DOI: 10.2436/20.1501.02.118
ISSN (paper): 0212-3037
http://revistes.iec.cat/index.php/TSCB

La biologia d'ahir i d'avui. Reflexions amb motiu del centenari de la Societat Catalana de Biologia (Francesc Piferrer, ed.)

Treballs de la SCB. Vol. 63 (2012), p. 183-198

BIOLOGIA I COMPUTACIÓ

Roderic Guigó

Centre de Regulació Genòmica i Universitat Pompeu Fabra

Adreça per a la correspondència: Roderic Guigó. Centre de Regulació Genòmica.
C. del Dr. Aiguader, 88. 08003 Barcelona. Tel.: 933 160 110.
Adreça electrònica: *roderic.guigo@crg.cat*.

RESUM

La biologia i la computació van esdevenir inextricablement unides durant la segona meitat del segle xx. Durant la primera dècada del segle xxi, els avenços en les tecnologies de seqüenciació dels àcids nucleics, juntament amb avenços més generals en automatització i robotització, han fet possible el monitoratge dels fenòmens de la vida amb una resolució sense precedents. La recol·lecció de dades sobre els sistemes vius a totes les escales, des de l'escala cel·lular i molecular fins als ecosistemes, es produeix a un ritme creixent que supera la capacitat dels sistemes computacionals. D'altra banda, la capacitat de monitorar els fenòmens de la vida d'una manera sistèmica està convertint la biologia, tradicionalment una ciència analítica en què la realitat és dissecada en els seus components més elementals per tal de ser compresa, en una ciència sintètica, en què el repte per a la intel·ligència humana és la integració d'informació múltiple i heterogènia per tal de generar coneixement sobre la realitat biològica. Mes enllà del seu paper com a magatzems de dades, els ordinadors tindran durant el segle xxi un paper d'importància creixent en totes les etapes de l'activitat científica, des de l'adquisició de dades fins a l'anàlisi i la interpretació, incloent-hi, possiblement, el raonament autònom i la generació i el contrast d'hipòtesis.

Paraules clau: bioinformàtica, biologia, computació.

BIOLOGY AND COMPUTATION

SUMMARY

Biology and Computation became inextricably linked during the second half of the 20th century. During the first decade of the 21st century, advances in methods to sequence nucleic acids, coupled with more general advances in automation, robotization, and multi-

Rebut: 16/06/2011. Acceptat: 15/07/2011.

plexing, have resulted in the capacity to survey the phenomena of life with unprecedented resolution. Data on life systems at all scales, from cells to ecosystems, is being collected at a pace that outstrips the capacity of current computational systems. On the other hand, the capacity to survey biological phenomena in a systemic way is turning Biology, traditionally an analytic science in which the natural world is dissected in its elemental components in order to be comprehended, into a synthetic science, where the challenge to human intelligence is to integrate multiple, heterogeneous, large-scale sources of information in order to generate biological knowledge. Beyond pure data storage and access, during the 21st century computers will play an increasingly important role in all steps of scientific endeavor, from data acquisition to analysis and interpretation —possibly including autonomous reasoning—, hypothesis generation and testing.

Key words: bioinformatics, biology, computation.

INTRODUCCIÓ

L'any 1912, l'any de la fundació de la Societat Catalana de Biologia, Stephane Leduc va publicar el llibre *La Biologie Synthetique*, terme amb el qual ens referim, avui dia, al disseny i la construcció de funcions i sistemes biològics que no es troben a la naturalesa, és a dir, a l'enginyeria de la vida. La biologia havia estat fins aleshores, i ho va continuar sent després, una ciència essencialment descriptiva. La recerca en biologia tenia majoritàriament com a objectiu la descripció dels fenòmens naturals. L'establiment de relacions de causalitat entre aquests fenòmens, és a dir, la transició d'una ciència purament descriptiva a una ciència teòrica i predictiva i, en conseqüència, la possibilitat d'actuar i intervenir sobre la realitat biològica, estava limitada per la poca resolució dels instruments d'observació d'aquesta realitat. Els processos fonamentals del fenomen de la vida es produeixen a escala molecular, i només poden ser compresos totalment a aquesta escala, i només la comprensió a aquesta escala fa possible una intervenció tecnològica efectiva. Però l'observació de la realitat a aquesta escala no és fora de la nostra capacitat natural. La invenció del microscopi al segle XVII —segurament l'avenç tecnològic amb un impacte més gran en el desenvolupament de la biologia— va incrementar substancialment la capacitat d'observació dels fenòmens biològics i va comportar, en conseqüència, canvis gairebé paradigmàtics, revolucionaris, en la nostra comprensió d'aquests fenòmens. L'escepticisme amb el qual la Royal Society va rebre les observacions inicials de Leeuwenhoek sobre organismes unicel·lulars, l'existència dels quals era desconeguda fins aleshores, n'és un exemple. L'escepticisme era tal que la Royal Society va decidir enviar un equip de metges i juristes a Delft per tal de determinar si les facultats mentals de Leeuwenhoek li permetien raonar amb lucidesa.

El microscopi va revelar un món desconegut i va proporcionar una base cel·lular a la fisiologia, la teoria sobre el funcionament dels éssers vius, però la naturalesa essencialment submicroscòpica de la vida era fora del seu abast. És només amb el descobriment del DNA i la seva estructura atòmica, a mitjan segle passat, que accedim per primer cop al nivell basal, fonamental, del fenomen de la vida: és quan els nucleòtids s'organitzen en la seqüència de DNA del genoma —aquell cristall aperiòdic constituït per la repetició d'un nombre petit d'unitats elementals, el codi Morse de la vida, com havia anticipat amb intuïció ex-

traordinària Erwin Schroëdinger— que sorgeix la vida, la biologia; per sota, en els nucleòtids individuals, no organitzats, només hi ha la química i la física.

Conèixer els principis moleculars sobre els quals s'organitza la seqüència de DNA del genoma no vol dir, tanmateix, ser capaç d'observar aquesta seqüència; haurien de passar encara dues dècades fins que l'any 1975, l'equip de Frederick Sanger a Cambridge fóra capaç d'obtenir, és a dir, «d'observar» per primer cop la seqüència del genoma d'un organisme viu, el virus *phi-X 174*. I encara dues dècades més fins que a cavall dels segles XX i XXI, hom fos capaç d'obtenir el primer esborrany complet de la seqüència del genoma humà. «El passat és el pròleg», escriu Shakespeare a *La tempesta*. I, efectivament, el Projecte del Genoma Humà ha estat només el pròleg de la revolució genòmica amb la qual ha començat el segle XXI. Avenços tecnològics extraordinaris —l'impacte dels quals en biologia podria ser similar al de la invenció del microscopi— estan convertint la seqüenciació d'àcids nucleics en un procés rutinari i econòmic. Així, per exemple, si el Projecte del Genoma Humà va tenir una durada superior a deu anys, va involucrar centenars de científics de tot el món i va tenir un cost aproximat de tres mil milions de dòlars, l'Institut de Genòmica de Pequín pot seqüenciar, al principi de l'any 2011, un genoma humà en cinc minuts amb un cost d'uns pocs milers de dòlars. L'època en la qual la seqüència del genoma de cada ésser humà sobre el planeta Terra serà coneguda —i restarà coneguda per sempre— acaba de començar.

La generalització de la seqüenciació del DNA, però, no serà l'única, ni potser la més important, de les conseqüències dels desenvolupaments en les tecnologies de seqüenciació dels àcids nucleics. Aquestes tecnologies ens permeten seqüenciar també molt eficientment, mitjançant una estratègia coneguda com a RNASeq, l'RNA cel·lular. L'RNA és el primer producte de l'activitat del genoma, la seva primera manifestació fenotípica, i fa de mitjancer necessari en tots els canvis fenotípics, que són causats per canvis en la seqüència del genoma. En aquest sentit, hi ha una relació gairebé unívoca entre l'estat (biològic) d'una cèl·lula i el seu contingut de RNA. Potser més sorprenent encara, gràcies a la invenció de protocols experimentals molt enginyosos, els instruments de seqüenciació de la generació actual no solament ens permeten obtenir la seqüència del DNA i de l'RNA dins la cèl·lula, sinó que ens permeten monitorar l'estatus bioquímic dels components de la cromatina (el conglomerat de DNA i proteïnes que constitueix els cromosomes). Canvis en aquest estatus, anomenat *epigenètic* —és a dir, que no afecta la seqüència de nucleòtids del genoma, però sí com aquesta és interpretada— modulen la síntesi de RNA a partir de DNA i confereixen l'especificitat cel·lular. Així, mitjançant la tècnica anomenada ChIPSeq, que combina immunoprecipitació de la cromatina amb seqüenciació massiva, és possible determinar els llocs del genoma en els quals s'uneix un determinat factor de transcripció. Mitjançant la mateixa tècnica és possible identificar les regions del genoma en les quals s'ha produït una determinada modificació bioquímica de les histones —les proteïnes que constitueixen els nucleosomes, les unitats estructurals bàsiques de la cromatina. Tècniques relacionades permeten determinar la posició exacta d'aquests nucleosomes i fins i tot les modificacions químiques, com ara la metilació, que afecten directament els nucleòtids de la seqüència del genoma.

Tots aquestes noves tecnologies generen una quantitat enorme de dades. La seqüència del genoma humà (o més precisament

la seqüència haploide del genoma nuclear d'una cèl·lula humana) té una mida aproximada de tres gigabases (és a dir, tres mil milions de nucleòtids, 3×10^9). A causa de les limitacions de les tecnologies actuals —amb les quals és impossible obtenir la seqüència contínua de molècules de DNA molt llargues— la seqüenciació acurada d'un genoma requereix alt grau de redundància. En la pràctica, és necessari seqüenciar l'equivalent a unes 200 Gb per tal d'obtenir la seqüència de 3 Gb del genoma humà. Aquesta és, per cert, la producció aproximada dels instruments de seqüenciació més avançats disponibles l'any 2011. El nombre de nucleòtids en l'RNA cel·lular és encara més gran, i alguns assaigs, com ara els de posicionament de nucleosomes o de metilació del DNA, requereixen una resolució més gran que la seqüenciació genòmica mateixa. El volum de dades que generen les tecnologies genòmiques és de tal magnitud, sense precedents en la història de la biologia, que el concurs de la computació, ni que sigui per a l'adquisició i emmagatzemant de les dades, és imprescindible. No es tracta, però, només de la genòmica. La creixent automatització i robotització de l'adquisició de dades (com ara el monitoratge continu de l'estat dels ecosistemes, l'enregistrament mitjançant vídeo del comportament d'animals, o l'obtenció d'imatges d'alta resolució durant el desenvolupament embrionari, per posar només alguns exemples), i les necessitats computacionals d'adquisició i emmagatzematge associades, s'han estès a tots els àmbits de la biologia. D'altra banda, aquesta capacitat creixent d'interrogar simultàniament tots els components del sistema fa que la biologia estigui deixant de ser una ciència exclusivament «analítica», en què la realitat és dissecada en els seus components més elementals per tal de ser compresa, i esdevingui cada cop més una ciència «sintètica», en què el repte està en la integració d'informació diversa per tal de comprendre el funcionament global del sistema biològic. Al principi del segle XXI, en conseqüència, l'esforç dels biòlegs s'està desplaçant de l'obtenció de dades (la tasca arquetípica del biòleg del segle XX) a l'anàlisi i interpretació —tasques per a les quals el concurs dels ordinadors és també imprescindible. En resum, els sistemes computacionals (els ordinadors i els sistemes de comunicació, i els programes informàtics) tenen un paper essencial en tots els processos des de l'adquisició i l'emmagatzematge fins a l'anàlisi i interpretació de les dades biològiques. Per fer front a aquestes necessitats una nova disciplina científica, la bioinformàtica, en la intersecció entre biologia i computació, ha emergit al final del segle XX, i, en poc temps ha esdevingut una de les disciplines centrals en la biologia (vegeu la figura 1).

LA CREIXENT INTERRELACIÓ ENTRE BIOLOGIA I COMPUTACIÓ DURANT LA SEGONA MEITAT DEL SEGLE XX

Tot i l'aparent novetat, les arrels de la relació íntima actual entre biologia i computació neixen gairebé amb la biologia molecular. De fet, els ordinadors tal com els entenem avui dia —és a dir, els ordinadors digitals programables en memòria— van entrar en funcionament al final dels anys quaranta, i són, en conseqüència, gairebé contemporanis al desxiframent de l'estructura del DNA per part de Watson i Crick i de la determinació per part de Sanger de la primera seqüència d'aminoàcids d'una proteïna. Dos esdeveniments que marquen d'alguna manera el naixement de la biologia molecular.

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 per tal que 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 anys 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, al final dels seixanta, l'*Atlas* contenia prop de tres-centes seqüències de proteïnes. Aquests *Atlas*, encara llibres impresos 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 funcionalment relacionades, i d'acord amb el grau de semblança que presentaven. Per cada família construïren el que s'anomena un alineament múltiple (vegeu la figura 2). Un alineament múltiple és una organització matricial d'una col·lecció de seqüències, en la qual cada fila correspon a 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 més sense massa dificultat. De fet, Dayhoff va construir alineaments per a grups de proteïnes que

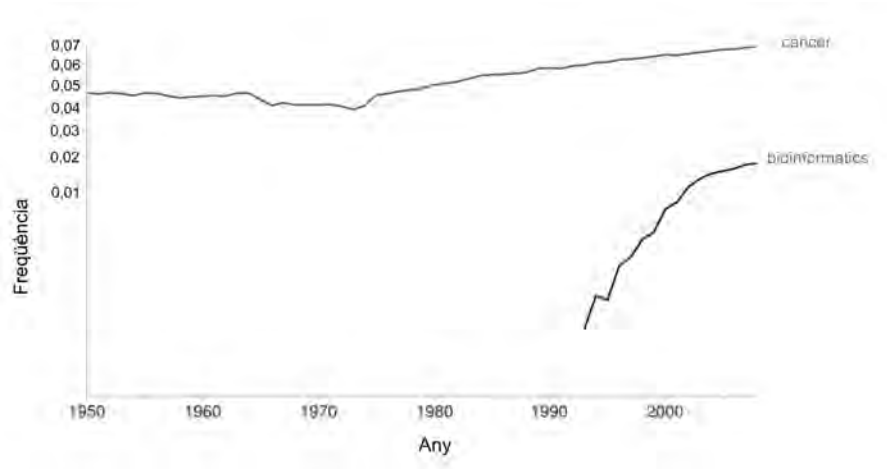

FIGURA 1. Freqüència d'articles a la base de dades bibliogràfica Medline en els quals apareixen el terme *cancer* o *bioinformatics*. Mentre que la freqüència d'articles amb la paraula *cancer* s'ha mantingut constant al voltant del 5 % des dels anys cinquanta, el terme *bioinformatics* no apareix fins la dècada dels noranta, però des d'aleshores la seva freqüència ha crescut de manera exponencial. L'any 2011, el terme *bioinformatics* apareix en més de l'1 % dels articles a Medline. Extret de ML-Trends (*http://www.ogic.ca/mltrends*).

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 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 seria penalitzat 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 3 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 tots dos aminoàcids es produeix amb menys freqüència que l'esperada (i que es tracta, en conseqüència,

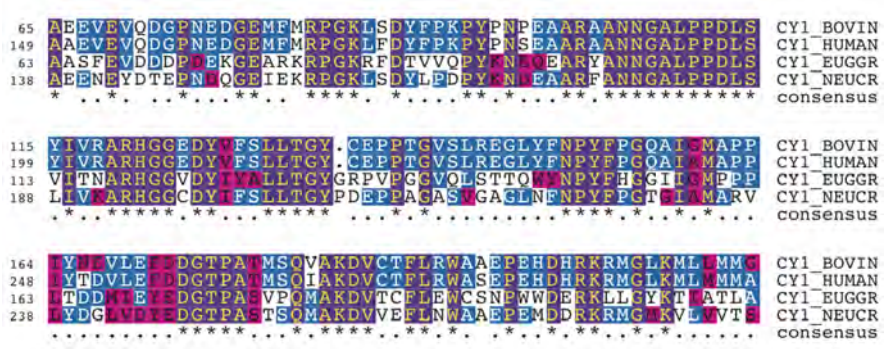

FIGURA 2. Alineament de les seqüències del citocrom C1 en diferents organismes: *Homo sapiens*, *Bos taurus* (vaca), *Euglena gracilis* (una alga unicel·lular) i *Neurospora crassa* (un fong). Alguns residus estan completament conservats (identificats amb un asterisc), en uns altres observem substitucions relacionades (identificades amb un punt), mentre que altres posicions són completament variables. En general, aquelles 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.

d'un intercanvi penalitzat per l'evolució). La matriu en la figura 3 és la matriu anomenada PAM250, perquè les freqüències esperades de canvi es calculen durant un període evolutiu en el qual s'ha produït una mitjana de 2,5 canvis en cada posició de la seqüència de les proteïnes comparades. Matrius de substitució han estat calculades per a diferents distàncies evolutives distintes i fent servir mètodes diferents, i són al nucli dels programes de recerca de semblança més usats avui dia.

L'alineament de seqüències

A mesura que el nombre de seqüències de proteïnes augmentava, també augmentava 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 d'aminoàcids ARNDCQ i ARDCK. Sembla que hi hauria consens 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 deleció en l'altre.) Però, què ocorre si les dues seqüències que volem alinear són ARNDCQ i SKEAE? En aquest cas no és tan senzill hipotetitzar l'alineament que reflecteix millor la història evolutiva comuna. Quin d'aquests tres alineaments, per exemple, és el que la reflecteix millor? Quin és el més versemblant 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

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	-5	-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

FIGURA 3. Matriu de substitució PAM 250 construïda per Dayhoff (extreta de *http://www.biotec.tu-dresden.de*). Vegeu el text per a una explicació del significat.

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 de 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 tots 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 3), 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 a 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'optimitat per tal de construir l'alineament entre dues seqüències: donades dues seqüències, l'alineament òptim, és a dir, aquell més versemblant des del punt de vista evolutiu, és l'alineament que té la puntuació màxima de tots els alineaments possibles entre totes dues seqüències.

Aquest criteri objectiu d'optimitat ens proporciona un procediment simple per tal de 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 hi ha 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 els 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 (en què 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, cadascun es descomponen en tres (sub)possibilitats, i així successivament. Aquest procés recursiu de descomposició acaba quan s'arriba a la seqüència buida X_0, que correspon a un gap, la qual s'alinea amb cadascuna de les subseqüències de l'altra seqüència (Y_1, …, Y_n). La seqüència buida Y_0 s'alinea, al seu torn, amb totes les subseqüències 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 100 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 hom pot obtenir el que s'anomena alineament global entre dues seqüències, és a dir, l'alineament que inclou la totalitat dels residus de cadascuna 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 tal d'obtenir el millor (o millors) alineaments locals 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 tal d'identificar seqüències conegudes que puguin eventualment estar 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, un problema d'origen biològic es planteja en termes computacionals i es desenvolupen tècniques informàtiques especifiques per tal de resoldre'l.

Les bases de dades de seqüències

Malgrat que al final dels anys seixanta s'havia compilat ja la seqüència d'aminoàcids d'alguns centenars de proteïnes, la seqüenciació d'àcids nucleics romania elusiva. Al principi 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 Departa-

ment 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 en podem veure la relació: la seqüència del genoma seria impossible sense Internet. Què difícil és anticipar cap a on anirà la ciència!

Al principi 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 gaire 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, hom crearia el DNA Data Bank of Japan (DDBJ). La primera versió de la base de dades d'EMBL, al juny de 1982, contenia 582 seqüències que sumaven poc menys de 600.000 nucleòtids. Des d'aleshores el seu ritme de creixement ha estat exponencial (vegeu la figura 4).

Cerques de similitud en bases de dades

L'existència de compilacions electròniques de seqüències va facilitar extraordinàriament l'anàlisi computacional. Va ser precisament mentre feia comparances entre les seqüències emmagatzemades en les recentment creades bases de dades electròniques que Doolittle va descobrir el 1983 la similitud entre la seqüència d'un oncogen i la seqüència d'un factor de creixement, una relació que havia passat desapercebuda als investigadors de Harvard i de Caltech i que contribuïa a la comprensió dels mecanismes moleculars involucrats en el

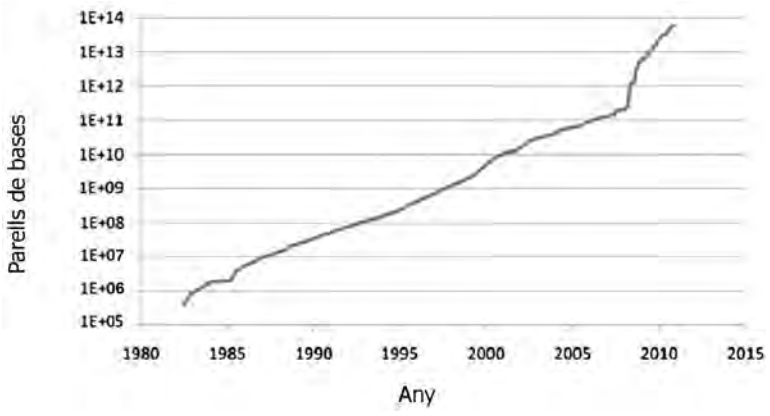

FIGURA 4. Creixement al llarg del temps de les bases de dades que contenen les seqüències d'àcids nucleics conegudes. La gràfica representa, a escala logarítmica, el nombre de nucleòtids emmagatzemats a la International Nucleotide Sequence Database Collaboration (INSDC; *http://www.insdc.org*) des de la creació de les primeres bases de dades, al principi dels anys vuitanta. Cal fer notar el punt d'inflexió que s'ha produït al final de la primera dècada del segle XXI, amb la generalització dels instruments de seqüenciació massivament paral·lels (extret de Cochrane *et al.*, 2010).

càncer. Aquest, i altres resultats semblants, en els quals la funció d'un gen era (almenys parcialment) inferida a partir de la similitud de la seva seqüència amb seqüències de funció coneguda, van demostrar la importància de les cerques de similitud en les bases de dades. A mesura que augmentava la grandària de les bases de dades de seqüències, però, els algorismes de programació dinàmica desenvolupats per Needleman i Wunsch, i Smith i Waterman, es van demostrar massa lents per portar a terme, de manera eficient, cerques de similitud entre una nova seqüència 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'algorismes 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 resum per tal d'accelerar la comparació i l'alineament de seqüències. Una taula resum 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 5). En una taula resum de dimensió 2 es registra l'aparició de cada diaminoàcid (o dinucleòtid), i així successivament. Aleshores, ateses dues seqüències, hom construeix la taula resum d'una d'aquestes. La segona seqüència, aleshores, es compara contra aquesta taula. L'avantatge és que els ordinadors poden indexar les taules resum 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 similitud 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 resum 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 algorismes de recerca de similitud 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.

Seqüència problema

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

Taula resum $K = 1$

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

FIGURA 5. Taula resum $k = 1$ d'una seqüència d'aminoàcids. En aquest cas, la taula resum simplement indica en quines posicions de la seqüència apareixen els diferents nucleòtids.

El Projecte del Genoma Humà: biologia i computació

Atesos els avenços en la dècada anterior, quan l'any 1990 començava oficialment el Projecte del Genoma Humà, la contribució de la informàtica ja era considerada indispensable. Com pot llegir-se en un dels documents que al principi dels noranta va elaborar el Departament d'Energia (DOE), l'organisme que juntament amb els National Institutes of Health (NIH) ha estat responsable als Estats Units del desenvolupament del Projecte del Genoma Humà: «Els sistemes computacionals tenen un paper essencial en tots els aspectes de la investigació genòmica [...]. Sense ordinadors potents i sistemes apropiats per al tractament de les dades, la investigació genòmica és impossible.»

Al mateix temps que s'iniciava el Projecte del Genoma Humà, al principi dels anys noranta, científics del CERN (Organització Europea per a l'Energia Nuclear) van inventar la tecnologia World Wide Web (WWW) sobre Internet. Durant la dècada següent, en paral·lel al progrés del Projecte del Genoma, el desenvolupament de la WWW va estendre l'accés i la utilització d'Internet gairebé a tota la humanitat i la WWW va acabar esdevenint, al principi del segle XXI, la manera natural d'accedir i interactuar a Internet. Quan el Projecte del Genoma Humà va culminar amb l'obtenció del primer esborrany de la seqüència, Internet, per mitjà del WWW, ja havia esdevingut el laboratori virtual en el qual els científics investigaven la seqüència del genoma humà. Sistemes com ara ENSEMBL (*http://www.ensembl.org*) i el *Genome browser* (*http://genome.ucsc.edu*) permeten accedir tant a la seqüència genòmica com a la informació que se'n recull sobre la funcionalitat, per exemple per mitjà de projectes com ara ENCODE (*http://genome.ucsc.edu/ENCODE*). Altres sistemes, com ara Galaxy (*http://main.g2.bx.psu.edu*), fan possible que els investigadors duguin a terme anàlisis d'aquesta informació utilitzant de manera transparent la infraestructura que hi ha a Internet.

ELS REPTES DE LA BIOLOGIA COMPUTACIONAL

Han passat deu anys des que el primer esborrany de la seqüència del genoma humà va ser publicat. Moltes promeses i esperances van ser posades en el coneixement d'aquesta seqüència (ens guariria, es deia, de les malalties, i ens allargaria la vida gairebé de manera indefinida). Deu anys més tard, tanmateix, moltes d'aquestes expectatives no han estat satisfetes. Si bé és cert que en la seqüència del genoma es troben les instruccions que codifiquen les característiques biològiques dels éssers vius, també és cert que avui dia, al principi del segle XXI, la manera com aquesta codificació es produeix ens és majoritàriament desconeguda. «Tot és el resultat de comparacions» escrivia Champollion al seu germà des de Grenoble, l'any 1818. Champollion es refereix al desxiframent del codi jeroglífic egipci, mitjançant la comparació del mateix text inscrit en la pedra de Rosetta en tres alfabets diferents: jeroglífic, demòtic i grec. Es tracta, en definitiva, del raonament inductiu, el qual ens permet generalitzar relacions de causalitat entre fenòmens a partir de l'observació repetida de la seva coocurrència. De manera anàloga, en genòmica, la comparació dels genomes de molts individus i de la correlació dels canvis en el genoma (per exemple, els canvis en un determinat nucleòtid) amb els canvis fenotípics (per exemple, el color dels ulls) ens ajuda a «desxifrar» les instruccions codificades en la seqüència del

genoma. Paral·lelament, la comparació dels epigenomes i de les poblacions de RNA cel·lular en molts individus i condicions diferents ens revela com aquestes instruccions es despleguen i operen a escala molecular. És aquest coneixement sobre el funcionament a escala molecular de les instruccions del genoma el que ens ha de permetre eventualment actuar sobre aquestes instruccions, per tal de modificar-les quan no funcionen com volem (i guarir malalties, per exemple), o per inventar-les i crear noves formes de vida. Fa deu anys, la capacitat tecnològica necessària per dur a terme aquest monitoratge molecular exhaustiu dels éssers vius, i per tant la comparació inductiva a la qual es referia Champollion, no existia. Ara ja existeix. L'obtenció de la seqüència del nostre genoma és ja al nostre abast i no és agosarat afirmar que abans que acabi aquesta dècada una bona part de la societat occidental, la part més opulenta almenys, tindrà accés a la seqüència del seu genoma. També durant aquesta dècada es generalitzaran els assaigs epigenòmics i transcriptòmics. Atès el seu paper privilegiat com a indicador fenotípic, el monitoratge de RNA mitjançant RNASeq, per exemple, s'usarà aviat més enllà de la recerca bàsica, en la medicina, l'agricultura, la biotecnologia i altres aplicacions tècniques de la biologia. RNASeq podria ser usat, per exemple, per al monitoratge ambulatori de la resposta dels tumors als tractaments. Podria esdevenir un component estàndard de les anàlisis de sang, un únic assaig capaç de monitorar moltes més variables, a un cost molt més reduït, que les anàlisis que es duen a terme avui dia. Fins i tot podríem imaginar, en un escenari gairebé de ciència-ficció, microdispositius que monitoressin en temps real l'RNA i l'epigenoma de determinats teixits o òrgans, per proporcionar d'aquesta manera informació actualitzada de l'estat cel·lular, i en conseqüència, de l'estat fisiològic dels individus. Tot i que aquesta mena d'intervencions resultarien potser massa oneroses en la vida diària, podrien ser útils en els casos en què els cossos són subjectes a condicions extremes, com ara la cirurgia, l'atletisme o l'exploració espacial.

Les possibles aplicacions de la seqüenciació d'àcids nucleics són encara inimaginables. La implantació, però, podria estar limitada, irònicament, per la capacitat computacional. La generalització de l'ús de les tecnologies de seqüenciació massiva al principi de la segona dècada del segle XXI ha suposat un punt d'inflexió en el creixement del volum de dades que genera la recerca genòmica (vegeu la figura 4). Si des de la seva creació, al principi dels vuitanta, fins ben entrada la primera dècada del segle XXI, el volum de la informació emmagatzemada en les bases de dades de seqüències ha crescut de manera exponencial, i s'ha doblat, de manera més o menys constant, cada divuit mesos, des de l'any 2008 fins al 2011, en només tres anys, el volum s'ha multiplicat per mil. Aquest creixement superexponencial supera de molt el creixement de la capacitat dels ordinadors, els quals, d'acord amb la llei de Moore, dupliquen la seva capacitat aproximadament cada dos anys. És a dir, el creixement de la capacitat dels ordinadors no és suficient per fer front al creixement de la capacitat d'informació genòmica. El model d'acord amb el qual totes les seqüències de nucleòtids obtingudes sobre la Terra són emmagatzemades per sempre de manera centralitzada no sembla sostenible. De fet, l'any 2011 potser passarà a la història (de la bioinformàtica) per ser l'any en el qual hom ha acceptat la impossibilitat de superar aquesta limitació: GenBank, la base de dades d'àcids nucleics als Estats Units, ha anunciat que, al final de l'any, deixarà d'emmagatzemar gran part de la

informació genòmica que es genera als laboratoris de tot el món. Cal desenvolupar noves eines i sistemes (tant a escala de maquinari com de programari) per tal de fer front a aquesta explosió de dades genòmiques, i de fet, han estat recentment desenvolupats alguns sistemes de compressió de dades específics per a la informació genòmica (Fritz *et al.*, 2011). Tanmateix, tot i que els problemes d'emmagatzematge de les dades biològiques poguessin eventualment ser resolts, un altre problema roman, de naturalesa més radical. Sense el procés de reflexió intel·lectual que duen a terme els científics —d'interrogació de la natura mitjançant l'experimentació i d'interpretació de les dades resultants mitjançant hipòtesis explicatives— les dades no es converteixen en coneixement. Però, mentre el volum i la complexitat de les dades biològiques creix de manera superexponencial, la intel·ligència global del planeta (en aquest cas, el nombre i la capacitat dels científics) no creix ni molt menys al mateix ritme. Les dades biològiques, obtingudes sovint de manera automàtica sense intervenció humana directa, comencen ja a acumular-se a manca de científics capacitats per analitzar-les. En conseqüència, el procés científic s'interromp i la traducció de la informació primer en coneixement i després en capacitat tecnològica s'alenteix. Davant aquesta situació, una de les solucions que ha començat a assajar-se en biologia consisteix en l'ampliació del paper que tenen les màquines dins el procés d'investigació científica. Així, l'any 2004, Ross King i els seus col·laboradors van crear un robot científic (King *et al.*, 2004). Aquest robot no solament porta a terme experiments —una tasca que els robots duen a terme de manera comuna en molts laboratoris de biologia— sinó que interpreta els resultats i genera noves hipòtesis, a partir de les quals planteja, al seu torn, nous experiments. El robot va ser dissenyat per tal d'elucidar la funció d'un conjunt de gens en el genoma del llevat. Sembla que va dur a terme aquesta tasca de manera més eficient que un investigador humà; tant el robot com l'investigador van arribar a les mateixes conclusions, però per arribar-hi el robot va plantejar un nombre menor d'hipòtesis i, en conseqüència, va fer un nombre menor d'experiments. Amb el robot científic traslladem a les màquines activitats, com ara la interpretació de la realitat, la reflexió, la planificació..., les quals consideràvem patrimoni de l'ésser humà. Es tracta de la mecanització del pensament que, en certa manera, Ramon Llull ja havia anticipat fa més de set-cents anys.

Tot i la intuïció de Ramon Llull, anticipar la ciència és molt difícil. Els fundadors de la Societat Catalana de Biologia, fa cent anys, difícilment podien anticipar la seqüència del genoma, i Internet com el laboratori per a la seva investigació. Hom desconeixia aleshores les bases moleculars de la vida —i en particular, el caràcter essencialment computacional de la seqüència del genoma— i, malgrat els avenços substancials en informàtica, i sobretot en comunicacions (amb l'extensió de l'ús del telèfon i el telègraf), era molt difícil de preveure l'existència d'ordinadors digitals. De fet, no es tracta tant que sigui difícil distingir entre allò que serà possible i impossible en el futur (el teletransport, la immortalitat, la vida extraterrestre...). És que és impossible imaginar el futur. Quan en el cinquantè aniversari del desxiframent de l'estructura del DNA, l'any 2003, van demanar a Sir Francis Crick si, en el moment del seu descobriment, ell i James Watson havien anticipat la possibilitat d'obtenir la seqüència completa del genoma —la qual havia estat desxifrada l'any 2001, un parell d'anys abans— la resposta de Crick no va ser només que no; és que no s'havien plantejat

aquesta qüestió, és que ni tan sols era possible plantejar-se-la. És molt difícil anticipar, en conseqüència, com serà el món d'aquí a cent anys. L'evolució cientificotecnològica recent, però, ens permet aventurar que durant el proper segle es produiran dos fets transcendentals en la història de la biosfera. D'una banda, serem capaços de dissenyar i crear de manera dirigida noves formes de vida. L'evolució, almenys des d'un punt de vista antropocèntric, deixarà de ser un procés exclusivament aleatori. D'altra banda, capacitats cognitives que considerem específicament humanes, com ara la intel·ligència, o almenys algunes de les seves manifestacions, deixaran de ser patrimoni dels éssers humans. Les màquines ocuparan àmbits cada cop més extensos, d'allò que anomenem humanitat. Dins l'activitat científica, tindran cada cop un paper més rellevant. I menys subordinat. En qualsevol cas, d'aquí a cent anys el món serà molt diferent. Els fundadors de la Societat Catalana de Biologia formaven part d'una generació que va intentar redreçar la situació de menysteniment i submissió en què es trobava el nostre país des de feia segles. Es tractava, suposo, de posar-lo al costat dels països més avançats, l'esforç dels quals fa possible el progrés cientificotecnològic, i fa el món en general més predictible i confortable. No van reeixir: en repassar la història recent de la informàtica i de la biologia molecular al segle XX —dues disciplines que al segle XXI canviaran el món substancialment— l'absència total d'investigadors catalans és desoladora. Podem atribuir-ho, certament, a la nostra dissortada història, i, en particular, a l'impacte, comparativament terrible, que el feixisme europeu va tenir per al desenvolupament de la ciència al nostre país. Al principi del segle XXI, però, hem de mirar el futur amb esperança i ambició. Com a científics i com a catalans el nostre desig hauria de ser que, quan d'aquí a cent anys els nostres compatriotes mirin endarrere, puguin sentir-se orgullosos de pertànyer a un país que, amb el seu esforç, i al costat dels altres, ha contribuït modestament, però significativa, a fer del món un lloc millor per viure.

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.

Cochrane, G.; Karsch-Mizrachi, I.; Nakamura, Y. (2010). «The International Nucleotide Sequence Database Collaboration». *Nucleic Acids Research*, 39: D15-D18.

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*, 10: 5303-5318.

Fritz, M. H. Y.; Leinonen, R.; Cochrane, G.; Birney, E. (2011). «Efficient storage of high throughput sequencing data using reference-based compression». *Genome Research*, 21: 734-740.

King, R. D.; Whelan, K. E.; Jones, F. M.; Reiser, P. G. K.; Bryant, C. H.; Muggleton, S. H.; Kell, D. B.; Oliver, S. G. (2004). «Functional genomic hypothesis generation and experimentation by a robot scientist». *Nature*, 427: 247-252.

Lipman, D. J.; Pearson, W. R. (1985). «Rapid and sensitive protein similarity searches». *Science*, 227: 1435-1441.

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*, 48: 443-453.

Smith, T. F.; Waterman, M. S. (1981). «Identification of common molecular subsequences». *Journal of Molecular Biology*, 147: 195-197.

SOBRE L'AUTOR

Roderic Guigó i Serra (Barcelona, 1959). Va estudiar biologia i filosofia a la Univer-

sitat de Barcelona i es va doctorar al Departament d'Estadística d'aquesta universitat l'any 1988. Durant el seu doctorat va investigar models matemàtics i de simulació en ecologia evolutiva. Va dur a terme estades postdoctorals a les universitats de Harvard i de Boston i a Los Alamos National Laboratory. Durant aquests anys es va formar en genòmica computacional, la disciplina en la qual s'emmarca la seva investigació avui dia. L'any 1994 es va incorporar com a investigador a l'Institut Municipal d'Investigació Mèdica. Des de l'any 2006 és coordinador del programa de bioinformàtica i genòmica del Centre de Regulació Genòmica. També és catedràtic de bioinformàtica de la Universitat Pompeu Fabra. Ha participat en nombrosos projectes genòmics, inclòs el Projecte del Genoma Humà. Actualment participa en el projecte ENCODE, un projecte internacional finançat pels instituts nacionals de la salut dels Estats Units, que té com a objectiu identificar totes les regions funcionals en la seqüència del genoma humà. L'any 2003 va rebre el Premi Ciutat de Barcelona a la investigació científica.