La computació i el projecte Genoma Humà

RODERIC GUIGÓ
Institut Municipal d'Investigació Mèdica de Barcelona (IMIM)

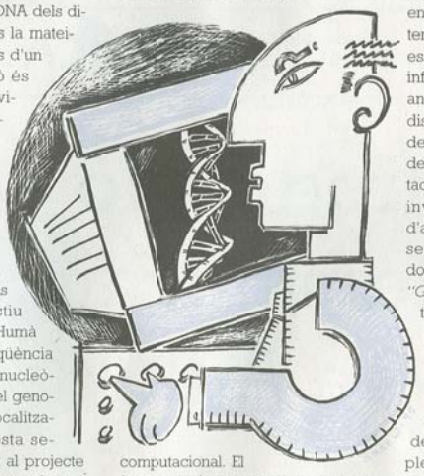
L'objectiu del projecte Genoma Humà és la caracterització del material genètic dels éssers humans. El projecte permetrà aprofundir el nostre coneixement del funcionament molecular dels organismes vivents, i tindrà un profund impacte en la medicina, en l'agricultura i en molts processos industrials. Atès el volum i la naturalesa de la informació que generarà, el projecte és impensable sense el concurs de la computació.

El cos humà és constituït per milers de milions de cèl·lules. En les cèl·lules es troben els cromosomes, molècules llarges d'Àcid Desoxirribonucleic (DNA). El DNA és constituït per la successió de molècules elemental anomenades nucleòtids. Una de les funcions importants del DNA és "codificar" les proteïnes. Les proteïnes són els components estructurals de les cèl·lules i fan possible els processos bioquímics subjacents a les manifestacions de la vida (la respiració, el llenguatge, ...). Les proteïnes són cadenes lineals de centenars d'unitats elemental anomenades aminoàcids, que poden ser de 20 tipus diferents. D'acord amb la seqüència d'aminoàcids, aquestes cadenes es pleguen i adopten una conformació tridimensional, de la qual depèn la seva funció. Hom creu que existeixen uns cent mil tipus diferents de proteïnes en l'organisme humà. La seqüència concreta d'aminoàcids de les proteïnes d'un organisme és "codificada" per la seqüència de DNA en els cromosomes d'aquest organisme. Cada grup de 3 nucleòtids consecutius en la seqüència de DNA especifica un aminoàcid en la seqüència d'una proteïna. La maquinària cel·lular "llegeix" la seqüència de DNA dels cromosomes d'acord amb aquest codi i sintetitza les proteïnes corresponents. Per això hi ha qui es refereix al DNA com al programa que especifica la construcció d'un organisme. Proteïnes diferents són codificades per segments diferents del DNA, anomenats "gens".

La seqüència de DNA dels diferents cromosomes és la mateixa en totes les cèl·lules d'un mateix individu, però és diferent en les d'individus diferents. La seqüència de DNA característica és el que anomenem el genoma. El genoma d'individus d'espècies diferents varia en el nombre, tipus i organització dels gens en els cromosomes. L'objectiu del projecte Genoma Humà és l'obtenció de la seqüència dels 3.000 milions de nucleòtids que constitueixen el genoma dels humans i la localització dels gens d'aquesta seqüència. Paral·lelament al projecte Genoma Humà, estan sent desenvolupats projectes de seqüenciació del genoma d'altres organismes i el projecte de la Diversitat del Genoma, que pretén determinar el grau de variabilitat del genoma humà.

El projecte Genoma Humà tindrà molta influència en el desenvolupament de la Biologia. Permetrà, primer, un millor coneixement del funcionament de l'organisme a nivell molecular i de la causa bioquímica de malalties amb un component genètic. Segon, la seqüenciació dels genomes d'organismes diferents permetrà una millor comprensió de les relacions filogenètiques entre les espècies i la identificació de la component genètica subjacent als trets característics d'una espècie. Tercer, l'estudi de la variabilitat genòmica dins l'espècie humana permetrà determinar la importància dels factors genètics *versus* els ambientals en la diversitat dels individus. D'altra banda, el projecte Genoma tindrà conseqüències tecnològiques importants, com ara la possibilitat de modificar genomes o de dissenyar-ne de nous.

La Informàtica juga un paper molt important en la investigació en Biologia Molecular. La informació que produeix aquesta investigació la fa apropiada a l'anàlisi



computacional. El fet que en la seqüència d'una biomolècula es troba implícita la seva funció, fa aquesta anàlisi extraordinàriament rellevant. Avui, la utilització de la computació ha esdevingut rutinària als laboratoris de Biologia Molecular. Així, quan els investigadors obtenen una nova seqüència, l'anàlisi computacional és utilitzat per localitzar-ne dominis funcionals, per predir-ne el plegament, per establir-ne relacions filogenètiques, però sobretot per comparar-la amb les seqüències ja emmagatzamades a les bases de dades, de manera que la similaritat amb una seqüència ja coneguda en pugui suggerir la funció. Característic de la utilització de la informàtica en Biologia Molecular és que la computació es realitza essencialment a Internet. Tant les bases de dades de bioseqüències actualitzades diàriament, com els programes i recursos d'anàlisi de seqüències es troben distribuits per Internet, s'hi accedeix utilitzant la infraestructura proporcionada pel WWW.

És evident que en els projectes de seqüenciació genòmica, la Informàtica hi jugarà un paper crucial. En aquests moments, GenBank, la base de dades de les seqüències de nucleòtids als EUA, rep al voltant de 20.000 connexions diàries per Internet. Conté 492.000 seqüències de més de 11.000 espècies, que totalitzen 354 milions de nucleòtids. Tot i així, la longitud de les seqüències emmagatzemades a GenBank representa només un 10% de la longitud del genoma d'un únic individu humà. El projecte del Genoma generarà, doncs, un enorme volum d'informació; els sistemes informàtics juguen un paper essencial en l'adquisició d'aquesta informació, la seva administració, anàlisi i interpretació, i una nova disciplina científica, la "Informàtica del Genoma", està sorgint per tal de fer front als problemes computacionals que es generen durant la investigació genòmica. Alguns d'aquests problemes només poden ser tractats amb supercomputadors, i han estat identificats com a *"Grand Challenges"*. En aquest sentit, ordinadors paral·lels han estat recentment utilitzats en algunes àrees de l'anàlisi de seqüències, com ara la reconstrucció de filogènies moleculars, la recerca de similaritats en les bases de dades de bioseqüències i el plegament tridimensional de les proteïnes i seran utilitzats aviat en àrees com ara l'anàlisi comparativa de genomes, l'anàlisi de seqüències de DNA de mida cromosòmica i l'extracció automàtica de coneixement de les bases de dades.

En definitiva, el projecte del Genoma Humà representa un punt d'inflexió a història de la Biologia. Amb ell, la Biologia deixa de ser una ciència purament experimental per passar a tenir un fort component teòric o computacional. Com ha escrit recentment John Maddox, l'editor de la revista *Nature*, "la computació i la biologia molecular, ja interdependents, estan a punt d'unir-se inextricablement... Els ordinadors són cada cop més un dels mitjans a través dels quals els problemes de biologia molecular poden ser resolts". Però no es tracta només que la computació pugui contribuir a resoldre determinats problemes en Biologia Molecular, sinó que aquests problemes només poden ser plantejats en termes computacionals. Despres de tot, la síntesi de proteïnes a partir del DNA és una computacio, el desxiframent del programa de la qual es l'objectiu últim del projecte del Genoma Humà.