Bioinformàtica: una eina essencial per als biòlegs del segle xxi

Marc A. Marti-Renom

Structural Genomics Group, CNAG-CRG, The Barcelona Institute of Science and Technology, Institució Catalana de Recerca i Estudis Avançats

Correspondència: Marc A. Marti-Renom. Centre Nacional d'Anàlisi Genòmica. Carrer de Baldiri i Reixac, 4. 08028 Barcelona. Adreça electrònica: *mmarti@pcb.ub.cat*.

DOI: 10.2436/20.1501.02.153
ISSN (ed. impresa): 0212-3037
ISSN (ed. digital): 2013-9802
http://revistes.iec.cat/index.php/TSCB
Rebut: 07/11/2014
Acceptat: 07/04/2015

Resum

La bioinformàtica i la biologia computacional són dues cares d'una mateixa disciplina, que fa ús de tècniques i algoritmes informàtics per estudiar processos biològics. Ambdós termes tenen, però, matisos diferents. Històricament, el terme de *biologia computacional* s'ha emprat més en ambients de simulació computacional en què la física i les matemàtiques tenen un paper molt rellevant, mentre que el terme de *bioinformàtica* s'ha usat més en l'estudi de grans quantitats de dades en què l'estadística és essencial. Assignem el terme que assignem, l'augment de les dades òmiques juntament amb una comprensió més acurada dels processos estudiats, ha fet que la bioinformàtica resulti imprescindible en el currículum del biòleg. Com a conseqüència, en aquest segle que tot just estrenem, no tots els biòlegs hauran de ser bioinformàtics, però sí que hauran de saber fer servir la bioinformàtica.

Paraules clau: bioinformàtica, biologia computacional, simulació, modelització.

Bioinformatics: an essential tool for biologists of the 21st century

Summary

Bioinformatics and *computational* biology are parallel terms that refer to the application of computational science to the study of biological processes. However, these two terms have certain historical connotations. *Computational biology* has referred more specifically to a scientific approach to simulating biology in which physics and mathematics play an important role. *Bioinformatics* is often used to designate a discipline in which computational software allows the vast amounts of data now produced by biologists to be processed. Regardless of which term is used, the increase of biological data and the more accurate insights into the theory behind biological processes make bioinformatics essential for any biologists of the 21st century. Consequently, today not all biologists need to be bioinformaticians but they certainly all need to be proficient in bioinformatics.

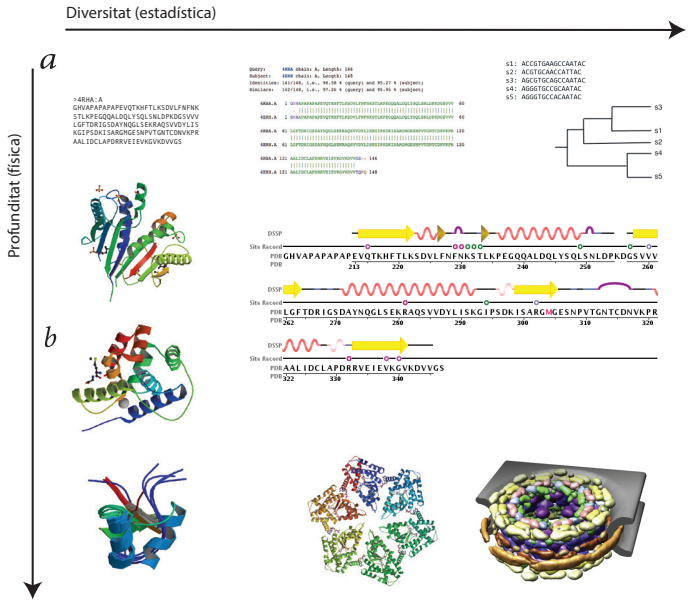
Keywords: bioinformatics, computational biology, simulation, modeling.

Un abans i un després de la bioinformàtica en l'era òmica

Al principi dels anys setanta del segle xx, Paulien Hogeweg i Ben Hesper van començar a usar el terme *bioinformàtica* per referir-se a «l'estudi de processos informàtics en sistemes biòtics». Aquesta definició original es referia més a un camp concret de la biologia que estudia com els sistemes biològics produeixen, usen i emmagatzemen informació (Hogeweg, 2011). Aquesta definició original contrasta amb la que podem trobar avui dia en l'*Enciclopèdia Catalana*: «Disciplina científica que aplica els principis de la informàtica a l'estudi de fenòmens biològics», o en l'*Oxford English Dictionary*: «The science of collecting and analysing complex biological data such as genetic codes». Aquestes aparents diferències en la definició del terme *bioinformàtica* ens donen una idea de com en els darrers quaranta anys la producció de grans quantitats de dades biològiques ha fet canviar aquest camp de la biologia.

Durant les dècades dels setanta i vuitanta, no es podia disposar de les quantitats de dades que l'era òmica portaria en el nou segle. Eren, doncs, anys en què els bioinformàtics (o biòlegs computacionals) feien ús de les eines informàtiques per estudiar molècules concretes d'organismes específics. Un dels camps de més producció en aquells moments va ser l'estudi del plegament de macromolècules, i en particular de proteïnes. Aquests estudis es feien principalment amb mètodes físics en què les interaccions entre els àtoms d'una molècula es tractaven a partir dels principis fonamentals de Newton (Brooks, 1995; Elber, 1996). Aquestes simulacions, però, eren (i són) molt costoses en hores de computació i, per tant, aplicables sols a un nombre limitat de macromolècules a la vegada. Aquestes limitacions de càlcul es van abordar a partir de l'ús de l'estadística per estudiar el plegament de proteïnes. Nous mètodes comparatius aprofitaven el creixement en dades al Protein Data Bank (PDB, Berman *et al*., 2000) per tal d'extreure estadístiques de conservació de l'estructura i la seqüència de proteïnes (Sali i Overington, 1994). Al mateix temps, els mètodes comparatius de seqüència començaven a ser suficientment ràpids per ser aplicats a gran escala, en què no sols es comparaven dues seqüències d'àcid desoxiribonucleic (DNA) o proteïnes, sinó que ja es podien fer cerques d'homòlogues contra bancs de seqüències (Altschul *et al*., 1990). Aquesta dicotomia, ja aparent en els anys noranta, entre la biologia computacional d'estructura basada principalment en la física i la bioinformàtica de seqüència basada en estadística, ens acompanyarà fins a l'aparició de mètodes híbrids en què la integració de dades i d'aproximacions per a l'estudi de sistemes complexos és primordial (vegeu la figura 1).

De totes maneres, en aquelles dècades, la visió òmica de la biologia era pràcticament inexistent fins que a mitjans anys noranta comencen a aparèixer els estudis genòmics en molts dels organismes



↑ Figura 1. Bioinformàtica entre la física i l'estadística. La bioinformàtica ha fet ús de la física i de l'estadística per abordar problemes complexos. La física ha estat principalment emprada per estudiar l'estructura de macromolècules i complexos d'aquesta. L'estadística ens ha ajudat a entendre la diversitat de les seqüències i estructures de macromolècules, que és essencial per alinear-les i inferir relacions d'homologia evolutiva. L'ús de mètodes físics per a l'estudi de l'estructura s'ha associat clàssicament al terme biologia computacional. L'ús de l'estadística per a l'estudi de la diversitat de seqüència s'ha associat clàssicament al terme bioinformàtica. Mètodes híbrids recents que integren dades diverses per estudiar complexos de macromolècules fan ús combinat de l'estadística i de la física. Figura adaptada de Luscombe *et al.*, 2001.

model usats en el laboratori. De fet, els primers articles de l'era òmica apareixen al final dels anys noranta (vegeu la taula 1). Una excepció d'aquesta regla és l'estudi del contingut genètic d'un organisme (o la genòmica), terme que apareix en articles ja en els anys quaranta, però que no és fins a la introducció del Projecte del Genoma Humà (1984) que pren una força rellevant. De fet, a partir de l'ús del terme *genòmica* altres camps de la biologia adopten el sufix *-òmic* per referir-se a l'estudi complet de les característiques dins un sistema biològic (vegeu la taula 1 per a una llista d'òmiques i el seu impacte).

En el segle xxi, la bioinformàtica ja es pot considerar adulta (Ouzounis, 2012) i està present en la majoria dels laboratoris moderns en biologia, ja sigui perquè tenen bioinformàtics immersos en el laboratori o per la construcció de ponts de col·laboració amb grups de recerca en bioinformàtica. De fet, sembla que el moment de «moda» de la bioinformàtica ha passat i que ens trobem davant d'una disciplina de la biologia en un estat d'estabilització. Comparativament, el terme *bioinformatics* és cercat a Google d'una manera molt similar a termes com *molecular biology* o *cell biology* (vegeu la figura 2). Aquest fet indicaria que, igual com amb termes associats amb òmiques (vegeu la figura 2), a partir dels anys 2009-2010 s'estabilitza «l'interès» de la població, la qual percep notícies associades a la bioinformàtica amb més normalitat.

Sigui com sigui, i gràcies a la bioinformàtica, avui dia podem fer recerca explorant sistemàticament la bibliografia (*text mining*), dissenyar nous fàrmacs (informàtica química), estudiar el plegament de proteïnes (biologia estructural computacional), o alinear milions de seqüències per cercar mutacions associades a malalties (genòmica informàtica), entre d'altres. En totes aquestes aplicacions, la bioinformàtica afrontarà tres grans reptes o objectius, que descric en la secció següent.

Objectius de la bioinformàtica

A grans trets, la bioinformàtica té tres objectius (Luscombe *et al.*, 2001): *a*) organitzar i emmagatzemar dades biològiques, *b*) desenvolupar eines informàtiques per estudiar i analitzar les dades emmagatzemades, i *c*) extreure informació biològicament rellevant de les dades.

Organització de les dades biològiques

Gràcies al concepte de *relació comparativa* i a la conservació evolutiva de les macromolècules, els bioinformàtics hem emmagatzemat i organitzat les dades biològiques basant-nos en la seva similitud. Per exemple, està demostrat que evolutivament la funció de dues proteïnes està més conservada que la seva estructura i, a la vegada, la seva estructura està més conservada que la seva seqüència (Chothia i Lesk, 1986; Rost, 1999). Tant és així, que es calcula que els milions de seqüències proteiques conegudes adopten poc més d'un miler d'estructures diferents (Yeats *et al.*, 2006). Gràcies a aquesta conservació evolutiva, juntament amb l'aparició de programari que ens permet comparar seqüències, es comença a classificar i organitzar el coneixement de macromolècules biològiques. De fet, dos programes com el BLAST (Altschul *et al.*, 1990) o el CLUSTAL (Higgins i Sharp, 1988), desenvolupats per comparar les seqüències biològiques d'una manera ràpida i acurada, es troben dins dels cent articles més referenciats en la història de la ciència (Noorden *et al.*, 2014). Tot i que es coneix molt menys l'estructura de les macromolècules que la seva seqüència, un dels primers bancs de dades biològiques correspon al PDB que, des del principi dels setanta, emmagatzema i organitza totes les dades d'estructures de proteïnes i d'àcids nucleics (Berman *et al.*, 2002). Similar a l'estructura de proteïnes, la gran quantitat de seqüències conegudes (per exemple, el banc de dades Universal Protein Resource (UniProt Consortium, 2010) conté actualment més de noranta milions de seqüències) ens ha permès classificar al voltant de quinze mil famílies de proteïnes a Pfam (Finn *et al.*, 2014). El PDB, UniProt o Pfam són exemples clars d'un camp de la bioinformàtica que té una rellevància cabdal per organitzar la gran quantitat de dades biològiques. El darrer exemplar de *Nucleic Acids Research* (NAR) sobre bancs de dades recull més de 1.550 bancs de dades que es poden consultar a *http://tinyurl.com/m5v9fw8* (Fernandez-Suarez *et al.*, 2014).

Desenvolupar eines per a l'estudi i l'anàlisi de dades biològiques

Qualsevol macromolècula en biologia es pot analitzar per la seva seqüència, estructura o funció. És precisament en aquests tres aspectes que els bioinformàtics hem desenvolupat eines per a l'estudi de les dades biològiques. Aquestes eines es basen en la física, l'estadística i els principis evolutius per comparar, classificar, simular i modelitzar macromolècules biològiques. Així mateix, aquests mètodes computacionals s'especialitzen en el tipus de macromolècula que estudien. Per exemple, els bancs de seqüències genòmiques, especialment les referents a l'àcid ribonucleic (RNA), s'han analitzat per tal de poder discernir entre regions codificants de proteïnes d'aquelles que no en codifiquen. Recentment, una gran quantitat de treballs apunten que l'RNA no es limita sols a la transferència d'informació entre el gen i la proteïna. Això ha fet que hi hagi un recent augment de programari que

adreça moltes qüestions sobre l'RNA i les seves múltiples funcions en la cèl·lula. Finalment, les proteïnes han estat clàssicament estudiades per la seva estructura, funció i interacció amb altres macromolècules o petites molècules (compostos químics). Tot plegat, ha fet que hi hagi actualment una llarga llista de mètodes computacionals que, de nou, la revista *Nucleic Acids Research* cataloga en el seu volum anual de *Web Servers* (Benson, 2014). Actualment, hi ha centenars de servidors a Internet que serveixen per estudiar les múltiples facetes de les macromolècules biològiques. Aquest servidors es poden consultar al directori d'enllaços bioinformàtics del Canadà: *http://bioinformatics.ca/links_directory*. De totes maneres, el lector ha de ser avisat que l'avanç dels mètodes computacionals, juntament amb el finançament irregular de les iniciatives per desenvolupar aquests mètodes, fa que aproximadament el 10 % dels servidors tinguin una existència efímera a la Xarxa (Schultheiss *et al*., 2011).

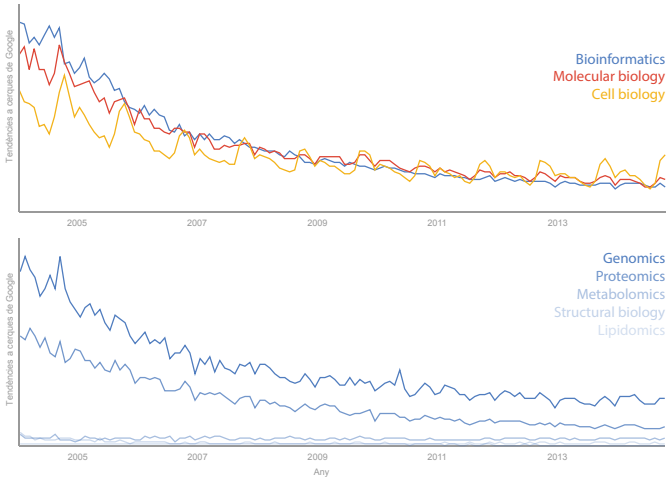
Extracció d'informació biològicament rellevant

Finalment, un cop hem emmagatzemat i estudiat les dades de macromolècules, s'ha de dur a terme una anàlisi exhaustiva per extreure'n informació de rellevància biològica. Aquest darrer punt és el menys específic dels tres grans objectius de la bioinformàtica. La pregunta de «què és biològicament rellevant?» té mil respostes, depenent de l'interlocutor que busca respondre a preguntes específiques. Aquestes poden anar des d'assignar la funció a una nova estructura resolta, ja sigui per mètodes experimentals o computacionals, fins a identificar quina combinació de factors de transcripció és la responsable de la transdiferenciació cellular. És doncs, aquest tercer aspecte de la bioinformàtica, en què es requereix un coneixement de la biologia més profund i en què l'aspecte més pràctic (o menys tècnic) pren rellevància.

El bioinformàtic del segle xxi

El camp de la bioinformàtica ha estat sempre un camp interdisciplinari en què es mesclen investigadors amb carreres tan dispars com la informàtica, les matemàtiques, la química, la física o la biologia, entre d'altres. En el nostre país, actualment les universitats UPF, UAB, UOC

Figura 2. *Google trends* (*http://google.com/trends*) per a termes associats a la bioinformàtica. La gràfica superior mostra la comparativa de mitjana de cerques amb termes com *bioinformatics*, *molecular biology* i *cell biology* des de 2005. La gràfica inferior mostra la comparativa de termes associats a mètodes òmics..



Taula 1. Òmiques. Taula adaptada i actualitzada de la pàgina web del Gerstein Group (*http://tinyurl.com/owh52bv*). Dades a la taula extretes de PubMed i de Google durant el novembre de 2014.

Terme	Descripció	Google	Pubmed	Any 1a publicació
Genome	Contingut genètic d'un organisme	65.500.000	925.543	1943
Proteome	Contingut proteic d'un organisme	7.320.000	32.918	1998
Transcriptome	Contingut de mRNA en un organisme	3.300.000	22.616	1997
Phenome	Identificació qualitativa de la forma i funció derivats dels gens	647.000	295	1995
Interactome	Llista d'interaccions de totes les macromolècules en una cèl·lula	487.000	1.767	1991
Metabolome	Contingut de compostos químics i petites molècules en un organisme	426.000	4.546	1998
Secretome	Contingut de productes gènics secretats de la cèl·lula	241.000	1.208	2000
Orfeome	Contingut de tots els ORF en un genoma	213.000	104	2002
Glycome	Contingut de molècules de carbohidrats en una cèl·lula	85.700	353	2000
Physiome	Descripció de les condicions fisiològiques en un organisme	76.600	133	1998
Regulome	Descripció de la xarxa de regulació gènica en una cèl·lula	25.600	44	2002
Fluxome	Descripció del contingut proteic i els seus fluxos	20.200	49	1999
Morphome	Descripció quantitativa de l'anatomia, bioquímica i composició química d'un organisme	18.600	4	2000
Translatome	Contingut de mRNA en un organisme ponderat pels seus nivells	13.200	62	2001
Cellome	Contingut i interaccions de totes les molècules en una cèl·lula	8.660	46	2002
Pseudome	La població de pseudogens en un genoma	7.750	0	–
Operome	La caracterització de proteïnes sense funció coneguda	7.750	1	2002
Functome	La població de productes gènics classificats per la seva funció	5.610	2	2001
Transportome	La població de productes gènics que són transportats	5.170	18	2004
Localizome	Localització de tots els productes gènics en una cèl·lula	3.090	10	2002
Foldome	Descripció dels productes gènics classificats per la seva estructura	2.810	1	2009
Unknome	Tots els gens de funció desconeguda	2.060	0	–
Ribonome	Llista de regions del genoma que codifiquen amb RNA complexos ribonucleoproteics	1.520	10	2002

i UVic ofereixen màsters o postgraus en bioinformàtica. Aquests màsters o postgraus aviat s'acompanyaran de graus en bioinformàtica gràcies a iniciatives com el Bioinformatics Barcelona (BIB, *http://bioinformaticsbarcelona.eu*). De totes maneres, i com deia en el resum al principi d'aquest article, no tots els biòlegs hauran de ser bioinformàtics però sí que hauran de conèixer la bioinformàtica.

Els tres objectius principals de la bioinformàtica (és a dir, emmagatzemar i organitzar les dades, analitzar les dades, i extreure conclusions biològiques rellevants) requereixen perfils de biòlegs/informàtics diferenciats. Des del punt de vista més tècnic, es requeriran bioinformàtics amb coneixements computacionals alts per tal de construir bancs de dades eficients que ens permetin accedir a les dades ràpidament o fins i tot en temps real. D'altres bioinformàtics necessitaran coneixements

en programació per desenvolupar eines eficients per comparar, classificar i analitzar les dades. Aquest segon gran grup de bioinformàtics, però, haurà de tenir uns coneixements de biologia més fonamentats per tal d'interpretar les dades experimentals correctament. Finalment, no haurien d'oblidar que les dades que analitzem provenen d'experiments sobre macromolècules biològiques de les quals hem de conèixer com funcionen i com es relacionen. És principalment en aquest tercer objectiu en què un biòleg o un bioinformàtic, per fer un estudi complet, haurà de conèixer bé l'ús d'eines bioinformàtiques per extreure el màxim profit de les dades.

Agraïments

Aquest article ha estat inspirat per lectures de revisions històriques del camp de la bioinformàtica, incloent, entre altres, la sèrie de *Bioinformatics Roots* de la revista *PLOS Computational Biology* (*http://www.ploscollections.org/rootsofbioinformatics*) i un parell d'articles excellents del grups de Mark Gerstein (Luscombe *et al.*, 2001) i Christos A. Ouzounis (Ouzounis, 2012).

Bibliografia

Altschul, S. F. [*et al.*] (1990). «Basic local alignment search tool». *J. Mol. Biol.*, 215 (3): 403-410.

Benson, G. (2014). «Editorial. Nucleic *Acids Research* annual *Web Server* Issue in 2014». *Nucleic Acids Res.*, 42 (*Web Server* issue): W1.

Berman, H. M. [*et al.*] (2000). «The Protein Data Bank». *Nucleic Acids Res.*, 28 (1): 235-242.

Brooks, C. L. 3rd. (1995). «Methodological advances in molecular dynamics simulations of biological systems». *Curr. Opin. Struct. Biol.*, 5 (2): 211-215.

Chothia, C.; Lesk, A. M. (1986). «The relation between the divergence of sequence and structure in proteins». *Embo J.*, 5 (4): 823-826.

Elber, R. (1996). «Novel methods for molecular dynamics simulations». *Curr. Opin. Struct. Biol.*, 6 (2): 232-235.

Fernandez-Suarez, X. M. [*et al.*] (2014). «The 2014 *Nucleic Acids Research Database* issue and an updated NAR online Molecular Biology Database Collection». *Nucleic Acids Res.*, 42 (*Database* issue): D1-6.

Finn, R. D. [*et al.*] (2014). «Pfam: the protein families database». *Nucleic Acids Res.*, 42 (*Database* issue): D222-230.

Higgins, D. G.; Sharp, P. M. (1988). «CLUSTAL: a package for performing multiple sequence alignment on a microcomputer». *Gene*, 73: 237-244.

Hogeweg, P. (2011). «The roots of bioinformatics in theoretical biology». *PLoS Comput Biol.*, 7 (3): e1002021.

Luscombe, N. M. [*et al.*] (2001). «What is bioinformatics? A proposed definition and overview of the field». *Methods Inf. Med.*, 40 (4): 346-358.

Noorden, R. [*et al.*] (2014). «The top 100 papers». *Nature*, 514 (7524): 550-553.

Ouzounis, C. A. (2012). «Rise and demise of bioinformatics? Promise and progress». *PLoS Comput Biol.*, 8 (4): e1002487.

Rost, B. (1999). «Twilight zone of protein sequence alignments». *Protein Eng.*, 12 (2): 85-94.