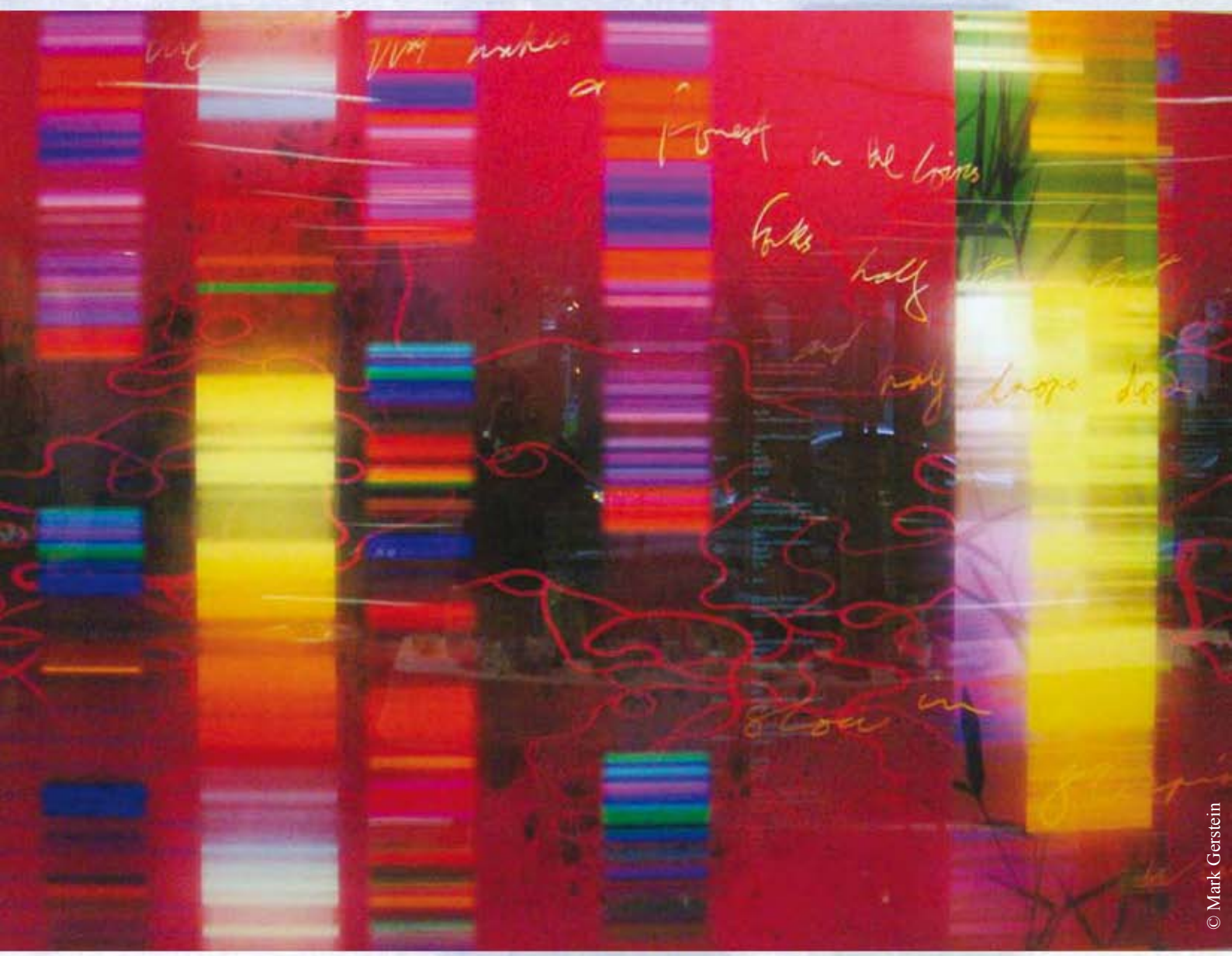
# BIOLOGIA I COMPUTACIÓ

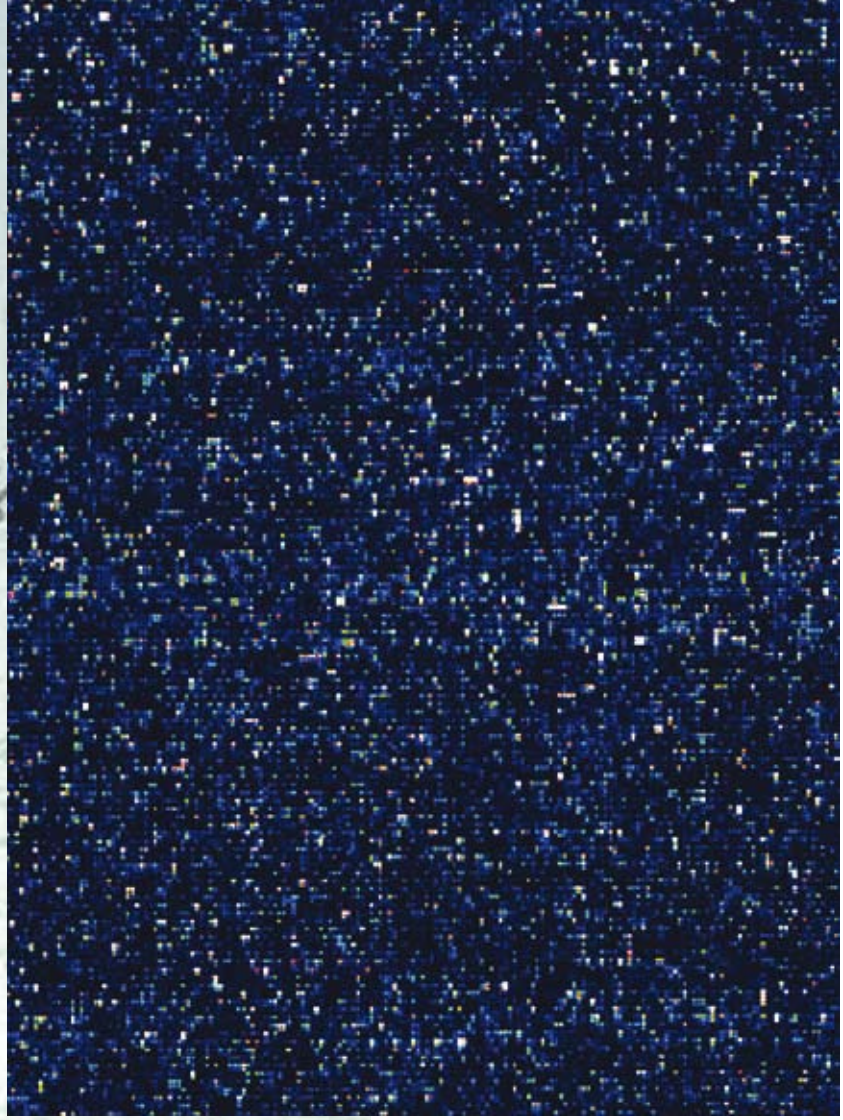


© Mark Gerstein

Escrit per:

**Roderic Guigó i Serra**

**CRG i IMIM. Universitat Pompeu Fabra**



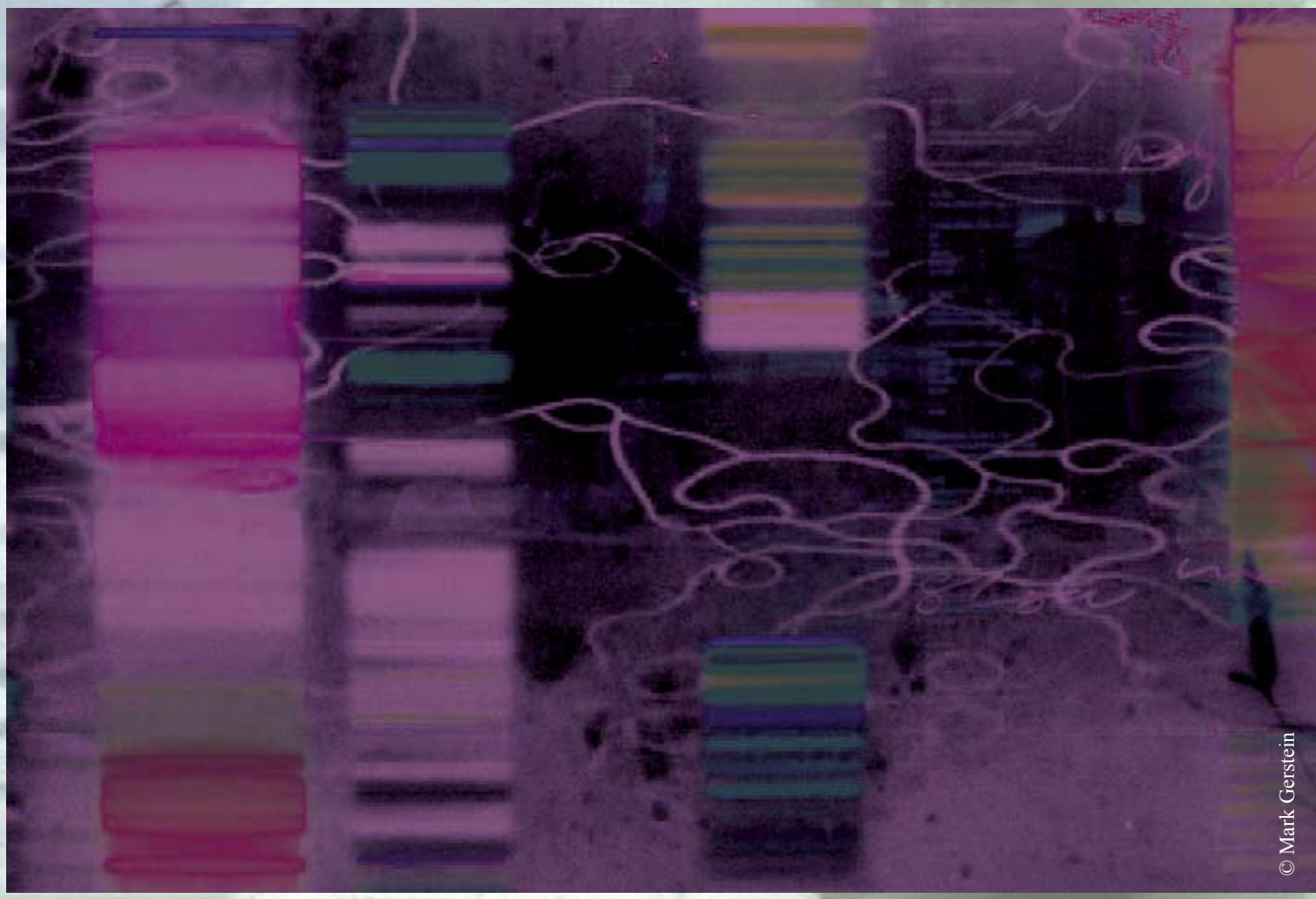
Cortesia de la UCTS, VAll d'Hebron

Si introduïm al *Google* la paraula *bioinformatics* —el terme que denota en anglès la disciplina científica en la qual la informàtica és utilitzada en la investigació en biologia— la recerca subsegüent d'Internet ens proporciona (a finals del 2006) prop de 40 milions de documents. Un nombre que no és molt diferent del nombre de documents a Internet que contenen paraules com genètica, bioquímica, fisiologia, ecologia, etc; tots ells termes que denoten disciplines de la biologia amb una història centenària. La bioinformàtica, no obstant això, és una disciplina molt nova: *Medline* —la base de dades que compila la literatura científica de la biologia i la medicina— no inclou cap article en el qual aparegui el terme *bioinformatics* amb anterioritat a l'any 1990. En aquests moments són prop de 12.000 articles, dels quals 2.500 han estat publicats el darrer any. Hem estat, doncs, testimonis en el curt espai d'una dècada, de l'explosió d'una nova disciplina científica —una explosió possiblement insòlita en a la història de la ciència.

Hom atribueix normalment al desenvolupament de les tecnologies de la genòmica el paper clau que la bioinformàtica juga avui dia en la investigació biològica. En efecte, les tecnologies de la genòmica generen un volum de dades sobre els fenòmens de la vida, d'una magnitud sense precedents en el camp de la biologia. Els experiments amb matrius de DNA (els DNA *microarrays* en anglès) són il·lustratius en aquest sentit. Mitjançant aquestes matrius, el suport físic de les quals tot just supera uns pocs centímetres quadrats, és possible obtenir en poques hores dades sobre el comportament simultani de milers de gens sota unes condicions determinades. Milers d'experiments amb matrius de DNA es porten a terme diàriament a tot el món, molts d'ells de forma gairebé automàtica. Fa només una dècada, tanmateix, l'obtenció d'aquestes mateixes dades sobre un únic gen era el resultat del treball continuat d'un investigador (o d'un equip d'investigadors) durant mesos, sovint durant anys. La biologia ha passat, doncs, en molt poc temps, de ser una ciència en la qual l'esforç humà s'orientava principalment cap a l'obtenció de (poques) dades, a ser una ciència que genera un volum literalment vertiginós de dades sense pràcticament cap intervenció humana. L'esforç de l'investigador s'ha desplaçat, en conseqüència, des de la producció cap a l'anàlisi de les dades. I és en aquest desplaçament on els mètodes informàtics juguen doncs un paper essencial, tant en la planificació dels experiments, com en la seva execució, i també, sobretot, en l'emmagatzematge i anàlisi dels seus resultats.

L'eclosió recent de la bionformàtica, però, no es produeix pas espontàniament; no sorgeix del no-res. La història de la biologia molecular i de la informàtica, després de la Segona Guerra Mundial, és, de fet, la història d'una interdependència creixent. Sempre és difícil posar una data exacta al naixement d'una determinada disciplina científica, però en el cas de la informàtica i la biologia molecular podríem dir que s'originen gairebé al mateix temps. Si, per posar alguna data, la biologia molecular neix l'any 1953, amb el desxiframent de l'estructura del DNA per part de Watson i Crick, i amb la determinació per primer cop, el mateix any, de la seqüència d'aminoàcids d'una proteïna per part de Sanger, la informàtica podríem dir que neix pocs any abans, amb el desenvolupament els primers ordinadors digitals programables amb memòria —és a dir, els ordinadors tal com els entenem avui dia.

Caldria esperar gairebé una dècada per a comprendre com la seqüència de nucleòtids del DNA exerceix la seva funcionalitat primària. A principis dels seixanta, els experiments de Korana, Brenner,

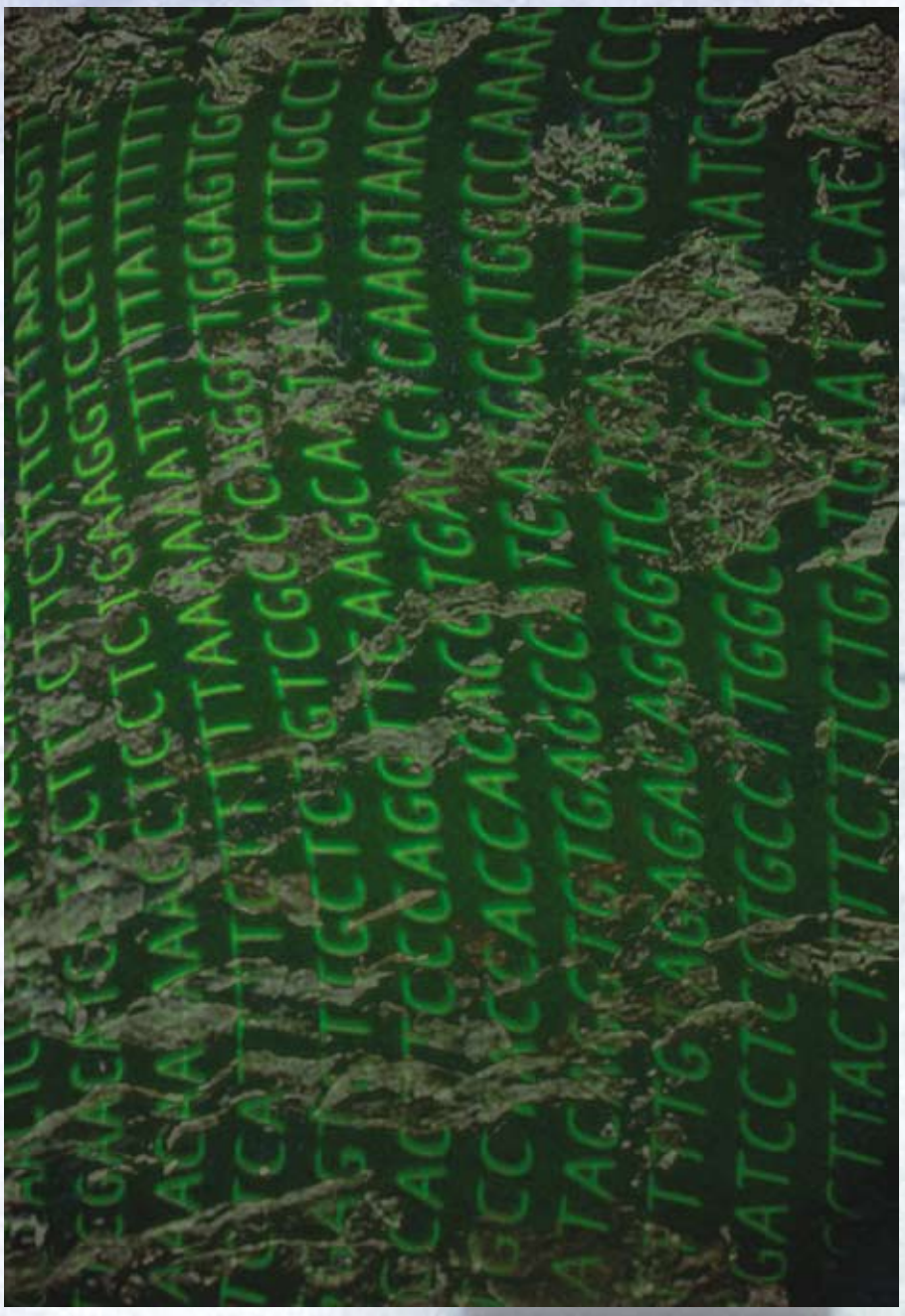



Kornberg i Ochoa entre d'altres, van permetre desxifrar l'anomenat codi genètic, el conjunt d'instruccions mitjançant les quals la seqüència de nucleòtids del DNA especifica la seqüència d'aminoàcids de les proteïnes. Pocs anys abans, hom inventa el primer llenguatge de programació d'alt nivell, el FORTRAN. Els llenguatges de programació d'alt nivell permeten escriure les instruccions mitjançant les quals l'ordinador resol un determinat problema, sense necessitat de conèixer com l'ordinador resol el problema. Fins llavors pràcticament només els enginyers que construïen els ordinadors eren capaços d'utilitzar-los. No és casualitat que en el vocabulari bàsic de la biologia molecular que comença a construir-se aleshores trobem tantes paraules que tenen un clar origen computacional: traducció, transcripció, codi, missatge, programa...

A principis dels seixanta, d'altra banda, els transistors comencen a substituir els tubs de buit en els circuits dels ordinadors i aquests són cada vegada més petits, ràpids i econòmics. A mitjan anys seixanta, la majoria de les grans empreses processaven la informació financera utilitzant ja ordinadors digitals. Mentrestant, el nombre de proteïnes de les quals es coneixia la seqüència d'aminoàcids augmentava sense parar. També a mitjan anys seixanta, Margaret Dayhoff i els seus col·laboradors van començar a compilar les seqüències d'aminoàcids conegudes. Van presentar aquestes compilacions als altres investigadors mitjançant els anomenats *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 finals dels seixanta, l'Atlas

contenia prop de tres-centes seqüències de proteïnes. L'Atlas es pot considerar, doncs, l'antecedent de les actuals bases de dades de seqüències, sense les quals la recerca en biologia seria impossible.

A la fi dels anys seixanta, amb l'aparició dels circuits integrats, els ordinadors es fan encara més petits, ràpids i econòmics. 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 estar a l'abast de les universitats i centres d'investigació. El fet que els ordinadors esdevinguessin més assequibles i la popularització dels llenguatges de programació d'alt nivell va fer que la computació comencés a convertir-se, en molts camps, part habitual de la pràctica científica. En el cas de la biologia, la repercussió va ser major en aquells camps en els quals l'anàlisi estadística o la modelització matemàtica juguen un paper més rellevant, com en la genètica, l'ecologia o la fisiologia.

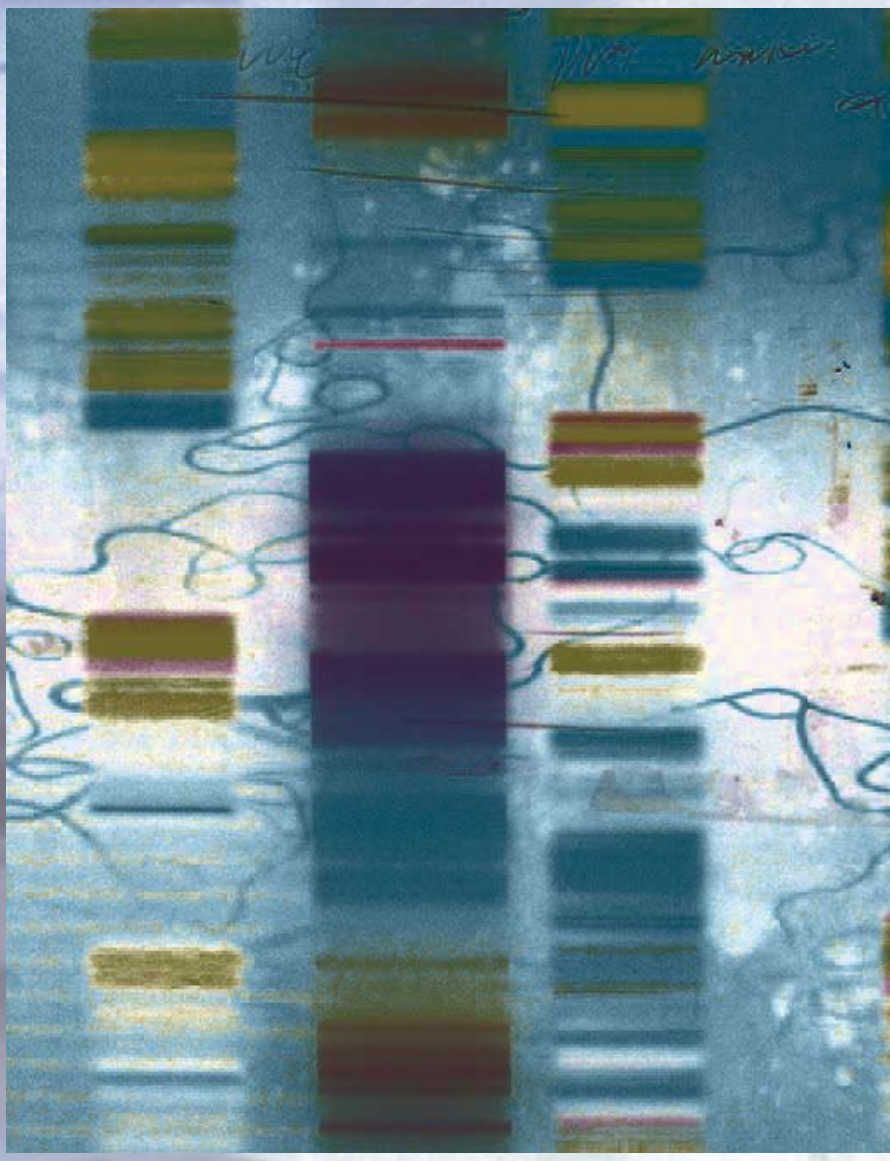


En particular, els ordinadors van permetre portar a terme anàlisis molt més exhaustives de les compilacions de seqüències d'aminoàcids de Dayhoff i col·laboradors. En particular, hom va poder comprovar que proteïnes amb funcions semblants exhibeixen sovint seqüències d'aminoàcids semblants. I al revés, seqüències d'aminoàcids semblants exhibeixen normalment funcions semblants. Mitjançant l'estudi sistemàtic dels canvis d'aminoàcids que es produeixen en proteïnes molt semblants, i la subsegüent construcció de matrius de substitució evolutives, hom aconsegueix quantificar la semblança entre dues seqüències. Aquesta estreta relació entre seqüència i funció i la capacitat de quantificar, de forma biològicament raonable, la semblança entre seqüències han estat pilars fonamentals sobre els quals s'ha bastit l'edifici de la biologia molecular moderna. En aquest sentit, si hem de posar alguna data al naixement de la bioinformàtica, aquesta seria a finals del anys seixanta amb els treballs pioners de Dayhoff.

Malgrat que a finals dels anys seixanta s'havia compilat ja la seqüència d'aminoàcids d'alguns centenars de proteïnes, la seqüenciació d'àcids nucleics romania per descobrir. A principis dels setanta, però, la situació canvia i, gràcies als treballs de Maxam i Gilbert d'una banda i de Sanger d'una altra, es posen a punt mètodes que permet finalment la seqüenciació d'àcids nucleics. Curiosament, això ocorre pràcticament al mateix temps que el Departament de Defensa dels Estats Units desenvolupava ARPAnet, una xarxa experimental d'ordinadors, que més tard esdevindria Internet, l'omnipresent xarxa d'ordinadors que tant a modificat les nostres vides. Dos esdeveniments que es produeixen de manera totalment independent i dels quals només ara podem veure la relació: la seqüència del genoma seria impossible sense Internet. Que difícil es anticipar cap on anirà la ciència —el món, en general!

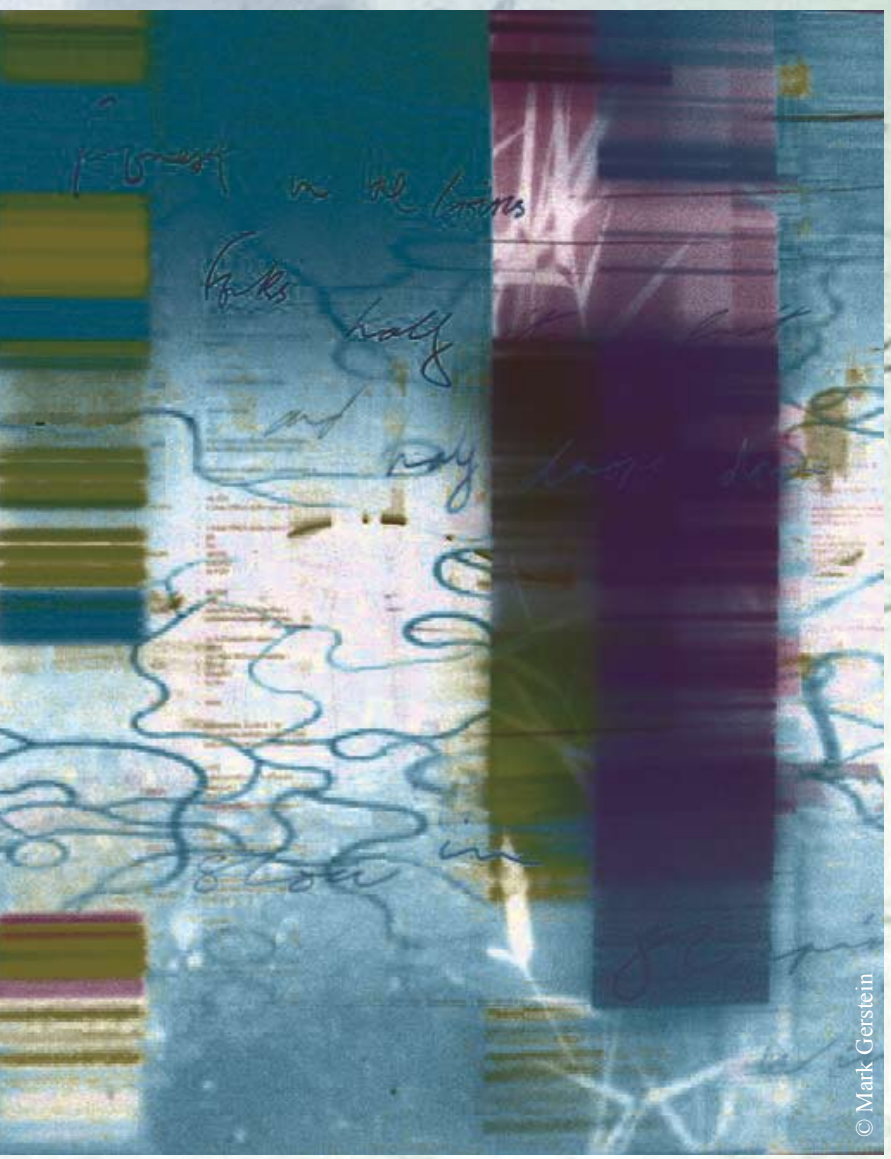
A principis dels vuitanta, el nombre de seqüències d'àcids nucleics havia crescut de manera espectacular. Era evident que la distribució de les col·leccions de seqüències en format imprès no podia

continuar per molt més temps. Així, l'any 1982 es creaven a Los Alamos National Laboratory, a Nou Mèxic, la base de dades americana de seqüències d'àcids nucleics en format electrònic, GenBank, i al Laboratori Europeu de Biologia Molecular (EMBL), a Heidelberg, l'equivalent europea. La primera versió de la base de dades d'EMBL, al juny de 1982, contenia 582 seqüències que sumaven poc menys de 600.000 nucleòtids. En aquests moments (desembre de l'any 2006) conté prop de 84 milions de seqüències i més de 150 mil milions de nucleòtids. L'existència de compilacions electròniques de seqüències va facilitar extraordinàriament la seva anàlisi computacional. Perquè, entre altres coses, el mateix any que es creaven les bases de dades electròniques de seqüències de DNA, IBM treia al mercat el primer ordinador personal, el PC. Els ordinadors començaven a ocupar les taules dels investigadors. I va ser precisament utilitzant el seu ordinador personal que Doolittle va descobrir, mentre realitzava comparacions entre les seqüències emmagatzemades en les bases de dades electròniques recentment creades, la similitud entre la seqüència d'un oncogen i la seqüència d'un factor de creixement. Una relació que havia passat desapercebuda als investigadors de Harvard i de Caltech i que contribuïa substancialment a la comprensió dels mecanismes moleculars involucrats en el càncer. Aquest i altres estudis semblants, en els quals la funció d'un gen era (almenys parcialment) inferida a partir de la similitud de la seva seqüència amb seqüències de funció coneguda, van demostrar la importància de les eines computacionals en la recerca en biologia.



I així, quan l'any 1990 començava "oficialment" el projecte del Genoma Humà, hom ja considerava indispensable el concurs de la computació. Com pot llegir-se en un dels document que a principis dels noranta va elaborar el Departament d'Energia (DOE), l'organisme que al costat dels *National Institutes of Health* (NIH), ha estat responsable als Estats Units del desenvolupament del projecte del Genoma Humà: "Els sistemes computacionals juguen un paper essencial en tots els aspectes de la investigació genòmica, des de l'adquisició de les dades fins a la seva anàlisi i manipulació. Sense ordinadors potents i sistemes apropiats per al tractament de les dades la investigació genòmica és impossible."

I certament, la quantitat de dades que genera la recerca genòmica, fa que avui en dia no puguem pensar la biologia sense la informàtica. Pot argumentar-se, tanmateix, que la generació massiva de dades no és patrimoni de la investigació en biologia, i que el mateix fenomen té lloc en àrees tan diverses de la ciència com l'astronomia, l'economia, la física d'altes energies, la meteorologia etc. I que, en totes elles, els mètodes computacionals juguen un paper essencial en el tractament i interpretació de les dades. És cert. ¿Però per què, doncs, com podem comprovar de nou al *Google*, existeixen tants pocs documents a Internet en els quals apareguin termes com *astroinformatics*, *meteoinformatics*, *econoinformatics*, o semblants, en contrast amb les desenes de milions




Roderic Guigó va obtenir el seu títol de doctor a la Universitat de Barcelona el 1988. Va treballar al Departament d'Estadística de la Facultat de Biologia en el desenvolupament de models matemàtics i informàtics d'ecologia evolutiva i genètica de poblacions. Va treballar al *Molecular Biology Computer Research Resource* del *Dana Farber Cancer Institute* a la Universitat de Harvard, i també al *BioMolecular Engineering Research Center* de la Universitat de Boston, on va estar treballant en el camp de l'anàlisi de seqüències. El 1992 va anar al *Los Alamos National Laboratory*, on va estudiar problemes relacionats amb l'anàlisi genòmica. Des del 1994 és investigador a l'Institut Municipal d'Investigació Mèdica, al Grup de Recerca en Informàtica Biomèdica (GRIB). Des del 2005 coordina el programa de Bioinformàtica i Genòmica del Centre de Regulació genòmica de Barcelona.

de documents en els quals apareix el terme *bioinformatics*? En la meva opinió, això és en part així, perquè la relació entre biologia i computació s'estableix a un nivell més íntim que no pas simplement el de la quantitat de les dades i té a veure amb la "qualitat" —entesa com la naturalesa— d'aquestes dades. La vida comença quan els nucleòtids s'organitzen en la seqüència del genoma. Per sota de la seqüència del genoma hi ha la química i la física. I és l'ordre particular dels nucleòtids d'aquesta seqüència —més que no pas en les seves característiques fisicoquímiques— el codi, com tan bé va anticipar Shröedinger als anys quaranta, que dicta les característiques biològiques dels ésser vius. I la vida, el desplegament en el món d'un ésser viu, és aleshores una computació, gairebé en un sentit paradigmàtic, sobre la seqüència del genoma. Un dels grans reptes de la biologia del segle XXI és, precisament, desxifrar els múltiples codis mitjançant els quals es produeix aquesta computació.