La biologia a l'alba d'un nou mil·lenni
(J. Bertranpetit, ed.)

Treballs de la SCB. Vol. 50 (2000) 97-106

LA BIOLOGIA COMPUTACIONAL

RODERIC GUIGÓ

Institut Municipal d'Investigació Mèdica (IMIM) i Facultat de Ciències de la Salut i de la Vida. Universitat Pompeu Fabra.

Adreça per a la correspondència: Institut Municipal d'Investigació Mèdica (IMIM). Doctor Aiguader, 80. 08003 Barcelona. Adreça electrònica: *rguigo@imim.es*

ELS INICIS

Després de la Segona Guerra Mundial es produeix l'eclosió de dues disciplines científiques, la informàtica i la biologia molecular, el progrès de les quals des d'aleshores ha estat espectacular i ha afectat profundament les nostres vides. A la fi dels anys quaranta entraven en funcionament els primers ordinadors digitals programables en memòria —és a dir, els ordinadors tal com els entenem avui en dia. Poc després, l'any 1953, Watson i Crick publicaven el famós article en el qual es descrivia l'estructura del DNA, el mateix any que Sanger determinava per primera vegada la seqüència d'aminoàcids d'una proteïna. L'any 1959, Perutz i Kendrew aconseguien determinar per primer cop l'estructura tridimensional d'una proteïna. És també als voltants d'aquest any que apareix el primer llenguatge de programació d'alt nivell d'utilització general, el FORTRAN. Amb els llenguatges d'alt nivell hom pot escriure les instruccions per tal que un ordinador resolgui un determinat problema, sense necessitat de conèixer com l'ordinador resol realment el problema: la utilització dels ordinadors deixa d'estar limitada als enginyers que els dissenyen.

Al principi dels anys seixanta els transistors començaven a substituir els tubs de buit en els circuits dels ordinadors i, en conseqüència, aquests esdevenien més petits, ràpids i econòmics. Cap a mitjan anys seixanta, la majoria de grans empreses processaven ja la informació financera utlitzant ordinadors digitals. D´altra banda, durant la primera meitat d'aquella dècada, experiments de Korana, Brenner, Ochoa i altres permeteren desxifrar el codi d'acord amb el qual la seqüència de nucleòtids del DNA especifica la seqüència d'aminoàcids de les proteïnes. L'any 1965 es determinava la primera seqüència d'un àcid nucleic: la seqüència d'una molècula de RNA de menys de cent nucleòtids. Mentrestant, el nombre de proteïnes de les quals hom havia pogut deduir la seqüència primària d'ami-

noàcids augmentava. A mitjan anys seixanta Margaret Dayhoff i Ledley van començar a compilar les seqüències d'aminoàcids conegudes. Tot i que inicialment aquesta compilació estava motivada per les necessitats de la recerca d'aquests investigadors, successives actualitzacions d'aquestes compilacions van ser posades de seguida a disposició de la comunitat científica; eren els *Atlas of Protein Sequence and Structure,* llibres en els quals Dayhoff presentava les seqüències agrupades en famílies de proteïnes funcionalment homòlogues. En la seva quarta edició a la fi dels anys seixanta, l'*Atlas* contenia al voltant de tres-centes seqüències de proteïnes.

A la fi dels anys seixanta, amb l'aparició dels circuits integrats, els ordinadors esdevenien encara més petits, ràpids i assequibles. La possibilitat de disposar d'ordinadors no es limitava ja a les grans empreses i als centres d'investigació i desenvolupament militar, sinó que els ordinadors començaven a ser a l'abast també de les universitats i dels centres de recerca. Amb l'assequibilitat dels ordinadors i la popularització dels llenguatges de programació d'alt nivell, la computació va anar esdevenint, en molts camps, part habitual de la pràctica científica. En el cas de la biologia, la repercussió va ser més gran en aquells camps en els quals l'anàlisi estadística o la modelització matemàtica tenen un paper més rellevant, com ara la genètica, l'ecologia o la fisiologia. En particular, els ordinadors van permetre dur a terme anàlisis més exhaustives de la compilació de seqüències d'aminoàcids de Dayhoff i col·laboradors. Durant els seixanta, aquestes seqüències s'utilitzaren sobretot per establir arbres filogenètics. Aquestes filogènies es construïen sota l'assumpció que la distància evolutiva entre dues espècies era proporcional a la freqüència de canvis en la seqüència d'aminoàcids de les seves proteïnes: naixia la filogènia molecular. L'any 1969 es produïa un altre fet rellevant: investigadors del Bell Laboratories creaven el sistema operatiu UNIX. Per les seves característiques, UNIX esdevindria amb els anys la plataforma sobre la qual es desenvoluparien la major part d'aplicacions computacionals en biologia molecular.

Durant els anys setanta, la tendència a la progressiva miniaturització dels components dels ordinadors i, en conseqüència, a l'increment de la seva potència i a la disminució del seu preu, va continuar. Cap a la meitat d'aquesta dècada, els fabricants de computadores van treure al mercat els primers miniordinadors: els ordinadors deixen de ser monopoli de les grans universitats i centres de recerca, i comencen a ser presents als departaments i laboratoris. La proliferació d'ordinadors i els avenços paral·lels en programació van conduir a l'aparició i el creixement de les xarxes informàtiques. És durant aquests anys que es produeix el desenvolupament d'ARPAnet, la xarxa del Departament de Defensa dels Estats Units, basada en el protocol TCP/IP. Hom considera habitualment ARPAnet la xarxa predecessora d'Internet. Pel que fa a la biologia molecular, l'any 1973 s'establia a Brookhaven la base de dades de les coordenades estructurals de les proteïnes. La confrontació de les seqüències de les proteïnes amb les seves estructures va permetre desenvolupar, a la fi dels setanta, els primers mètodes estadístics per a la predicció de l'estructura secundària de les proteïnes. Aquests mètodes es basaven en la propensió diferencial dels aminoàcids a trobar-se en els diferents elements estructurals. És també durant els anys setanta que es desenvolupen les tècniques de clonatge del DNA i de DNA recombinant. La seqüenciació de molècules de DNA, però, romania elusiva: si a la fi dels anys seixanta s´havien compilat ja alguns centenars de seqüències de proteïnes, el

nombre de molècules de DNA seqüenciades no era superior a la vintena. Fins i tot, hom havia arribat a afirmar que la seqüenciació del DNA era intrínsecament impossible. Aquesta situació va canviar radicalment a partir de l'any 1975 amb el desenvolupament dels mètodes de seqüenciació de Maxam i Gilbert, i de Sanger. L'any 1977, al laboratori de Sanger hom obtenia la seqüència completa del primer genoma d'un organisme: els gairebé 5.400 nucleòtids del genoma del virus ϕX174. En pocs anys, la seqüènciació de DNA esdevindria rutinària.

Al final dels anys setanta havia estat determinada ja la seqüència d'un nombre elevat de proteïnes i molècules de DNA, i els ordinadors eren potents i fàcilment programables: les eines necessàries per emmagatzemar i analitzar les noves dades que la recerca en biologia molecular genera s´han desenvolupat, doncs, paral·lelament, però independentment de les eines necessàries per generar aquestes dades. Tot i així, els mètodes estadístics i de modelització matemàtica, orientats essencialment al tractament de les quantitats, es mostraven limitats davant la peculiar naturalesa de les noves dades que la investigació en biologia molecular generava: seqüències de lletres. Dades de naturalesa diferent, que plantejaven nous problemes i requerien noves maneres de pensar. L'agrupació que Dayhoff havia fet de les seqüències d'aminoàcids en famílies homòlogues relacionades funcionalment o evolutiva havia demostrat que la seqüència concreta d'aminoàcids de les proteïnes i la seqüència de nucleòtids del DNA són portadores de gran quantitat d'informació sobre la funció i la història d'aquestes molècules: seqüències semblants indiquen una funció o una història similar. Així, el problema de determinar el grau de similitud entre dues seqüències esdevenia clau en biologia molecular. És en resposta a aquest problema que Needelman i Wunsch, primer, i Sellers després, durant els setanta, i Smith i Waterman, al principi dels vuitanta, varen elaborar els primers algoritmes de programació dinàmica per a l'alineament, comparació i determinació del grau de similitud de dues seqüències. En cert sentit podríem dir que és amb el desenvolupament d'aquests algoritmes que neix la biologia molecular computacional: en el moment en el qual les tècniques estadístiques o matemàtiques conegudes deixen de ser suficients per afrontar els nous problemes que les noves dades moleculars plantegen i es fa necessari desenvolupar noves tècniques matemàtiques, algorítmiques i computacionals.

LES GRANS COMPILACIONS DE SEQÜÈNCIES

Al principi dels vuitanta, el nombre de seqüències conegudes havia crescut de manera espectacular. Els avenços en tecnologia de seqüenciació del DNA havien convertit la seqüenciació del DNA en més simple que la seqüenciació de proteïnes, i cada cop era més freqüent inferir la seqüència d'aminoàcids d'una proteïna a partir de la seqüència de DNA que la codificava que no pas seqüenciar directament la proteïna. Es feia evident que la distribució de les col·leccions de seqüències en format imprès no podia mantenir-se gaire més temps. Així, l'any 1982 es creava a Los Alamos National Laboratory la base de dades americana de seqüències de nucleòtids en format electrònic, GenBank. Gairebé al mateix temps, el Laboratori Europeu de Biologia Molecular (EMBL) creava la seva pròpia base de dades electrònica de seqüències de DNA a Heidelberg. El primer *release* de la base de dades d'EMBL, el juny de 1982, contenia 582 seqüències que sumaven poc menys de 600.000 nucleòtids.

L'existència de compilacions electròni-

ques de seqüències va facilitar-ne enormement l'anàlisi computacional. Durant els primers anys de la dècada dels vuitanta, tècniques computacionals van ser desenvolupades tant per definir i identificar senyals sintàctics en les seqüències de DNA i proteïnes, com per mesurar regularitats estadístiques en la seva composició. En línies generals, els patrons sintàctics correspondrien als senyals efectivament responsables de les diverses funcions de les seqüències —llocs de *splicing*, elements promotors, llocs antigènics... —, mentre que les regularitats estadístiques —com ara les que caracteritzen les regions codificants en el DNA o les regions hidrofòbiques en les proteïnes— reflectirien el resultat sobre la composició de les seqüències de l'exercici d'aquestes funcions. La biologia molecular computacional es consolidava. L'existència de bases de dades en format electrònic, però, va fer possible sobretot la comparació sistemàtica de les noves seqüències amb les seqüències ja existents. L'any 1982, Lipman va proposar un mètode basat en la construcció de les anomenades *hash tables* per dur a terme recerques eficients en les bases de dades de seqüències. Aquest és el mètode en el qual estan basats programes tan populars, avui en dia, com FASTA o BLAST. El mateix any, IBM treia al mercat el primer ordinador personal, el PC: als laboratoris i als centres de recerca, els ordinadors començaven a ocupar les taules dels investigadors. Va ser precisament utilitzant el seu ordinador personal que Doolittle va descobrir l'any 1983, mentre duia a terme recerques rutinàries en les bases de dades de seqüències, la similaritat entre un oncogen i un factor de creixement. Una relació que havia passat desapercebuda a investigadors de Harvard i de Caltech, i que contribuïa substancialment a la comprensió dels mecanismes moleculars involucrats en el càncer. Aquest i altres resultats similars, en els quals la funció d'un gen era (almenys parcialment) inferida a partir de la similaritat de la seva seqüència amb seqüències de funció coneguda, van demostrar la importància de la biologia computacional. En la història de la biologia, la investigació purament teòrica (o, millor dit, no experimental) mai no s'havia demostrat tan rellevant. La recerca en biologia molecular començava a ser dependent de la computació.

Durant els anys vuitanta, el nombre de seqüències conegudes augmentava a un ritme exponencial. Tant a GenBank com a EMBL, però, les seqüències continuaven sent copiades a la base de dades directament dels articles on havien estat publicades. El manteniment i l'actualització de les bases de dades requeria, en conseqüència, una intervenció humana considerable. Òbviament, els recursos humans de què hom disposava per mantenir i actualitzar les bases de dades no podien créixer al mateix ritme que el nombre de seqüències. El resultat era que el període de temps transcorregut des de la publicació d'una seqüència a la seva introducció a la base de dades creixia continuadament. A la fi de l'any 1986, aquest endarreriment s'havia estimat en prop de dos anys, i no feia altra cosa que augmentar. El model original d'operació de les bases de dades era clarament insostenible. A mitjan anys vuitanta, però, les xarxes d'ordinadors —Internet, en particular— s'estenien per les universitats i centres de recerca. La concurrència d´aquest fet amb la creixent popularització dels ordinadors personals va fer possible la implantació d'un nou model de manteniment i actualització de les bases de dades: els autors enviaven directament a la base de dades les seqüències en format electrònic, amb freqüència creixent a través d'Internet. La intervenció humana necessària per mantenir i actualitzar les bases de dades es reduïa considerablement. Al principi dels noranta, aquest model

s'havia implantat plenament, i les bases de dades estaven, essencialment, actualitzades.

L'any 1990 començava «oficialment» el projecte del Genoma Humà als Estats Units. L'objectiu d'aquest projecte és l'obtenció de la seqüència dels tres mil milions de nucleòtids que constitueixen el genoma humà i la identificació dels aproximadament cent mil gens codificats en aquesta seqüència. Donada la magnitud del volum de dades que aquest projecte —i els projectes de seqüenciació dels genomes d'altres organismes— hauria de generar, i la rellevància que l'anàlisi computacional de seqüències havia demostrat, hom era conscient des que es va concebre que el projecte del Genoma Humà era impossible sense el concurs de la computació. Com es pot llegir en un document de principis dels noranta del Department of Energy (DOE), l'organisme que juntament amb els National Institutes of Health (NIH) és el reponsable del desenvolupament del projecte del Genoma Humà als Estats Units: «El Programa del Genoma Humà produirà grans quantitats de dades complexes tant sobre la seqüència de DNA com sobre els mapes que se'n construeixen. El desenvolupament de projectes informàtics en algoritmes, programari i bases de dades és crucial per a l'acumulació i la interpretació d'aquestes dades de manera robusta i automatitzada en els centres de seqüenciació genòmica... Els sistemes computacionals tenen un paper essencial en tots els aspectes de la recerca genòmica, des de l'adquisició de les dades fins a la seva anàlisi i manipulació. Sense computadors potents i sistemes apropiats per al tractament de dades, la recerca genòmica és impossible.»

CAP A UNA RECERCA INTEGRADA

Al principi dels noranta, però, no era només als centres de seqüenciació massiva on la informàtica s'havia convertit en essencial; la utilització dels ordinadors s'estenia a pràcticament qualsevol laboratori de biologia molecular, i la utilització, ja fos local o remota a través d'una xarxa informàtica, de programes d'anàlisi de seqüències esdevenia rutinària. Al marge dels projectes genòmics, doncs, durant la primera dècada dels noranta, el volum de dades moleculars acumulades a les bases de dades continuava creixent de manera exponencial. No es tractava només de seqüències de DNA i de proteïnes, sinó de moltes altres dades de diferents tipus: mapes físics i genètics, estructures de proteïnes, xarxes metabòliques, dades funcionals a diferents escales. Els investigadors en biologia molecular desitjaven accés immediat a tota aquesta informació. Per exemple, els investigadors interessats en un determinat gen voldrien conèixer la seqüència, la localització cromosòmica, la funció i l'estructura de les proteïnes codificades per gens similars, els teixits o estadis del desenvolupament en els quals el gen s'expressa, els potencials gens homòlegs en organismes models... Aquesta informació, tanmateix, es troba dispersa en dotzenes de bases de dades especialitzades independents, cada una amb la seva pròpia estructura i el seu peculiar mecanisme d'accés. La necessitat de disposar d'un sistema integrat i consistent d'accés transparent en aquestes bases de dades heterogènies esdevenia òbvia. L'any 1991, de manera totalment aliena a les necessitats de la biologia molecular, investigadors de l'Organització Europea per l'Energia Nuclear (CERN) inventaven el World Wide Web (WWW), un sistema *hipermedia* a Internet. Amb el desenvolupament l'any 1993 al National Center for Supercomputer Applications (NCSA) de Mosaic, un dels primers navegadors, WWW i, en conseqüència, Internet adquirien una difusió extraordinària. En pocs anys, Internet formaria part de la nostra vida quotidia-

na. Possiblement mai abans en la història de la humanitat, una nova tecnologia no s'havia imposat amb tanta celeritat. La infraestructura proporcionada per WWW a Internet resolia molts dels problemes d'accés, integració i anàlisi d'informació amb què s'enfrontava la investigació en biologia molecular cap a la meitat dels noranta. Internet permetia l'actualització constant de les bases de dades i WWW proporcionava als usuaris una interfície consistent per a l'accés, la consulta i fins i tot l'anàlisi d'aquestes dades. La tecnologia hipertext permetia la navegació transparent a través de bases de dades heterogènies i distants. Aprofitant aquesta tecnologia, dos sistemes integrats d'accés a les bases en biologia molecular a Internet, *Entrez* i *Sequence Retrieval System* (SRS), van ser desenvolupats gairebé simultàniament al National Center for Biotechnology Information (NCBI) i a l'Institut Europeu de Bioinformàtica (EBI), dues institucions públiques creades al final dels vuitanta i principi dels noranta als Estats Units i a Europa, respectivament, per fer front a les necessitats creixents de recursos computacionals de la recerca en biologia molecular. En cert sentit, hom podria dir que a partir de mitjan anys noranta una part important de la investigació en biologia molecular té lloc a Internet: Internet és el lloc on resideixen les dades i on es duen a terme les computacions. Certament, els projectes genòmics, tal com els entenem avui en dia —projectes cooperatius a escala mundial on la comunicació instantània entre productors de dades i entre aquests i els seus consumidors és essencial— no serien possibles sense Internet.

Cap a la meitat dels anys noranta, hom posava de manifest una nova vessant, fins a cert punt insospitada, de la relació cada cop més estreta entre biologia molecular i computació. L'any 1994, Adelman demostrava que el DNA té capacitat de calcular. Concretament, utilitzant tècniques estàndard de biologia molecular, va resoldre un problema típic en computació dels anomenats NP-complets: el de decidir si un graf és hamiltonià, és a dir, si existeix un camí que passa per tots els nodes del graf només una vegada. Tot i que la versió concreta del problema que va resoldre era molt senzilla, el fet que en la resolució intervinguessin només molècules de DNA i enzims implicats en el seu processament anticipava la possibilitat de construir ordinadors moleculars. Des d'aleshores s´han publicat dotzenes d'articles sobre el problema i hi ha qui manté que, en determinades aplicacions, els ordinadors basats en DNA poden constituir una alternativa viable als ordinadors digitals actuals.

Mentrestant, els projectes genòmics continuaven progressant. L'any 1992 es publicava la seqüència del cromosoma III de llevat (315.000 nucleòtids) — el primer cromosoma d'un organisme eucariota completament seqüenciat. L'any 1995, la companyia privada TIGR anunciava l'obtenció de la seqüència d'*Haemophilus influenzae* (1,8 milions de nucleòtids), el primer genoma no víric completament seqüenciat. Aquest fet posava de manifest l'interès de la indústria farmacèutica i biotecnològica en la recerca genòmica i, en conseqüència, en la biocomputació: un gran nombre de companyies realitzaven inversions —en alguns casos milionàries— per dotar-se de departaments de recerca i desenvolupament en bioinformàtica. L'any 1996 hom completava el genoma de llevat (dotze milions de nucleòtids), el primer d'un organisme eucariota; l'any 1997 el genoma d'*Escherichia coli* (4,5 milions de nucleòtids), al final de l'any 1998 el genoma de *Caenorhabditis elegans* (noranta-set milions de nucleòtids), el primer d'un organisme eucariota multicel·lular. En tots aquests casos, l'anàlisi computacional ha estat essencial per interpretar les seqüències d'aquests genomes, en particular per identificar-ne els

gens i per inferir la funció probable d'un gran nombre d'aquests. En el cas de *C. Elegans*, per exemple, els sis articles publicats a la revista *Science* que descriuen la seqüenciació i l'anàlisi del seu genoma contenen anàlisis computacionals, dos d'aquests articles depenen substancialment d'aquestes anàlisis i un d'ells —en el qual hom compara exhaustivament les proteïnes de llevat i les de *C. Elegans*— és exclusivament computacional.

Al principi de l'any 1999, una vintena de genomes, sense comptar els genomes vírics, han estat ja seqüenciats —la majoria d'organismes procariotes—, i l'etapa de seqüenciació massiva del projecte Genoma Humà ha començat. Durant els propers anys, milions, potser desenes de milions o, fins i tot, centenars de milions de nucleòtids seran seqüenciats diàriament arreu del món. Aquesta empresa afectarà de manera radical el desenvolupament de la biologia i tindrà un impacte extraordinari en la medicina, l'agricultura i molts processos industrials. El desenvolupament i l'aplicació de tècniques computacionals per a la col·lecció, emmagatzematge, anàlisi i interpretació de la informació evolutiva, estructural i funcional que les seqüències d'aquests genomes contenen —aquesta disciplina que he anomenat aquí biologia molecular computacional, però que hom també coneix com a bioinformàtica, biocomputació, informàtica del genoma o, simplement, anàlisi de seqüències— és part essencial d'aquesta empresa.

LA BIOLOGIA *IN SILICO*

Amb aquesta breu i incompleta revisió històrica de la creixent dependència durant la segona meitat del segle XX que la biologia molecular té de la computació, he volgut posar en relleu que la biologia del segle XXI és impensable sense la computació. Com escrivia John Maddox, l'antic editor de la revista *Nature*, l'any 1993, a l'alba dels projectes genòmics: «... La computació i la biologia molecular, ja interdependents, són a punt d'esdevenir inextricablement lligades... Els ordinadors són, cada cop més, un dels mitjans a través dels quals els problemes en biologia molecular poden ser resolts.»

Hom pot objectar que aquesta revisió és esbiaixada i que la computació ha estat i és rellevant en altres àmbits de la biologia diferents de la biologia molecular. Això és cert; la utilització de la computació té una llarga tradició en la recerca en biologia i biomedicina. L'anàlisi computacional i matemàtica ha contribuït notablement a l'avenç científic en camps tan diversos dins les ciències de la vida com l'ecologia, la genètica, l'antropologia, la fisiologia, la farmacologia... Segons la meva opinió, tanmateix, amb l'adveniment de les dades moleculars, la relació entre computació i biologia esdevé radicalment diferent. La peculiar naturalesa d'aquestes dades —seqüències de símbols— les fa particularment apropiades per a l'anàlisi computacional. El fet que les seqüències siguin per si mateixes portadores d'una enorme quantitat d'informació —si més no la que deriva del fet que seqüències similars exhibeixen usualment una funció i una història similars, una íntima i singular relació entre sintaxi i semàntica— fa aquesta anàlisi excepcionalment rellevant. En biologia molecular, els ordinadors, en conseqüència, no serveixen tant per modelitzar la realitat com per observar-la, analitzar-la i interpretar-la. És a dir, a diferència de la modelització matemàtica tradicional en biologia, on la realitat ha de ser sovint (extraordinàriament) simplificada per construir models simbòlics susceptibles de ser tractats matemàticament i computacionalment, en biologia molecular la realitat és intrínsecament simbòlica i l'ordinador és l'instrument mitjançant el qual aquesta realitat és observada sense intermediació. És per

això que en biologia molecular i genòmica, la computació no és només una eina per resoldre determinats problemes, sinó que molts problemes no poden ni tan sols ser plantejats si no és en termes computacionals. La biologia *in silico* emergeix amb la mateixa entitat que la biologia *in vivo* o que la biologia *in vitro*.

En el sentit restringit en el qual l'he entès aquí, la biologia computacional, com a disciplina cientificotecnològica recent i en rapidíssima evolució, manca encara de sistematització. Tanmateix, entre les àrees de recerca pròpies del seu domini en trobem algunes que han esdevingut gairebé clàssiques. Destaquen, entre d'altres, l'alineament i la similitud entre seqüències, els algoritmes exactes per l'alineament òptim de seqüències i l'estimació de la significació estadística de la similitud entre seqüències. Els algoritmes aproximats per a la recerca eficient de similituds en les bases de dades. El reconeixement de patrons en seqüències: la representació de senyals sintàctics i els algoritmes de recerca d'aquests senyals (*search by signal*). Els models estadístics de les seqüències i els algoritmes per la segmentació de seqüències en dominis d'acord amb aquests models (*search by content*). La reconstrucció de filogènies moleculars. La predicció de l'estructura de les proteïnes i dels àcids nucleics a partir de la seqüència. L'anàlisi global de genomes i l'anàlisi comparativa de genomes.

MIRANT CAP AL FUTUR

És difícil preveure quina serà l'evolució en el futur de la biologia computacional. Recordem que només vint-i-cinc anys després que, a causa de la complexitat tecnològica, hom considerés gairebé impossible la seqüenciació del DNA, s'ha obtingut la seqüencia de prop de cent milions de nucleòtids del genoma d'un animal multicel·lular. Així d'errònia pot arribar a ser la predicció dels esdeveniments científics. La història de la biologia molecular i la informàtica, a més, il·lustra fins a quin punt el progrés d'una disciplina científica depèn del desenvolupament d'altres disciplines aparentment no relacionades i, en conseqüència, la dificultat de predir-ne el desenvolupament sense considerar el desenvolupament global de la ciència. La història de la biologia molecular i la informàtica il·lustren d'altra banda, però, que tot allò que la humanitat imagina acaba, d'alguna manera, esdevenint realitat. L'obtenció de la seqüència del genoma humà no més tard de l'any 2005 serà la culminació d'un projecte de magnitud insòlita en biologia. Serà, però, al mateix temps, l'inici d'una nova manera de fer biologia. Després del genoma humà seguiran els genomes de moltes altres espècies. La comparació dels genomes a través de tot l'espectre filogenètic permetrà una reconstrucció més acurada de la història de la vida sobre la Terra. Aquesta és una història complexa i en la qual les recombinacions i les duplicaciones genòmiques han tingut un paper important. L'assumpció en què es basen els mètodes de reconstrucció filogenètica actuals, d'acord amb la qual similitud de seqüència implica proximitat evolutiva, haurà de ser flexibilitzada, i nous mètodes computacionals hauran de ser desenvolupats en els quals la proximitat entre genomes sigui funció no només de la similitud de la seva seqüència primària, sinò de la similitud d'elements estructurals de jerarquia superior, com ara l'ordre relatiu dels gens o la seva estructura exònica. La comparació dels genomes d'espècies properes permetrà definir aquells gens que defineixen una espècie a escala molecular. En particular, la comparació del genoma humà amb el genomes dels primats més pròxims permetrà identificar aquells (poquíssims) gens responsables de l'especi-

ficitat humana —la nostra manera peculiar de *viure en el món*.

A l'altre costat de l'espectre filogenètic, la comparació dels genomes de soques patògenes de microbis amb el d´aquelles que no ho són permetrà identificar els gens responsables d'aquesta patogenicitat, la qual cosa facilitarà, per exemple, el desenvolupament de nous antibiòtics. Amb els anys, allò que ara ha estat un costosíssim projecte internacional que s´ha desenvolupat durant més d'una dècada, esdevindrà un procés rutinari que hom durà a terme de manera automàtica en un xip minúscul: dels genomes de les espècies hom passarà a seqüenciar els genomes dels individus. Podrem quantificar, aleshores, l'aportació del component genètic en la individualitat i haurem d'encarar de manera més desapassionada la vella polèmica ambient *versus* herència. La seqüenciació del genoma dels individus generarà un volum de dades de magnitud gairebé incommensurable en relació al volum de dades que els projectes genòmics generen avui en dia, i que ja ens sembla difícilment tractable. Avenços en maquinari són, òbviament, imprescindibles per al tractament d'aquestes dades, però també són necessaris els avenços en programari: programes molt més eficients que permetin alinear la seqüència de cromosomes sencers, i prou sensibles per detectar de manera inequívoca variacions lleugeríssimes en aquesta seqüència.

Les dades moleculars són particularment rellevants quan poden ser correlacionades amb dades funcionals. És per això que bases de dades de funcions a diferents escales, com ara d'expressió gènica, de xarxes metabòliques, d'interacció entre proteïnes, de patrons de desenvolupament, d'efectes fenotípics etc., esdevindran cada cop més rellevants. Noves tecnologies informàtiques hauran de ser desenvolupades per tal d'integrar informació d'heterogeneïtat creixent i fer possible la inferència de generalitzacions a partir de les instàncies particulars acumulades a les bases de dades. En el cas del DNA, com que la funció a escala molecular és essencialment especificada en la seqüència de nucleòtids, la mineria de les bases de dades de seqüències (per *data mining* s'entén l'aplicació de tècniques informàtiques per extreure coneixement de manera més o menys automàtica de grans col·leccions sistemàtiques de dades, com ara bases de dades) pot contribuir a un millor coneixement dels mecanismes moleculars responsables de la seva funcionalitat, essencialment dels mecanismes implicats en els processos de transcripció (inclosa la seva regulació mitjançant la caracterització de les regions promotores dels gens), *splicing* i traducció. En aquest sentit, el recent descobriment i caracterització d'un nou tipus d'introns, els U12, que són processats de manera diferent per la maquinària de l'*splicing*, ha estat resultat d'una investigació essencialment computacional. D'una manera o d'una altra, doncs, la computació contribuirà a incrementar el nostre coneixement del funcionament dels organismes i, en particular, de la manera com la funció està codificada en la seqüència. La integració de dades funcionals a diferents escales i dades de seqüència farà possible en algun moment la simulació realista de la cèl·lula en l'ordinador. Podrem aleshores investigar en l'ordinador com els canvis en la seqüència es propaguen a l'organisme; en altres paraules, predir l'organisme a partir de la seqüència. La capacitat d'anticipar la naturalesa en fa possible la manipulació; en el cas dels genomes, l'enginyeria tindrà un component computacional essencial.

A causa de l'enorme volum i de la complexitat de les dades que genera la biologia, la utilització de l'ordinador esdevindrà imprescindible, tal i com ocorre en tantes altres disciplines científiques; però és, sobretot,

perquè la vida té a escala molecular un caràcter essencialment simbòlic, i perquè, en conseqüència, a escala molecular els processos de la vida són computacions en un sentit gairebè paradigmàtic, que biologia i computació romandran en el futur «inextricablement unides».

BIBLIOGRAFIA

BISHOP, M. J. Ed. (1998). *Guide to Human Genome Computing*. Nova York: Academic Press.

JUDSON, H. F. (1996). *The Eighth Day of Creation: Makers of the Revolution in Biology*. Nova York: Cold Spring Harbor Laboratory.

SHURKIN, J. N. (1996). *Engines of the Mind : The Evolution of the Computer from Mainframes to Microprocessors*. Nova York ; London: WW Norton & CO.

SMITH T. F. (1990). «The History of the Genetic Sequence Databases», *Genomics*, núm. 6, pàg. 701-707.

RECURSOS D'INTERNET

«Biocomputing Course Resource List: Course Syllabia». <http://www.techfak.uni-bielefeld.de/bcd/Curric/syllabi.html>

«A list of Bioinformatics Courses». <http://linkage.rockefeller.edu/wli/bioinfocourse.html>

NOTA A LA BIBLIOGRAFIA

En aquests moments hi ha un bon nombre de llibres sobre bioinformàtica al mercat i no se'n presenta un llistat exhaustiu. N'hi ha un de recollit, *Guide to Human Genome Computing,* que és molt emprat pels biòlegs moleculars que fan servir eines computacionals. Com que l'article té un fort component històric, hi ha una història de la biologia molecular (possiblement la més completa) i una història de la computació. La història del seu maridatge encara ha de ser escrita, tot i que l'article de Temple F. Smith és una crònica de primera mà del naixement d'aquesta relació. Tal com es reflecteix en l'article, el desenvolupament de la biologia molecular computacional es produeix (no de manera casual) paral·lelament al desenvolupament d'internet. A internet hi ha, literalment, milers de pàgines sobre bioinformàtica. Els dos llocs citats són simplement llistats de pàgines d'internet que contenen cursos de bioinformàtica. Alguns d'aquests cursos són introduccions excel·lents en aquesta disciplina.