. Milers d'experiments amb matrius de DNA es porten a terme diàriament en tot el món, molts de manera gairebé automàtica. Fa només una dècada, tanmateix, l'obtenció d'aquestes mateixes dades sobre un únic gen era el resultat del treball continuat d'un investigador (o d'un equip d'investigadors) durant mesos, sovint durant anys. La biologia ha passat, doncs, en molt poc temps, de ser una ciència en la qual l'esforç humà s'orientava principalment envers l'obtenció de (poques) dades, a ser una ciència que genera un volum literalment vertiginós de dades sense pràcticament intervenció humana. L'esforç de l'investigador s'ha desplaçat, en conseqüència, des de la producció cap a l'anàlisi de les dades. I és en aquest desplaçament en què els mètodes informàtics tenen un

paper essencial, tant en la planificació dels experiments com en la seva execució, i, sobretot, en l'emmagatzematge i anàlisi dels resultats.

L'eclosió recent de la bioinformàtica, amb el desenvolupament de la genòmica, no sorgeix, però, del no-res. La història de la biologia molecular i de la informàtica, després de la Segona Guerra Mundial és, de fet, la història d'una interdependència creixent. Sempre és difícil posar una data exacta al naixement d'una determinada disciplina científica, però en el cas de la informàtica i la biologia molecular podríem dir que s'originen gairebé al mateix temps. Si, per posar alguna data, la biologia molecular neix l'any 1953, amb el desxiframent de l'estructura del DNA per part de Watson i Crick, i amb la determinació per primer cop, el mateix any, de la seqüència d'aminoàcids d'una proteïna per part de Sanger, la informàtica podríem dir que neix pocs anys abans, amb el desenvolupament dels primers ordinadors digitals programables en memòria, és a dir, els ordinadors tal com els entenem avui dia.

LES PRIMERES COL·LECCIONS DE SEQÜÈNCIES. LA MODELITZACIÓ DE L'EVOLUCIÓ MOLECULAR

Haurien de passar gairebé vint anys, però, perquè, gràcies a la progressiva miniaturització dels seus components, els ordinadors esdevinguessin suficientment petits, ràpids i econòmics perquè el seu ús pogués generalitzar-se i estendre's a les universitats i centres de recerca. Uns anys durant els quals, d'altra banda, va augmentar de manera considerable el nombre de proteïnes de les quals, seguint l'exemple de Sanger, s'havia aconseguit desxifrar la seqüència d'aminoàcids. A mitjan dècada dels seixanta, Margaret Dayhoff i els seus col·laboradors van començar a compilar aquestes seqüències d'aminoàcids. Aquestes compilacions van ser donades a conèixer als altres investigadors mitjançant els anomenats *Atlas of protein sequence and structure*. En la seva quarta edició, a finals dels seixanta, l'*Atlas* contenia prop de tres-centes seqüències de proteïnes. Aquests atles, encara llibres impresos

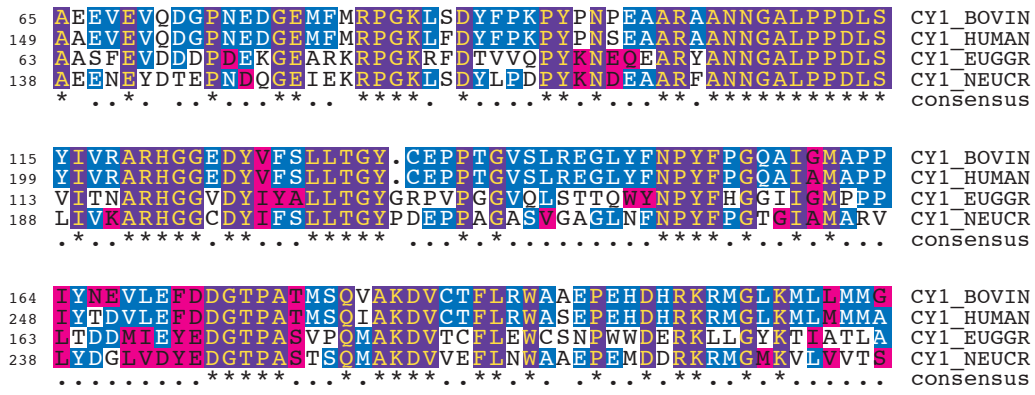

FIGURA 1. Alineament de les seqüències del citocrom C1 en diferents organismes: *Homo sapiens*, *Bos taurus* (vaca), *Euglena gracilis* (una alga unicel·lular) i *Neurospora crassa* (un fong). Alguns residus estan completament conservats (identificats amb un asterisc), en uns altres observem substitucions relacionades (identificades amb un punt), mentre que altres posicions són completament variables. En general, les posicions conservades en un alineament corresponen als aminoàcids més importants per al manteniment de la funció comuna en les proteïnes alineades. Per altra banda, la seqüència del citocrom C1 és molt més semblant entre els dos organismes mamífers, que entre aquests i les algues o els fongs. La similitud de seqüència constitueix, de fet, una bona indicació de proximitat filogenètica.

en paper, constitueixen possiblement les primeres bases de dades biomoleculars.

Dayhoff i els seus col·laboradors, però, no es van limitar a col·leccionar les seqüències, sinó que les van organitzar en famílies i superfamílies relacionades funcionalment, i d'acord amb el grau de semblança que presentaven. Per a cada família construïren el que s'anomena un *alineament múltiple* (vegeu la figura 1). Un alineament múltiple és una organització matricial d'una col·lecció de seqüències, en la qual cada fila es correspon amb una seqüència diferent (per exemple, la seqüència de la mateixa proteïna en espècies diferents) i cada columna correspon a una posició «equivalent» en les seqüències. Si les seqüències que es comparen són força similars, l'alineament múltiple es pot construir a mà sense gaire dificultat. De fet, Dayhoff va construir alineaments per a grups de proteïnes que compartien almenys el 85 % de la seva seqüència (Dayhoff *et al.*, 1978).

A partir d'aquests alineaments és possible investigar fins a quin punt determinades substitucions d'aminoàcids són tolerades per l'evolució. En efecte, Dayhoff assumia que els aminoàcids en una mateixa columna d'un alineament tenien el mateix origen evolutiu, és a dir, provenien d'un aminoàcid en una proteïna ancestral que havia mutat eventualment de manera diferent en les diferents proteïnes alineades. Sota aquesta assumpció, intercanvis entre aminoàcids que són observats sovint en la mateixa columna dels alineaments serien

C Cys	12																			
S Ser	0	2																		
T Thr	-2	1	3																	
P Pro	-3	1	0	6																
A Ala	-2	1	1	1	2															
G Gly	-3	1	0	-1	1	5														
N Asn	-4	1	0	-1	0	0	2													
D Asp	-5	0	0	-1	0	1	2	4												
E Glu	-5	0	0	-1	0	0	1	3	4											
Q Gln	-5	-1	-1	0	0	-1	1	2	2	4										
H His	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F Phe	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	-2	-3	-4	-5	-2	-6	0	0	17
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
	Cys	Ser	Thr	Pro	Ala	Gly	Asn	Asp	Glu	Gln	His	Arg	Lys	Met	Ile	Leu	Val	Phe	Tyr	Trp

FIGURA 2. Matriu de substitució PAM 250 construïda per Dayhoff. Vegeu el text per a una explicació del significat.

tolerats (*acceptats*, en la terminologia utilitzada per Dayhoff) per l'evolució, mentre que l'intercanvi entre aminoàcids que s'observen rarament en una mateixa columna dels alineaments serien penalitzats per l'evolució. Dayhoff va anar més lluny, i va quantificar aquesta tolerància a l'intercanvi d'aminoàcids durant l'evolució. A partir dels alineaments múltiples va construir les anomenades *matrius de substitució*. El valor dels coeficients d'aquestes matrius està relacionat amb la probabilitat d'observar durant un determinat període evolutiu la substitució d'un determinat aminoàcid per un altre. A la figura 2 hi ha representada una d'aquestes matrius. El valor zero és el valor neutral, és a dir, el valor que indica que la substitució d'un aminoàcid per l'altre en els alineaments múltiples ocorre amb la freqüència que esperaríem a l'atzar. Els valors positius (com ara, per exemple, el valor +2 entre arginina (Arg) i histidina (His)) indiquen que l'intercanvi entre aquests dos aminoàcids és observat amb una freqüència més elevada que l'esperada (i que, per tant, és un intercanvi favorable des del punt de vista evolutiu), mentre que els valors negatius (com, per exemple, el valor –7 entre glicina (Gly) i triptòfan (Trp)) indiquen que l'intercanvi entre els dos aminoàcids es produeix amb menys freqüència que l'esperada (i que es tracta, en conseqüència, d'un intercanvi penalitzat per l'evolució). La matriu de la figura 2 és la matriu anomenada PAM250, perquè les freqüències esperades de canvi es calculen durant un període evolutiu en el qual s'han produït una mitjana de 2,5 canvis en cada posició de la seqüència de les proteïnes comparades. Les matrius de substitució han estat calculades per a diferents distàncies evolutives amb mètodes diferents, i es troben al nucli dels programes de recerca de semblança de seqüència més usats avui dia.

L'ALINEAMENT DE SEQÜÈNCIES

A mesura que el nombre de seqüències de proteïnes augmentava, augmentava també l'interès per obtenir alineaments de seqüències cada cop més allunyades filogenèticament. Però, mentre que alinear dues seqüències molt similars és relativament simple, alinear dues seqüències allunyades és més complicat. Per exemple, suposem que volem alinear les seqüències ARNDCQ i ARDCK. Sembla que existiria consens en el fet que l'alineament

```
ARNDCQ
AR-DCK
```

és aquell que reflecteix millor la relació evolutiva entre els aminoàcids de les dues seqüències. (En aquest alineament, el caràcter «–», que es llegeix com a *gap*, indica la inserció d'un aminoàcid en una de les seqüències o la seva deleció en l'altra.) Però què ocorre si les dues seqüències que volem alinear són ARNDCQ i SKEAE? En aquest cas no és tan senzill fer una hipòtesi sobre l'alineament que reflecteix millor la seva història evolutiva comuna. Quin d'aquests tres alineaments, per exemple, és aquell que la reflecteix millor? Quin és el més semblant des del punt de vista evolutiu?

```
ARNDCQ     ARNDCQ     --ARNDCQ-
SK-EAE     -SKEAE      SKEA---E
```

O, més concretament, i fixant-nos només en els dos primers alineaments, quins esdeveniments són més probables durant el procés evolutiu, la substitució de A per S, de R per K i la inserció/deleció de N, com en l'alineament primer, o la inserció/deleció d'A, la substitució de R per S i la de N per K, com en el segon alineament? (la segona part de l'alineament és idèntica en els dos casos). Per respondre aquesta qüestió podem utilitzar les matrius de substitució construïdes

per Dayhoff, les quals quantifiquen precisament la tolerància de l'evolució als intercanvis entre aminoàcids. Així l'intercanvi entre una A i una S rep, d'acord amb la matriu PAM250 (vegeu la figura 2), una puntuació s(A, S) = 1. De la mateixa manera tenim, s(R, K) = 3, mentre que s(R, S) = 0 i s(N, K) = 1. Si suposem que la probabilitat d'una inserció/deleció és independent de l'aminoàcid inserit/delecionat, comprovem que les substitucions implícites en el primer alineament (A ↔ S i R ↔ K) són més probables que les substitucions implícites en el segon alineament (R ↔ S i N ↔ K), ja que tenen globalment valors positius més grans en la matriu de Dayhoff. Per tant, el primer alineament és més versemblant des del punt de vista evolutiu que no pas el segon. De fet, podem atorgar a cada alineament una puntuació, la qual és simplement la suma de les puntuacions, d'acord amb la matriu de Dayhoff, de les substitucions entre aminoàcids observades en cada posició. Així, la puntuació dels dos alineaments anteriors (suposant que l'alineament amb un *gap* tingui una puntuació de –1) seria:

```
 A R N D C Q          A R N D C Q
 S K - E A E          - S K E A E
+1+3-1+3-2+2=6       -1+0+1+3-2+2=3
```

El primer alineament té, efectivament, una puntuació superior al segon. La matriu de substitució de Dayhoff ens proporciona, en conseqüència, un criteri objectiu d'optimització per construir l'alineament entre dues seqüències: ateses dues seqüències, l'alineament òptim, és a dir, el més versemblant des del punt de vista evolutiu, és l'alineament que té la puntuació màxima de tots els alineaments possibles entre les dues seqüències.

Aquest criteri objectiu d'optimització ens proporciona un procediment simple per computar l'alineament òptim entre dues seqüències. Es tracta simplement de calcular la puntuació de tots els alineaments possibles entre les dues seqüències i escollir-ne un dels que tingui la puntuació màxima. El problema és que el nombre d'alineaments possibles entre dues seqüències és molt gran. Per exemple, el nombre d'alineaments possibles entre dues seqüències de cent aminoàcids cadascuna és aproximadament 10^{200}, un nombre impossible de calcular en un període raonable de temps amb cap ordinador (dels que existeixen avui dia o, fins i tot, dels que podrien arribar a existir). Per fer front a aquest problema, Nedleman i Wunsch (1970) van inventar un algoritme que permetia trobar l'alineament òptim entre dues seqüències, i requeria un nombre d'operacions «infinitament» més petit. Aquest algoritme es basa en una tècnica informàtica coneguda com a *programació dinàmica*, d'acord amb la qual, determinats problemes poden ser resolts molt més eficientment si són descompostos recursivament en subproblemes, a partir de la solució dels quals s'obté la solució del problema original. En el cas de l'alineament de dues seqüències, l'algoritme de Nedleman i Wuncsh es basa en el fet que l'alineament òptim entre dues seqüències X i Y (de longituds n i m) que acaben amb els residus x_n i y_m és l'alineament millor entre aquest tres alineaments possibles:

a) L'alineament òptim entre X_{n-1} i Y_{m-1}, seguit de l'alineament de x_n amb y_m (on X_j és la subseqüència de X que comença en la posició 1 i acaba en la posició j).

b) L'alineament òptim entre X_{n-1} i Y_m, seguit de l'alineament de x_n amb un *gap*.

c) L'alineament òptim entre X_n i Y_{m-1}, seguit de l'alineament de y_m amb un *gap*.

Els alineaments òptims entre X_{n-1} i Y_{m-1}, X_{n-1} i Y_m, i X_n i Y_{m-1} es troben, de la mateixa manera, descomponent-ne cadascun en tres (sub)possibilitats, i així successivament. Aquest procés recursiu de descomposició acaba quan un *gap* en una seqüència (és a dir, la seqüència buida X_0 s'alinea amb cada

una de les subseqüències de l'altra seqüència (Y_1, …, Y_n) i viceversa (Y_0 amb X_1, …, X_n). Les puntuacions d'aquests alineaments inicials són trivials de calcular, atesa la puntuació d'alineament amb un *gap* en la matriu de substitucions sota la qual es construeix l'alineament, i a partir d'aquestes es calculen recursivament els alineaments entre tots els parells de subseqüències de *X* i de *Y*. Tot i que inicialment aquest procediment pot semblar contraintuïtiu, és fàcil veure que dóna lloc, amb un nombre reduït de càlculs, a l'alineament òptim entre dues seqüències. De fet, el nombre d'operacions necessàries per obtenir l'alineament òptim entre dues seqüències utilitzant l'algoritme de Needleman-Wunsch és simplement proporcional al producte de la longitud de dues seqüències (és a dir, en el cas de dues seqüències de cent aminoàcids, un nombre d'operacions proporcional a 100^2; un nombre que és, en la pràctica, «infinitament» més petit que 10^{200}).

Mitjançant l'algoritme de Needleman i Wunsch es pot obtenir el que s'anomena *alineament global entre dues seqüències*, és a dir, l'alineament que inclou la totalitat dels residus de cada una de les dues seqüències comparades. Sovint, però, quan es comparen dues seqüències, només determinades regions exhibeixen una similitud de seqüència indicativa d'un origen evolutiu o d'una funcionalitat similar. Per exemple, quan es comparen dues seqüències ortòlogues entre genomes evolutivament allunyats, només les regions codificants exhibeixen una similitud de seqüència suficient perquè tingui sentit, des del punt de vista biològic, construir-ne l'alineament. L'any 1981 Smith i Waterman van desenvolupar una modificació de l'algoritme de programació dinàmica per obtenir el millor (o millors) alineament local entre dues seqüències (Smith i Waterman, 1981). La construcció d'alineaments locals entre una determinada seqüència i totes les seqüències emmagatzemades en una base de dades per identificar seqüències conegudes que puguin estar eventualment relacionades evolutivament o funcional amb la seqüència problema ha estat una de les tècniques més utilitzades en tota la biologia molecular en la darrera dècada del segle xx (vegeu més avall).

D'alguna manera, podríem dir que amb les matrius de substitució de Dayhoff i els algoritmes de Needleman-Wunsch i Smith-Waterman d'alineament de seqüències s'inaugura la disciplina de la bioinformàtica: per primer cop, es desenvolupen tècniques computacionals específiques per fer front a problemes d'origen biològic.

LES BASES DE DADES DE SEQÜÈNCIES

Malgrat que a finals dels anys seixanta s'havia compilat ja la seqüència d'aminoàcids d'alguns centenars de proteïnes, la seqüenciació d'àcids nucleics romania elusiva. A principis dels setanta, però, la situació canvia i gràcies als treballs de Maxam i Gilbert, d'una banda, i de Sanger, d'una altra, es posen a punt mètodes que permeten finalment la seqüenciació d'àcids nucleics. Curiosament, això ocorre pràcticament al mateix temps que el Departament de Defensa dels Estats Units desenvolupava ARPANET, una xarxa experimental d'ordinadors, que esdevindria més tard Internet, l'omnipresent xarxa d'ordinadors que tant ha modificat les nostres vides. Dos esdeveniments que es produeixen de manera totalment independent i dels quals només ara podem veure la relació: la seqüència del genoma seria impossible sense Internet. Què difícil és anticipar cap a on anirà la ciència!

A principis dels vuitanta, el nombre de seqüències d'àcids nucleics havia crescut de manera espectacular. Era evident que la distribució de les col·leccions de seqüències en format imprès, com ara els atles

compilats per Dayhoff, no podia continuar per molt més temps. Així, l'any 1982 es creava a Los Alamos National Laboratory a Nou Mèxic la base de dades americana de seqüències d'àcids nucleics en format electrònic, GenBank, i al Laboratori Europeu de Biologia Molecular (EMBL) a Heidelberg l'equivalent europea. Alguns anys més tard, al Japó, es crearia el DNA Data Bank of Japan (DDBJ). La primera versió de la base de dades d'EMBL, el juny de 1982, contenia 582 seqüències que sumaven poc menys de 600.000 nucleòtids. L'octubre de l'any 2007 conté més de cent milions de seqüències i prop de dos-cents mil milions de nucleòtids. El seu creixement continua sent exponencial (vegeu la figura 3). A més, multitud d'altres bases de dades (de seqüència i estructura de proteïnes, de regions promotores del gens, de processos bioquímics, d'experiments amb microxips, etc.) han estat creades des d'aleshores (vegeu, per exemple, l'edició especial de la revista *Nucleic Acid Research* dedicada a les bases de dades en biologia, que es publica el mes de gener de cada any (*http://nar.oxfordjournals.org/content/vol35/suppl_1/index.dtl*).

RECONEIXEMENT DE PATRONS EN SEQÜÈNCIES

Aviat la seqüenciació d'àcids nucleics reemplaçaria la seqüenciació de proteïnes; fins i tot per obtenir la seqüència d'una proteïna era més senzill obtenir primer la se-

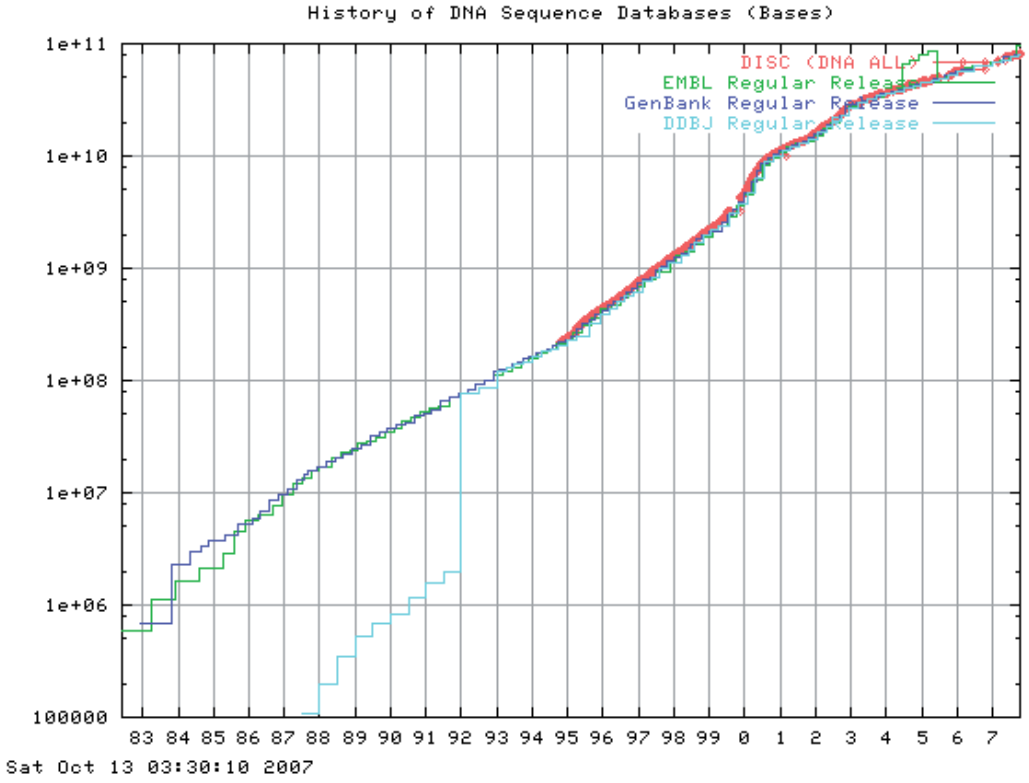

FIGURA 3. Creixement de la base de dades EMBL, que conté informació sobre les seqüències d'àcids nucleics conegudes. En el moment de la seva fundació, el 1982, les seqüències d'àcids nucleics dipositades a dins sumaven al voltant de 500.000 nucleòtids. L'octubre de 2007 sumaven al voltant de 200.000 milions.

qüència de DNA del gen que la codifica i a partir d'aquesta deduir la seqüència d'aminoàcids de la proteïna, que no pas seqüenciar la proteïna directament. Però mentre que la seqüència de les proteïnes és funcional en la seva totalitat, no ocorre el mateix amb la seqüència del DNA. En realitat, una gran part del genoma dels organismes eucariotes superiors sembla buit de funció. En particular, els gens eucariotes no es codifiquen de manera contínua en la seqüència de DNA, sinó que les regions que codifiquen, els exons estan separades per regions de longitud usualment molt major, els introns, de funció majoritàriament desconeguda. Senyals (motius, patrons) de seqüència delimiten les regions funcionals en la seqüència dels àcids nucleics i fan possible el seu reconeixement per part de la maquinària cel·lular. El problema del reconeixement de patrons en seqüències d'àcids nucleics (i també de proteïnes) comença a adquirir rellevància. L'any 1982, per exemple, Fickett (1982) desenvolupa un dels primers mètodes per identificar dominis funcionals en seqüències de DNA (regions codificants, en aquest cas concret). Aquests mètodes es basen en la definició d'un patró de seqüència (no sempre explícit), i en la implementació d'un algoritme per a la localització en seqüències problema.

Un patró de seqüència és un objecte que denota un conjunt de seqüències en un alfabet determinat (l'alfabet dels quatre nucleòtids en el cas del DNA, o l'alfabet dels vint aminoàcids en el cas de les proteïnes). Les expressions regulars constitueixen un dels mètodes més estesos per definir patrons

	-3	-2	-1	+1	+2	+3	+4	+5	+6
A	35.1	59.6	8.7	0.0	0.0	50.7	72.1	7.0	15.8
C	34.8	13.3	2.7	0.0	0.0	2.8	7.6	4.7	17.2
G	18.5	13.2	80.9	100.0	0.0	43.9	12.2	83.1	18.8
T	11.6	13.9	7.7	0.0	100.0	2.5	8.1	5.2	48.3
	C/A	A	G	G	T	A	A	G	T

FIGURA 4. *Part superior.* Matriu de puntuacions específica de posició (*position specific scoring matrix*, PSSM) corresponent als llocs donadors d'empalmament en gens humans. Els llocs donadors d'empalmament són els senyals que defineixen les fronteres entre les regions codificants (exons) i les no codificants (introns). L'intró comença en la posició +1. Com és possible observar, els dos primers nucleòtids de l'intró són sempre GT. Qualsevol nucleòtid pot apareixer en les altres posicions del senyal; no obstant això, alguns nucleòtids apareixen amb probabilitats molt distintes en algunes posicions. Així, per exemple, la probabilitat del nucleòtid G en la posició +4 és 0,83, mentre que la del nucleòtid C en aquesta posició és només 0,05. *Part inferior.* Representació gràfica en forma de logotip de la matriu de puntuacions. En cada posició, els nucleòtids apareixen en una grandària proporcional a la seva probabilitat. La grandària total de cada posició és proporcional al seu contingut informatiu. Com més allunyada de la distribució a l'atzar és la distribució de probabilitats dels nucleòtids en una posició, més informativa és aquesta posició.

de seqüència en aquests alfabets. En una expressió regular, l'alfabet original s'estén amb símbols addicionals que denoten, per exemple, la possibilitat de símbols alternatius en una determinada posició o el fet que una determinada posició pot no estar necessàriament present. Així el patró

C.[STA]..C[STA][^P]C

denota les seqüències d'aminoàcids constituïdes per la seqüència: cisteïna (C), qualsevol aminoàcid (.), serina, treonina o alanina ([STA]), qualsevol aminoàcid (.), cisteïna (C), serina, treonina o alanina ([STA]), qualsevol aminoàcid, llevat de prolina ([^P]) i cisteïna (C). Aquest patró constitueix la signatura específica de la regió d'unió al ferro de les proteïnes que pertanyen a la família de les ferredoxines. El patró prové, en aquest cas, de la base de dades PROSITE, una base de dades de patrons de seqüència d'aminoàcids molt usada en biologia molecular.

En ocasions és important especificar la probabilitat amb la qual els diferents símbols apareixen en cada posició, ja que probabilitats distintes en una determinada posició poden reflectir fenòmens biològics subjacents; per exemple, diferent afinitat en la interacció entre la seqüència que conté el patró i una altra bioseqüència distinta, depenent del nucleòtid o aminoàcid que ocupa aquesta posició. Aquestes probabilitats configuren les anomenades *matrius de puntuació específiques de posició* (*position specific scoring matrices*, PSSM) (vegeu la figura 4). Mitjançant aquestes matrius és possible calcular la probabilitat que una determinada

```
>sp|P01128|TSIS_SMSAV TRANSFORMING PROTEIN P28-SIS

          Length = 226

 Score =  140 bits (350), Expect = 2e-33
 Identities = 75/161 (46%), Positives = 100/161 (61%), Gaps = 11/161 (6%)

Query: 25  IPREVIERLARSQIHSIRDLQRLLEIDSVGSEDSLDTSLRAHGVHATKHVPEKRPLPIRR 84
           IP E+ + L+   I S  DLQRLL+ DS G ED  +  L     H+   +         R
Sbjct: 10  IPEELYKMLSGHSIRSFDDLQRLLQGDS-GKEDGAELDLNMTRSHSGGELESLA----RG 64

Query: 85  KRSI------EEAVPAVCKTRTVIYEIPRSQVDPTSANFLIWPPCVEVKRCTGCCNTSSV 138
           KRS+      E A+ A CKTRT ++EI R  +D T+ANFL+WPPCVEV+RC+GCCN  +V
Sbjct: 65  KRSLGSLSVAEPAMIAECKTRTEVFEISRRLIDRTNANFLVWPPCVEVQRCSGCCNNRNV 124

Query: 139 KCQPSRVHHRSVKVAKVEYVRKKPKLKEVQVRLEEHLECAC 179
           +C+P++V  R V+V K+E VRKKP  K+  V LE+HL C C
Sbjct: 125 QCRPTQVQLRPVQVRKIEIVRKKPIFKKATVTLEDHLACKC 165
```

FIGURA 5. Resultat de la comparació de la seqüència del *platelet derived growth factor* amb totes les seqüències d'aminoàcids en la base de dades amb el programa BLAST. El programa identifica totes les seqüències amb les quals la seqüència problema exhibeix alguna similitud i mostra el seu alineament local. En la figura es mostra, en concret, l'alineament entre la seqüència del *platelet derived growth factor* (*query*) i la *transforming protein P28-SIS* (*subject*), un oncogen. La seqüència intermèdia indica el grau de conservació dels residus alineats. Quan aquests estan conservats, l'aminoàcid conservat apareix en la seqüència intermèdia; quan ocorre una substitució conservada el signe «+» apareix en aquesta seqüència. Si no hi ha conservació, no apareix cap símbol. Els guions (–) indiquen *gaps* en una de les seqüències alineades. El *score* és la puntuació global de l'alineament (d'acord amb la matriu de substitució anomenada BLOSUM62, similar a la PAM250). *Expect* és el nombre d'ocasions en les quals esperem trobar per atzar alineaments amb una puntuació igual o superior a la puntuació en *score*. En aquest cas, la probabilitat de trobar per atzar un alineament com el de la figura és extraordinàriament baixa. Això ens indueix a pensar que les dues seqüències estan relacionades funcionalment. El descobriment per part de Doolittle d'aquesta semblança el 1983 va contribuir a millorar la nostra comprensió dels mecanismes moleculars involucrats en el càncer.

subseqüència pertanyi a un determinat patró. Poden utilitzar-se mètodes més sofisticats quan les posicions al llarg del patró no són independents. Una tècnica molt potent per representar patrons la constitueixen els anomenats *models de Markov ocults* (*hidden Markov models*, HMM). En aquests models s'assumeix que una determinada seqüència està constituïda per dominis funcionals diferents (per exemple, una seqüència gènica està constituïda per la successió d'exons i introns). El nombre i localització exactes d'aquests dominis o estats, però, són *a priori* desconeguts i el problema consisteix a identificar-los a partir només de la seqüència observada (de nucleòtids en aquest cas). Es coneixen, però, les probabilitats de cada residu característiques de cada domini particular, les anomenades *probabilitats d'emissió* (per exemple, els introns són en general més rics en nucleòtids A i T que els exons), així com les probabilitats de transició entre els diferents estats. Per tant, es pot calcular la probabilitat d'una seqüència observada, donada una determinada partició de la seqüència en estats. Existeixen algoritmes molt eficients per determinar quina és la partició de la seqüència en estats, que fan que la probabilitat de la seqüència observada sigui màxima. Aquesta partició és la que es considera la solució del problema.

Seqüència problema

1	2	3	4	5	6	7	8	9	10	11	12	13
W	A	T	S	N	A	N	D	C	R	I	C	K

Taula *hash K* = 1

A	C	D	I	K	N	R	S	T	W
2 6	9 12	8	11	13	5 7	10	4	3	1

FIGURA 6. Taula *hash k* = 1 d'una seqüència d'aminoàcids. En aquest cas, la taula *hash* simplement indica en quines posicions de la seqüència apareixen els diferents nucleòtids.

RECERQUES DE SEMBLANÇA EN BASES DE DADES

L'existència de compilacions electròniques de seqüències va facilitar-ne extraordinàriament l'anàlisi computacional. Va ser precisament mentre realitzava comparances entre les seqüències emmagatzemades en les recentment creades bases de dades electròniques que Doolittle va descobrir el 1983 la semblança entre la seqüència d'un oncogen i la seqüència d'un factor de creixement (vegeu la figura 5). Una relació que havia passat desapercebuda als investigadors de Harvard i de Caltech que havien estudiat aquest gen, i que contribuïa a la comprensió dels mecanismes moleculars involucrats en el càncer. Aquest i altres resultats semblants, en els quals la funció d'un gen era (almenys parcialment) inferida a partir de la semblança de la seva seqüència amb seqüències de funció coneguda, van demostrar la importància de les recerques de semblança en les bases de dades. A mesura que augmentava la grandària de les bases de dades de seqüències, però, els algoritmes de programació dinàmica desenvolupats per Needleman i Wunsch, i Smith i Waterman, van esdevenir massa lents per portar a terme, de manera eficient, recerques de semblança entre una seqüència nova i les seqüències prèviament emmagatzemades en les bases de dades. Programes com FASTA (Lipman i Person, 1988) i BLAST (Altschul *et al*., 1990) van resoldre aquest problema mitjançant la utilització d'algoritmes heurístics que proporcionaven alineaments generalment molt aproximats a l'alineament òptim, encara que no necessàriament l'alineament òptim, i que eren molt més ràpids.

Aquests algoritmes utilitzen una tècnica informàtica coneguda com a *taules hash* per accelerar la comparació i l'alineament de seqüències. Una taula *hash* de dimensió 1 d'una seqüència és una taula en la qual es registra la posició en què apareix cada un

dels vint aminoàcids (o dels quatre nucleòtids) en la seqüència (vegeu la figura 6). En una taula *hash* de dimensió 2 es registra l'aparició de cada diaminoàcid (o dinucleòtid), i així successivament. Aleshores, donades dues seqüències, es construeix la taula *hash* d'una d'aquestes. La segona seqüència, es compara aleshores amb aquesta taula. L'avantatge és que els ordinadors poden indexar les taules *hash* pels caràcters, en lloc de fer-ho per les posicions. És a dir, quan en la segona seqüència trobem el símbol «A» , un programa d'ordinador pot accedir directament (en una sola operació) el registre de la taula que conté les posicions en les quals el símbol «A» apareix en la primera seqüència. D'aquesta manera és possible identificar ràpidament regions d'alta semblança entre les dues seqüències, les quals s'utilitzen d'ancoratge per construir l'alineament, sense necessitat d'explorar totes les possibilitats que s'exploren implícitament en els algoritmes de programació dinàmica. El risc, però, és que l'ancoratge inicial porti a la construcció d'un alineament subòptim (és a dir, amb una puntuació inferior, d'acord amb la matriu de substitució emprada, que la puntuació màxima). Com més gran és la dimensió de la taula *hash*, més gran és el guany d'eficiència comparat amb l'algoritme de programació dinàmica, però més gran és també el risc d'obtenir un alineament subòptim.

L'enorme importància dels algoritmes de recerca de semblança en bases de dades en la investigació en biologia molecular queda reflectida en el fet que l'article que descriu el programa BLAST (Altschul *et al*., 1990) ha estat el més citat en biologia durant la dècada dels noranta.

EL PROJECTE GENOMA HUMÀ: BIOLOGIA I COMPUTACIÓ

Atesos els avenços en la dècada anterior, quan l'any 1990 començava «oficialment» el Projecte Genoma Humà, ja es considerava indispensable el concurs de la computació. Com pot llegir-se en un dels documents que a principis dels noranta va elaborar el Departament d'Energia (DOE), l'organisme que al costat dels Instituts Nacionals de Salut (NIH), ha estat responsable als Estats Units del desenvolupament del Projecte Genoma Humà: «Els sistemes computacionals tenen un paper essencial en tots els aspectes de la investigació genòmica, des de l'adquisició de les dades fins a l'anàlisi i manipulació. Sense ordinadors potents i sistemes apropiats per al tractament de les dades, la investigació genòmica és impossible.»

Tanmateix, el creixement exponencial de les bases de dades de seqüències durant la dècada dels vuitanta plantejava cada vegada problemes més greus de manteniment i actualització. El model original sobre el qual es van estructurar aquestes bases de dades requeria una considerable intervenció humana, un tipus de recurs que, òbviament, no podia créixer al mateix ritme. També, el nombre de bases de dades especialitzades creixia sense parar. No es tracta només de bases de dades de seqüències, sinó també de dades de molts altres tipus: mapes físics i genètics, estructures de proteïnes, xarxes metabòliques, dades funcionals a escales distintes, etc. Els investigadors necessiten accedir de manera immediata i transparent a aquesta informació. Per exemple, aquells investigadors interessats en un gen determinat volen conèixer la seqüència, la localització cromosòmica, la funció i l'estructura de gens similars, els teixits o estadis del desenvolupament en els quals s'expressa el gen, gens homòlegs en altres organismes, etc. A principis dels anys noranta aquesta informació es trobava dispersa en dotzenes de bases de dades especialitzades distintes, cadascuna amb la seva estructura pròpia i la seva manera peculiar d'accés.

Va ser precisament en aquells anys que

els científics del CERN (Organització Europea per a l'Energia Nuclear) van inventar la tecnologia *world wide web* (WWW) sobre Internet (la xarxa d'ordinadors desenvolupada als Estats Units gairebé dues dècades abans). WWW era la plataforma adequada que permetia resoldre molts dels problemes de manteniment, actualització, accés i integració de les bases de dades en biologia molecular. En certa manera, sense Internet i la tecnologia WWW el Projecte Genoma Humà no hauria estat possible, almenys, de la manera com els coneixem avui. De fet, Internet ha estat, i continua sent, el laboratori virtual en el qual té lloc la recerca genòmica. Tres sistemes, en particular, que utilitzen la tecnologia WWW, faciliten l'accés de tota la comunitat (i no solament de la comunitat científica) a la informació genòmica distribuïda en múltiples bases de dades: el sistema ENSEMBL (*http://www.ensembl.org*, vegeu la figura 7) a Europa, i el servidor de NCBI (*http://www.ncbi.nlm.nih.gov/Genomes/index.html*) i el Genome Browser (*http://genome.ucsc.edu/*) als Estats Units.

Certament, la quantitat de dades que genera la recerca genòmica fa que no puguem pensar, avui dia, la biologia sense la infor-

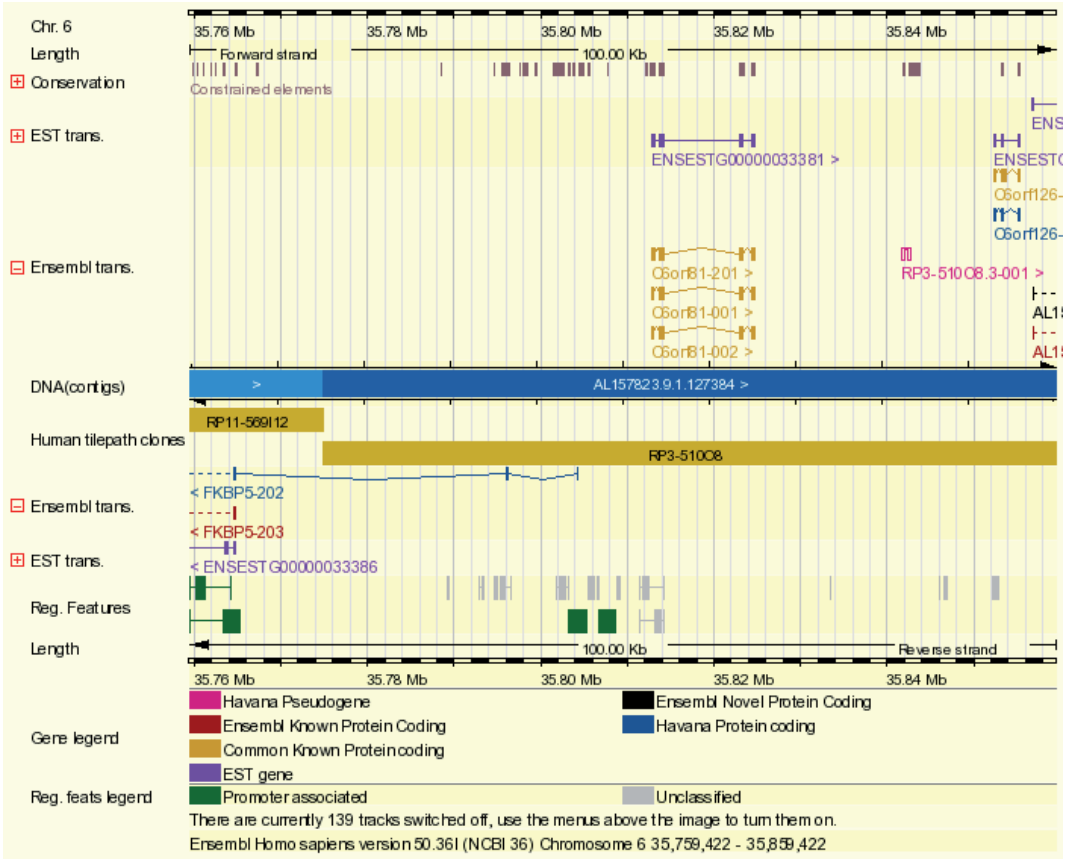

FIGURA 7. Visualització d'una regió del genoma humà a traves del navegador genòmic ENSEMBL (*www.ensembl.org*). Aquesta representació diagramàtica ens mostra els gens que han estat anotats o predits en una regió de cent mil nucleotids del cromosoma 6 humà.

màtica. Pot argumentar-se, tanmateix, que la generació massiva de dades no és patrimoni de la investigació en biologia, i que el mateix fenomen té lloc en àrees tan diverses de la ciència com l'astronomia, l'economia, la física d'altes energies, la meteorologia, etc. I que, en totes, els mètodes computacionals tenen un paper essencial en el tractament i interpretació de les dades. És cert. Però per què, doncs, com podem comprovar de nou a Google, existeixen tan pocs documents a Internet en els quals apareguin termes com *astroinformatics*, *meteoinformatics*, *econoinformatics*, o semblants, en contrast amb les desenes de milions de documents en els quals apareix el terme *bioinformatics*? En la meva opinió, això és en part així perquè la relació entre biologia i computació s'estableix a un nivell més íntim que no pas simplement el de la quantitat de les dades, i té a veure amb la «qualitat» (entesa com la naturalesa) d'aquestes dades. La vida comença quan els nucleòtids s'organitzen en la seqüència del genoma. Per sota de la seqüència del genoma hi ha la química i la física. I és l'ordre particular dels nucleòtids d'aquesta seqüència, més que no pas les seves característiques fisicoquímiques (el codi, com tan bé va anticipar Shröedinger als anys quaranta, que dicta les característiques biològiques dels éssers vius). I la vida, el desplegament en el món d'un ésser viu, és aleshores una computació, gairebé en un sentit paradigmàtic, sobre la seqüència del genoma. Un dels grans reptes de la biologia del segle XXI és, precisament, desxifrar els múltiples codis mitjançant els quals es produeix aquesta computació.

BIBLIOGRAFIA

Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. (1990). «Basic local alignment search tool». *Journal of Molecular Biology*, 215: 403-410.

Dayhoff, M.; Schwartz, R.; Orcutt, B. (1978). «A model of evolutionary change in protein». *Atlas of Protein Sequences and Structure*, 5: 345-352.

Fickett, J. W. (1982). «Recognition of protein coding regions in DNA sequences». *Nucleic Acids Research*, 105: 303-318.

Lipman, D. J.; Pearson, W. R. (1985). «Rapid and sensitive protein similarity searches». *Science*, 2271: 435-441.

Needleman, S. B.; Wunsch, C. D. (1970). «A general method applicable to the search for similarities in the amino acid sequence of two proteins». *Journal of Molecular Biology*, 484: 43-53.

Smith, T. F.; Waterman, M. S. (1981). «Identification of common molecular subsequences». *Journal of Molecular Biology*, 1471: 95-97.