Setzè Congrés de Metges i Biòlegs de Llengua Catalana
Llibre de Ponències (2000), p. 167-173

BIOINFORMÀTICA: LA INTERSECCIÓ ENTRE BIOLOGIA I COMPUTACIÓ

RODERIC GUIGÓ
Institut Municipal d'Investigació Mèdica. Universitat Pompeu Fabra. Barcelona

INTRODUCCIÓ

Som testimonis, en aquests anys, d'un canvi substancial en la manera com es practica la recerca en biologia. La biologia, tradicionalment una ciència descriptiva, ha passat a ser una ciència caracteritzada per la generació de gran quantitat d'informació. La invenció i el desenvolupament d'un nombre de tecnologies -al voltant de les quals s'articula aquesta nova disciplina científica que anomenem Genòmica- en són els responsables. L'automatització i robotització dels mètodes seqüenciació del DNA, que permeten l'obtenció de la seqüència completa dels genomes dels organismes vius o la seqüència parcial d'aquells, d'entre els gens codificats en aquests genomes, que s'expressen en una determinada estirp cel·lular, la invenció de les matrius de DNA (DNA *arrays*), les quals permeten, en particular, monitoritzar l'expressió simultània de milers de gens en condicions diferents, l'augment de precisió de les tècniques de gels bidimensionals i d'espectroscòpia de masses, que permeten la caracterització global de les proteïnes en què els gens són traduïts, les tècniques, com ara el sistema de dos híbrids en llevat, que permeten inferir globalment les interaccions entre aquestes proteïnes, la complexitat d'aquestes interaccions que sustenta els processos de la vida, i l'automatització i robotització de les tècniques de raigs X i de ressonància magnètica nuclear, que estan accelerant substancialment el desvetllament de les estructures tridimensionals d'un gran nombre d'aquestes proteïnes, en són possiblement els exemples més rellevants.

Els resultats que s'obtindran de l'aplicació d'aquestes tècniques transformaran profundament el nostre coneixement dels processos de la vida i tindran, possiblement, un impacte extraordinari en la medicina, en l'agricultura i en molts processos industrials. El resultat immediat de l'aplicació d'aquestes tècniques, però, és l'obtenció gairebé automàtica de quantitats immenses de dades, d'una magnitud insòlita en la història de la biologia. En aquest sentit, amb la genòmica, la biologia ha esdevingut en certa manera una ciència de la informació; tant en l'obtenció de les dades genòmiques primàries, com en el seu emmagatzematge, i la seva anàlisi, la informàtica juga, en conseqüència, un paper clau, i una nova disciplina científica, la bioinformàtica, en la intersecció entre biologia i computació ha emergit recentment per fer front a les especificitats que el tractament d'aquestes dades comporta.

En aquesta ponència voldria primer revisar la història de la vinculació creixent de la biologia amb la computació durant la segona meitat del segle XX, fins a principis dels anys noranta en què, amb el desenvolupament dels projectes genòmics, aquesta vinculació es fa indissoluble: la biologia ja no es pot pensar sense la computació. Voldria després destacar que durant la dècada dels noranta, Internet ha esdevingut el lloc on es desenvolupa part de la recerca en biologia: el lloc on resideixen les dades i on es duen a terme les anàlisis. Finalment, voldria emfatitzar que aquesta relació entre biologia i computació no es produeix només a causa de la «quantitat» d'informació que la recerca genòmica genera, sinó de manera molt més íntima, a partir de la mateixa naturalesa dels processos bàsics de la vida: amb el desenvolupament dels projectes genòmics culmina, d'alguna manera, el procés de reconeixement del caràcter simbòlic de la vida a nivell molecular, i del caràcter de computacions dels seus processos més bàsics.

INFORMÀTICA I BIOLOGIA MOLECULAR: UNA HISTÒRIA PARAL·LELA

Tot i que és amb la genòmica que la dependència de la biologia en la informàtica es fa més evident, la història d'aquestes dues disciplines durant la segona meitat d'aquest segle -des de l'eclosió de la informàtica i de la biologia molecular després de la segona guerra mundial -és, fins a cert punt, sorprenentment paral·lela i d'una vinculació creixent. Podríem, d'una manera una mica arbitrària, datar el naixement de la Informàtica i de la biologia molecular pràcticament al mateix temps -a finals dels anys quaranta, principis dels cinquanta- quan entren en funcionament els primers ordinadors digitals programables (veieu, per exemple, SHURKIN, 1996) i quan es desxifra l'estructura molecular del DNA (WATSON & CRICK, 1953). Anys més tard, però també gairebé al mateix temps -a finals dels anys cinquanta i principis dels seixanta-, mentre els informàtics dissenyaven els primers llenguatges d'alt nivell -mitjançant els quals podem especificar les instruccions que permeten la utilització dels ordinadors sense necessitat de conèixer el seu mecanisme de funcionament; és a dir, fan possible la separació entre programari i maquinari-, els biòlegs desxifraven l'anomenat codi genètic (veieu, per exemple, JUDSON, 1996), les instruccions d'acord amb les quals la seqüència de nucleòtids del DNA especifica la seqüència d'aminoàcids de les proteïnes. És amb aquests desenvolupaments paral·lels i independents que s'estableix la primera relació, conceptual si bé no encara operativa, entre biologia i computació. S'utilitzen aleshores per tal de descriure els processos bàsics de la vida, termes extrapolats de la Informàtica: codi, programa, instrucció, llenguatge, traducció, transcripció, desxiframent ..., termes alguns dels quals han passat a formar part del llenguatge bàsic de la biologia. I és, un altre cop, gairebé al mateix temps -a mitjans dels anys seixanta- que es produeixen dos fets en biologia i informàtica, la importància dels quals és avui indiscutida: d'una banda, el desenvolupament d'ARPANET, la xarxa d'ordinadors predecessora d'Internet, d'una altra, el desenvolupament de les tècniques de seqüenciació dels àcids nucleics (SANGER & COULSON, 1975; MAXAM & GILBERT, 1977). Dos esdeveniments que només ara, en reconèixer que el genoma com a entitat operativa existeix només a Internet, ens és possible de relacionar. Comença aleshores el procés de desxiframent de la seqüència de nucleòtids del DNA dels organismes vius, un procés el creixement del qual ha continuat des d'aleshores a un ritme exponencial. Ja a principis dels anys vuitanta i per fer-ne front a l'acumulació, es creen gairebé simultàniament als Estats Units i a Europa (i més tard, al Japó) les bases de dades que contenen les seqüències de tots els àcids nucleics coneguts. Gairebé el mateix any que IBM treu al mercat el primer ordinador personal, el popular PC i culmina una altra etapa d'aquest procés constant de miniaturització i increment de potència dels ordinadors digitals.

I és perquè els ordinadors han esdevingut suficientment potents, que l'acumulació de seqüències de DNA i de proteïnes esdevé rellevant. Els ordinadors permeten comparar les seqüències i, en comparar-les, es descobreix un fet fonamental: que seqüències semblants tenen també una funció o una història semblant, i que, per tant, 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. Aquest fet justifica un dels procediments que, fins a cert punt inadvertidament, s'ha demostrat més fructífer en la recerca en biologia molecular de les dues darreres dècades: el procediment que consisteix a comparar en un ordinador la seqüència d'un nou gen amb la seqüència dels gens prèviament coneguts, i inferir-ne la seva funcionalitat a partir de la similaritat potencial amb gens de funcionalitat coneguda. La rellevància d'aquest mètode es va posar de seguida de manifest: a principis dels vuitanta, hom va descobrir duent a terme recerques rutinàries en les bases de dades de seqüències, la similaritat entre un oncogen i un factor de creixement (Doolittle, 1983). Una relació que contribuïa substancialment a la comprensió dels mecanismes moleculars involucrats en el càncer. La importància de la computació en el progrés del coneixement dels processos de la vida esdevenia indiscutible.

És així, doncs, que el problema de determinar el grau de similaritat entre dues seqüències esdevé clau en la recerca en biologia molecular. Un problema que implicava dades de naturalesa peculiar -seqüències de símbols- i enfront del qual els mètodes estadístics i de modelització matemàtica utilitzats comunament en altres àrees de la biologia es mostraven limitats. És en resposta a aquest problema que durant els setanta i a principis dels vuitanta hom elabora els primers algoritmes per a l'alineament, comparació i determinació del grau de similaritat de dues seqüències (Needelman & Wunsch, 1970; Sellers, 1974; Smith & Waterman, 1981). En cert sentit podríem dir que és amb el desenvolupament d'aquests algoritmes que neix la bioinformàtica: en el moment en el qual les tècniques estadístiques o matemàtiques conegudes deixen de ser suficients per adreçar els nous problemes que les noves dades moleculars plantegen, i es fa necessari desenvolupar noves tècniques matemàtiques, algorítmiques i computacionals. Així durant els anys vuitanta, nombroses tècniques computacionals van ser desenvolupades per tal de tractar de manera eficient l'acumulació de dades moleculars. En particular, van ser desenvolupats mètodes 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 d'*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. Amb el desenvolupament d'aquestes tècniques, però sobretot, amb el desenvolupament de mètodes més precisos i eficients per fer recerques de similaritat en las bases de dades de bioseqüències (Pearson & Lipman, 1987;Altschul *et al.*, 1990), la bioinformàtica consolidava el seu domini de recerca.

Arribem als anys noranta, quan comença «oficialment» el Projecte del genoma humà. 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 de la seva concepció 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, *software* 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 juguen 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".

A principis dels noranta, però, no era només en la investigació genòmica 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 nivells. 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 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, hom inventava la *World Wide Web* (WWW), un sistema hipermèdia a Internet. Amb la WWW, Internet adquiria una difusió extraordinària. En particular, 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 a mitjans 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 RECERCA GENÒMICA A INTERNET

En cert sentit, des de mitjans dels anys noranta, bona part 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 nivell mundial on la comunicació instantània entre productors de dades i entre aquests i els seus consumidors és essencial— no serien possibles sense Internet. En aquests moments, el genoma de desenes d'organismes procariotes, i d'uns pocs eucariotes ha estat seqüenciat i aviat ho serà la seqüència del genoma humà. Per a la majoria d'investigadors, però, aquests genomes —seqüències de milers de milions de lletres, en alguns casos—, només existeixen a Internet. És a Internet on s'accedeix i és a Internet on es

duen a terme bona part de les anàlisis que permeten transformar aquestes seqüències interminables de lletres en coneixement rellevant sobre els processos de la vida.

Així, en un escenari típic en aquests moments, suposem que com a investigadors hem deduït la seqüència d'aminoàcids d'un fragment d'un gen humà en el qual estem interessats. Possiblement, allò que ens interessaria primer és localitzar la regió en el genoma humà en què aquest gen es troba, per tal de caracteritzar-lo completament. Podem accedir directament a un centre de seqüenciació genòmica, al Sanger Center per exemple (http://www.sanger.ac.uk) i utilitzant un programa com ara TBLASTN comparar la seqüència parcial del nostre gen amb tota la seqüència del genoma humà. Un cop el fragment del genoma que codifica el gen ha estat identificat, podrem utilitzar alguns programes de predicció de gens per tal de completar i refinar l'estructura exònica del gen d'interès. Podríem accedir, per exemple, a Metagene (http://ares.ifrc.mcw.edu/MetaGene/) un servidor www que distribueix seqüències problema a un nombre de servidors www que implementen programes de predicció de gen, i que n'integra els resultats. Amb l'estructura del gen refinada i la seqüència completa d'aminoàcids deduïda, voldríem segurament obtenir informació sobre la funció probable d'aquest gen. La manera habitual de fer-ho és comparar la seqüència d'aminoàcids del nostre gen amb les seqüències d'aminoàcids de tots els gens coneguts. Les bases de dades que compilen totes les seqüències d'aminoàcids conegudes estan actualitzades diàriament i s'hi pot accedir, entre d'altres, a través de l'Institut Europeu de Bioinformàtica (EBI, http://www.ebi.ac.uk) i al National Center for Biotechnology Information (NCBI) dels National Health Institutes (NIH) als Estats Units (http://www.ncbi.nlm.nih.gov). Per tal de comparar la seqüència del gen d'interès amb la seqüència de tots els gens coneguts, hom pot fer servir programes de diversa complexitat. Si la similaritat amb gens de funció coneguda és molt alta, comparacions simples a nivell de seqüència primària són suficients, però si la similaritat és debil caldrà dur a terme recerques de similaritat iteratives, o utilitzar bases de dades específiques de motius funcionals (com ara PROSITE, per exemple, http://www.expasy.ch/prosite/). Un cop hem reconegut un conjunt de seqüències relacionades amb la proteïna codificada pel nostre gen —la família a la qual el nostre gen pertany— és habitual realitzar un alineament múltiple de totes aquestes seqüències per tal de localitzar els residus conservats en totes elles i que són, probablement aquells residus implicats en la funcionalitat característica dels membres d'aquesta família. Podem utilitzar, per exemple, el programa CLUSTALW, del qual existeixen multitud de servidors (un d'ells és http://bioweb.pasteur.fr/seqanal/interfaces/clustalw-simple.html). En ocasions, la seqüència primària no és suficient per inferir-hi funcionalitat i cal recórrer a comparacions a nivell estructural. El primer pas és realitzar una predicció d'estructura secundària (hi ha un gran nombre de programes per dur a terme aquesta tasca. En podem trobar alguns, per exemple PredictProtein, http://www.embl-heidelberg.de/predictprotein/predictprotein.html). Caracteritzada l'estructura secundària, podem dur a terme una predicció de la classe estructural a la qual la proteïna té més probabilitats de pertànyer (el servidor PSA, http://bmerc-www.bu.edu/psa/index.htm, n'es ún exemple). D'altra banda, potser estem interessats a caracteritzar la regió promotora del gen, per tal d'investigar-ne la regulació de la seva expressió. Podríem, per exemple, intentar localitzar els elements promotors en la regió 5' del gen, analitzant-ne la seqüència a la base de dades TRANSFAC de factors de transcripció i elements promotors (http://transfac.gbf.de/TRANSFAC/). El resultat de les anàlisis ens proporcionaria informació sobre els factors de transcripció, hipotèticament implicats en la regulació de l'expressió del nostre gen.

És a dir, sense moure'ns de l'ordinador amb la infraestructura que proporciona Internet, i

només a partir de la seqüència primària, podem obtenir gran quantitat d'informació i de coneixements sobre el gen en el qual estem interessats. Aquesta informació ens pot ajudar a plantejar hipòtesis i a dissenyar de manera més eficaç els nostres experiments.

VIDA I COMPUTACIÓ

Voldria, per acabar, emfatitzar que amb la genòmica, la importància de la computació en la biologia no prové només del fet que l'enorme volum i la complexitat de les dades que les tecnologies desenvolupades al voltant d'aquella disciplina generen, i que fan que la utilització de l'ordinador esdevingui imprescindible, com, d'altra banda ocorre, avui en dia, amb tantes disciplines científiques. Amb la genòmica, la relació entre biologia i computació no es fonamenta només en la «quantitat» de dades, sinó que s'estableix de manera més íntima i radicalment diferent, a partir de la naturalesa peculiar d'informació genòmica primària: la seqüència de nucleòtids del DNA. La peculiar naturalesa d'aquesta informació —seqüències de símbols— la fa particularment apropiada 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 genòmica, 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 tal de construir models simbòlics susceptibles de ser tractats matemàticament i computacionalment, en genòmica 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 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. En definitiva, amb la genòmica culmina el procès de reconeixement que la vida té, a nivell molecular, un caràcter essencialment simbòlic, i que, en conseqüència, a nivell molecular, els processos de la vida són computacions en un sentit gairebé paradigmàtic.

Tot i que en aquesta ponència la revisió de les relacions entre biologia i computació ha estat esbiaixada en una direcció (la biologia depenent de la computació), aquest caràcter intrínsecament «computacional» de la vida fa aquesta relació efectivament bidireccional. Així, a mitjans dels anys noranta hom demostrava en la pràctica que el DNA té efectivament capacitat de calcular (ADELMAN, 1994). Des d'aleshores, la investigació en la possibiltat de construir ordinadors basats en DNA no s'ha aturat i hi ha qui manté que, en determinades aplicacions, aquests ordinadors poden constituir una alternativa viable als ordinadors digitals actuals. És cert, doncs, biologia i computació són ja, i romandran en el futur, inextricablement unides.

BIBLIOGRAFIA

ADELMAN, L. (1994) "Molecular Computation of Solutions to Combinatorial Problems," *Science*, 266:1021-1023.

ALTSCHUL, S. F. *et al.* (1990) "Basic Local Alignment Search Tool". *Journal of Molecular Biology*,215:403-410.

DOOLITLE, R. F. *et al.* (1983) "Simian sarcoma virus onc gene, v-sis, is derived from the gene (or genes) encoding a platelet-derived growth factor". *Science*, 221:275-277.

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

MAXAM, A. M.; W. GILBERT (1977) "A new method for sequencing DNA". *Proc. Natl. Acad. Sci. USA*, 74: 560-564.

NEEDELMAN, S. B.; C. D. WUNSCH (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.

PEARSON, W. R.; D. J. LIPMAN (1987) "Improved tools for biological sequence comparison". *Proc. Natl. Acad. Sci. USA*, 85: 2444-2448.

SANGER, F.; A. R. COULSON (1975) "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase". *Journal of Molecular Biology*, 94: 441-448.

SELLERS, P. H. (1974) "On the theory and computation of evolutionary distances" *SIAM J. Appl. Math*, 26: 787-793.

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.; M. S. WATERMAN (1981) "Identification of Common Molecular Subsequences". *Journal of Molecular Biology*, 147: 195-197.

WATSON, J. D.; F. H. C. CRICK (1953) "A Structure for Deoxyribose Nucleic Acid". *Nature*, 171: 737-738.