Bioinformàtica: on es gira la truita!

Oscar Conchillo Solé
Institut de Biotecnologia i Biomedicina Laboratori de Bioinformàtica Grup de Biologia Computacional
Universitat Autònoma de Barcelona

Proteïnes i ADN dues paraules que tot sovint sentim a les noticies, les primeres catalitzen tots els processos químics responsables de que ens autoanomenem éssers vius, el segon conté tota la informació necessària per a la construcció de les primeres.

Aquí parlaré d'una manera planera i senzilla sobre com els científics estudien les seqüències i estructures d'ambdues macromolècules i com *Linux* i el *Programari lliure* estan ajudant a que això sigui possible. Em focalitzaré sobretot en les proteïnes, doncs son aquestes l'objectiu principal del meu treball.

INTRODUCCIÓ

Tot i que per a molts *Linux* és un sistema operatiu pensat per informàtics per a informàtics i que mai podrà imposar-se en un món dominat pel "monopoli" *Microsoft* m'agradaria presentar aquí un món on les coses són tot just el contrari.

La Bioinformàtica és una ciència nova, una ciència nascuda com a tal durant els anys 80, una ciència on el "*grep*", el "*FORTRAN*", el "*CC*", el "*vi*" i les famoses "*GL*" de *Silicon Grafics* eren les eines bàsiques amb les quals acabar-ho fent tot.

Per altra banda, és també una ciència que, tot i que ara comença a tenir més difusió, ha estat condemnada als soterranis i habitacions fosques de la ciència durant molts anys (si més no en aquest país). No és estrany doncs que si sumem aquestes dues característiques obtinguem aquest resultat final: *LINUX*!

Linux i el Programari Lliure han portat la "supercomputació en *clusters"* assequible, la possibilitat de muntar laboratoris amb suficients estacions de treball per a tots els estudiants que les necessitin, la possibilitat d'utilitzar eines i "software" que si s'haguessin hagut de pagar mai hauríem arribat a veure.

La nostra ciència té moltes branques, però totes tenen una cosa en comú: els ordenadors. Ells son les nostres pipetes, les nostres provetes, els nostres cultius. Ells són el lloc on desenvolupar els nostres experiments, que ja no són ni "*in vivo*" ni "*in vitro*", sinó "*in silico*".

Sentim cada dia a les noticies parlar del Genoma humà, de la Genòmica i Proteòmica i dels avenços que això comportarà, però ... que és tot això? El Genoma humà no es més que pàgines i pàgines i més pàgines on el seu únic contingut són paraules cada una de les

quals conté entre 1000 i 10000 lletres, sense espais i amb molt poca varietat, doncs només la A, la T, la G i la C formen aquestes paraules. Una proteïna o complex proteic pot contenir entre 300 i milions d'àtoms, cada un dels quals amb tres coordenades cartesianes si en volem fer una representació tridimensional. Evidentment tot això són dades, moltes dades, masses perquè ningú les pugui processar sense l'ajuda d'ordenadors, una eina indispensable, però només una eina. Al igual que en la ciència experimental clàssica calen protocols i mètodes per extreure conclusions de tot allò que es objecte d'estudi, per això ens cal un ampli ventall de programari molt del qual s'ha d'anar creant sobre la marxa, per a emmagatzemar, ordenar, classificar i estudiar totes aquestes dades que abans mencionava.

Molts són els mètodes d'estudi, alguns portaran prediccions, alguns portaran conclusions sobre el funcionament de diferents parts de la vida i d'altres simplement curiositats, molts donaran errors que s'arrestaran durant anys i altres simplement teoritzaran, però tots, en un moment o altre hauran passat per una màquina *Linux*!

PERÒ...QUE ESTUDIEM ?

En tots, absolutament tots els processos químics que fan que els éssers vius ho estiguin hi intervé una o varies proteïnes, actuant fonamentalment com a catalitzadors, però essent a la vegada transportadors, reguladors, transmissors de senyals, activadors, inhibidors, i una llarga llista de noms acabats en "or". Aquestes estan formades per 20 aminoàcids, cada un dels quals amb propietats i estructura diferents, disposats en seqüència i estructura també diferents. La informació necessària per a la construcció de cada una d'aquestes està codificada en les cadenes d' ADN (Àcid DesoxiriboNucleic).

La composició d'una proteïna li donarà les propietats necessàries per a complir la seva funció i això inclou les propietats reactives de determinats aminoàcids i l'estructura (la "forma en l'espai") necessària per a que aquests puguin actuar allà on hagin de fer-ho.

El treball d'un bioinformàtic consisteix bàsicament en desenvolupar i/o utilitzar diferents programes per descobrir aquesta relació. Per què determinada seqüència proteica dóna aquesta estructura i no una altra? Per què determinada seqüència d'ADN només es troba en seqüències que donaran lloc només a proteïnes d'una determinada classe? Per què aquesta estructura només dona lloc a proteïnes amb una única funció tot hi que tinguin seqüències diferents? Quan una proteïna es troba en un medi determinat, per què l'adopció de diferents conformacions significa canvis en la possibilitat de realitzar millor o pitjor les seves funcions? Per què diferents maneres de plegar la mateixa proteïna donaran lloc a variacions en funció i estabilitat?

Aquestes són algunes de les preguntes que intenta contestar la bioinformàtica.

ALGUNS METODES D'ESTUDI

Tal com deia durant la introducció nosaltres treballem amb dades, seqüències de lletres i coordenades espacials majoritàriament. Una de les aproximacions a les respostes dins

d'aquest camp és la recerca de patrons o perfils presents en algunes seqüències o en altres per així abstreure'n determinades relacions entre diferents proteïnes. Comparant-les les unes amb les altres hem trobat patrons que són presents en totes les proteïnes d'una mateixa família, patrons que indiquen alguna de les funcions que portarà a terme la proteïna que els presenti i moltes altres coses. *PROSITE* [1] és una base de dades amb tot un conjunt de programes completament gratuïts i amb disponibilitat del seu codi font per a la recerca d'aquests patrons dins de seqüències "problema".

Això mateix que hem fet amb les seqüències podem fer-ho també amb les estructures, buscant zones semblants entre unes i altres. A ningú no se li escapa que no es el mateix comparar una seqüència de lletres que representa una proteïna que tot el conjunt de coordenades dels milers d'àtoms que poden arribar a formar-la. Existeixen multitud de programes disponibles a la xarxa amb els seus respectius codis font per a realitzar aquestes comparacions. El *CE* [2] i el *XAM* [3], en són uns clars exemples. Un cop s'han agrupat les estructures segons les seves semblances aquestes són classificades, ordenades i emmagatzemades en diferents bases de dades segons el criteri d'aquell que les fa i del motiu de la seva recerca. Després altres programes (com l'*archtype* [4] o el *modloop* [5]) utilitzaran aquestes bases de dades per a realitzar altres classificacions més especifiques o fins hi tot prediccions d'estructures no presents en cap base de dades.

Actualment quan es descobreix una nova proteïna es seqüencia i es compara amb tota la base de dades per a veure les característiques de les més semblants, la qual cosa es du a terme amb determinats algorismes de recerca (que solen portar el mateix nom que el programa que els implementa), els mes coneguts són BLAST [6] i FASTA [7], cap dels dos és GPL o GNU, però sempre es van subministrar els seus codis fon sense massa problema, per a institucions acadèmiques o inclòs usuaris individuals (tot hi que en la nova versió del BLAST això ja no és així, continuen circulant , i utilitzant-se, els codis fonts del original) La seva utilització està molt estesa i poden ser utilitzats per a la elaboració de qualsevol "software" propi sense cap mena de problema.

Quan estem treballant amb varies seqüències i en volem descobrir la relació existent, una bona manera de començar és fer un alineament múltiple de totes elles, aconseguint així veure quins aminoàcids son comuns i quins diferents, per a fer això utilitzem programes com el *CLUSTALW* [8], el codi font del qual és completament disponible per a aquell que el vulgui.

Algunes vegades amb la seqüència de les proteïnes en tens prou per a obtenir les respostes que busques, però moltes vegades no, moltes vegades la informació que necessites es troba en la estructura, i obtenir-la no és una cosa senzilla, la majoria de vegades impossible per moltes raons . La bioinformàtica també te respostes a això, tot i que encara no coneixem les regles que porten a una determinada estructura apartir d'una seqüència tenim diverses maneres d'aconseguir models força creïbles, alguna de les possibles de la mà del programa *MODELER* [9], el codi font del qual no està disponible *a priori*, però si es demana als seus creadors no és gens difícil d'aconseguir i per suposat el

permís per a desenvolupar-ne *software* derivat.

Tota molècula pertanyent a un ésser viu està a una determinada temperatura, sol estar envoltada d'aigua o per al contrari d'algun ambient hidrofòbic, cosa que pot variar la seva estructura i les seves propietats i per tant la seva capacitat per a realitzar les funcions que realitzen. Per estudiar com varien les macromolècules (molècules grans com l' ADN i les proteïnes que son el principal objecte dels nostres estudis) sotmeses a les diferents variacions de temperatura, ph i hidrofobicitat, que es poden trobar en el seu ambient natural, utilitzem tècniques de dinàmica molecular. Programes com Gromos [10], Gromacs [11], Amber [12] i altres simulen el moviment dels àtoms tenint en compte els seus enllaços. Si be algun d'aquests programes es distribueix sota llicencia GNU, la majoria cedeixen el codi font per un preu molt reduit destinat a cobrir costs bàsics. Programes com el *VMD* [13] permeten visualitzar tots aquests processos i fer-ne representacions.

Quan parlem d'estructures tridimensionals ens és necessari representar-les, per a fer-ho tenim programes "freeware" dels que en destacaríem el *Swiss-pdb Viewer* [14] o totalment lliures com el *RASMOL* [15] o el *PYMOL* [16], un programa fet en *python* [17] que et permet representar molècules de tot tipus de moltes maneres, seleccionar quins fragments representar i com i fins-hi tot fer-ne animacions.

Fins ara he parlat de diferents programes específics d'alguns dels camps que inclou aquesta nova ciència, la veritat és que n'hi han molts més per a fer les mateixes funcions o moltes altres que ni tan sols he esmentat aquí i que potser hauria d'haver-ho fet (Els versats en la matèria reconeixeran que *autodock* [18] o *HMMER* [19] són algunes de les grans absències) però tampoc vull expendre'm en excés en això. Tot seguit m'agradaria comentar altres programes més comuns i freqüents per a la comunitat "Open Source" d'arreu.

ALGUNES EINES GENERALS

Com ja s'ha dit Linux i les eines GNU que inclouen les distribucions són el suport, la base sense la qual res d'això seria possible, però moltes eines GPL en són també responsables.

L'existència de *mysql* [20] ha facilitat molt el treball a molts de nosaltres a l'hora d'integrar i ordenar els resultats dels anàlisis de multitud de bases de dades. La capacitat d'un llenguatge com el *perl* [21] ha reduït moltíssim el temps que els investigadors dediquen a la creació de programari tant per la seva gran integració amb *mysql* com pel fet que incorpori moltes de les funcions típiques de la bioinformàtica gràcies al seu mòdul *bioperl* [22]. L'*apache* [23] ha proporcionat la interfície més adequada per a que tothom pugui utilitzar totes aquestes eines, així com el *mozilla* [24] i altres navegadors han permès connectar-s'hi, a part, és clar, de buscar i trobar informació diversa i articles científics (i llegir-los amb l'xpdf [25]). L'OpenPBS [26] (i els antics NQS [27] i DQS [28]) ens han portat "els clusters de computadores". El *wget* ens permet descarregar grans

bases de dades sense preocupar-se massa per l'estabilitat de la xarxa i la llista seguiria i seguiria.
Potser sembla evident però no per això menys important, però l'*openoffice.org* [29] ha estat de gran utilitat, doncs des de la seva aparició no cal canviar-se d'ordenador per a llegir tots els ".doc" que envia la administració i/o la direcció dels centre, no cal canviar-se d'ordenador per a l'elaboració de presentacions per a seminaris i congressos com el que ens ocupa i no cal canviar-se d'ordenador per moltes altes coses que segur que a qualsevol que llegeixi això li vindran al cap.

CONCLSIÓ

Com crec que s'haurà pogut comprovar amb tot l'escrit fins aquí molts són els científics a tot el món que tenen molt a agrair a l'existència del codi lliure i a tota la gent que dedica part de les seves vides al seu desenvolupament. Potser ara, en mig de la batalla per les patents de "software" europees seria un bon moment per recordar-los-hi, doncs molts d'ells s'obliden que la majoria de les coses explicades aquí seran impossibles el dia que siguin una realitat i potser així i només així s'adonaran que són més importants els perjudicis que no pas els beneficis que esperen treure'n patentant els seus programes desenvolupats amb eines de Programari Lliure.

BIBLIOGRAFIA

PROSITE [1] http://ca.expasy.org/prosite/ *Hulo N., Sigrist C.J.A., Le Saux V., Langendijk-Genevaux P.S., Bordoli L., Gattiker A., De Castro E., Bucher P., Bairoch A.* **Recent improvements to the PROSITE database** *Nucl. Acids. Res. 32:D134-D137(2004).*

CE [2] http://cl.sdsc.edu/ce.html *Shindyalov IN, Bourne PE* **Protein structure alignment by incremental combinatorial extension (CE) of the optimal path**. *Protein Engineering 11(9) 739-747.(1998).*

XAM [3] *Xia, T.H.* **PhD Thesis No 9831** *ETH Zurich Switzerland B1961149J* .

archtype [4] http://bioinf.uab.es/archdb *Oliva, B. Bates, P.A. Querol, E. Avilés, F.X. And Stemberg* **An automated Classification of the structure of protein loops** *J. Mol. Biol 266, 814-830 (1997).*

modloop [5] http://alto.compbio.ucsf.edu/modloop//modloop.html *Fiser A, Do RK, Sali A.* **Modeling of loops in protein structures**.*Protein Sci. Sep;9(9):1753-73 (2000).*

BLAST [6] http://www.ncbi.nlm.nih.gov/BLAST/ *Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman* **Basic local alignment search tool** *J. Mol Biol 215(3):403-10.(1990).*
Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res. 25:3389-3402.(1997)* .

FASTA [7] http://www.ebi.ac.uk/fasta33/ *W. R. Pearson and D. J. Lipman* **Improved Tools for Biological Sequence Comparison.** *PNAS 85:2444- 2448. (1988)*
W. R. Pearson **Rapid and Sensitive Sequence Comparison with FASTP and FASTA** *Methods in Enzymology 183:63 – 98.(1990).*

CLUSTALW [8]http://www.ebi.ac.uk/clustalw *Higgins D., Thompson J., Gibson T.Thompson J.D., Higgins*

D.G., Gibson T.J. **CLUSTAL W: improving the sensitivity of progressivemultiple sequence alignment through sequence weighting,position-specific gap penalties and weight matrix choice.***Nucleic Acids Res. 22:4673-4680 (1994).*

MODELER [9] http://salilab.org/modeller/modeller.html *M.A. Marti-Renom, A. Stuart, A. Fiser, R. Sánchez, F. Melo, A. Sali.* ***Comparative protein structure modeling of genes and genomes****. Annu. Rev. Biophys. Biomol. Struct. 29, 291-325, (2000).*

Gromos [10] http://www.igc.ethz.ch/gromos/ *van Gunsteren WF, Billeter SR, Eising AA, Hu¨nenberger PH, Kru¨ ger P, Mark AE, Scott WRP, Tironi IG.* **Biomolecular simulation: The GROMOS96 manual and user guide***; vdf Hochschulverlag AG an der ETH Zu¨ rich and BIOMOS b.v. Zu¨ rich: Groningen; 1996.*
Gromax [11] http://www.gromacs.org/ *E. Lindahl, B. Hess and D. vand der Spoel:* **GROMACS 3.0: A package for molecular simulation and trajectory analysis.** *J. Mol. Mod. 7 pp. 306-317 (2001)*

Amber [12] http://amber.scripps.edu/ *D.A. Pearlman, D.A. Case, J.W. Caldwell, W.R. Ross, T.E. Cheatham, III, S. DeBolt, D. Ferguson, G. Seibel and P. Kollman.* **AMBER, a computer program for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to elucidate the structures and energies of molecules.** *Comp. Phys. Commun.* ***91,*** *1-41 (1995).*

VMD [13] http://www.ks.uiuc.edu/Research/vmd/ *Humphrey, W., Dalke, A. and Schulten, K.,* ***VMD - Visual Molecular Dynamics****, J. Molec. Graphics, 1996, vol. 14, pp. 33-38.*

SPDBV14] http://www.expasy.org/spdbv/ *Guex, N. and Peitsch, M.C.* **SWISS-MODEL and the Swiss-Pdb Viewer: An environment for comparative protein modeling.***Electrophoresis18, 2714-2723 (1997).*

RASMOL [15] http://openrasmol.org/ Roger A. Sayle, E. J. Milner-White ***RasMol: Biomolecular graphics for all****, Trends in Biochemical Sciences* ***20****(Sept):374-376, 1995.*

PYMOL [16] http://pymol.sourceforge.net/

python [17] http://www.python.org/

autodock [18] http://www.scripps.edu/pub/olson-web/doc/autodock/index.html *Morris, G. M., Goodsell, D. S., Halliday, R.S., Huey, R., Hart, W. E., Belew, R. K. and Olson, A.* **Automated Docking Using a Lamarckian Genetic Algorithm and and Empirical Binding Free Energy Function.** *J. J. Computational Chemistry, 19: 1639-1662. (1998).*

HMMER [19] http://hmmer.wustl.edu/ *S. Eddy* **Profile Hidden Markov Models** *Bioinformatics 14 755-763 (1998).*

mysql [20] http://www.mysql.com/

perl [21] http://www.perl.org/

bioperl [22] http://bioperl.org/

apache [23] http://httpd.apache.org/

mozilla [24] *http://www.mozilla.org*

xpdf [25] http://www.foolabs.com/xpdf/

OpenPBS [26] http://www.openpbs.org/

NQS [27] http://www.gnqs.org/oldgnqs/

DQS [28] http://www.scri.fsu.edu/~pasko/dqs.html

openoffice.org [29] http://www.openoffice.org/