Del genoma als gens

Ferriol Calvet[1] i Roderic Guigó[1, 2]

[1] Centre de Regulació Genòmica (CRG), Barcelona Institute of Science and Technology (BIST)
[2] Universitat Pompeu Fabra (UPF)

Correspondència: Roderic Guigó. Centre de Regulació Genòmica. C. del Dr. Aiguader, 88. 08003 Barcelona. Adreça electrònica: *roderic.guigo@crg.cat*.
Ferriol Calvet. Institut de Recerca Biomèdica. C. de Baldiri Reixac, 10. 08028 Barcelona. Adreça electrònica: *ferriol.calvet@irbbarcelona.org*.

DOI: 10.2436/20.1501.02.212
ISSN (ed. impresa): 0212-3037
ISSN (ed. digital): 2013-9802
http://revistes.iec.cat/index.php/TSCB
Rebut: 02/05/2022
Acceptat: 14/06/2022

Resum

Durant el segle xx es va establir la naturalesa molecular dels gens —les unitats bàsiques de l'herència biològica— i dels processos involucrats en el flux d'informació dels gens a les proteïnes. La generalització de les tècniques de seqüenciació va facilitar l'obtenció de la seqüència dels genomes, però la seva utilitat és limitada sense un mapa dels gens. Per tal de construir-lo, hom utilitza una combinació de mètodes experimentals i computacionals. Les noves tecnologies que permeten l'obtenció de la seqüència completa dels RNA missatgers són les preferides, però no sempre es poden utilitzar. Per aquest motiu, els mètodes computacionals són essencials. Aquests mètodes inclouen la localització de regions en el genoma amb uns biaixos en la composició de la seqüència característics de les regions codificants de proteïnes. Els projectes en marxa que seqüenciaran el genoma de totes les espècies eucariotes faran necessari el desenvolupament de mètodes computacionals cada cop més acurats i eficients.

Paraules clau: anotació, seqüenciació, gens, predictors *ab initio*.

From genome to genes

Abstract

In the 20th century, scientists established the molecular nature of genes – the basic units of biological inheritance – and of the processes involved in the flow of information from genes to proteins. The generalization of sequencing techniques facilitated the determination of the sequence of genomes, but their usability was limited without a map of all the genes involved. In order to build it, a combination of experimental and computational methods are used. The new sequencing technologies that allow the determination of the complete sequence of messenger RNAs are preferred, but they cannot always be used and this is why computational methods are essential. These methods include the identification of regions in the genome with sequence composition biases similar to those in the regions known to code for proteins. The ongoing projects that will sequence the genome of all the eukaryote species will make it necessary to develop more accurate and efficient computational methods.

Keywords: annotation, sequencing, genes, *ab initio* predictors.

La naturalesa molecular dels gens

El terme *gen* va ser introduït l'any 1905 per Wilhem Johansen, per referir-se als «factors» hereditaris que Gregor Mendel havia proposat, mig segle abans, com a responsables dels trets observables dels individus. A principis del segle xx, es va determinar que els gens estaven situats als cromosomes, i el primer mapa genètic es va construir l'any 1913. Aquest avenç va ser seguit pel desenvolupament de mapes de més resolució i el càlcul de distàncies entre gens. Tots aquests models es van construir sense conèixer la naturalesa química del material hereditari, però dos grups de científics, Oswald Avery, Colin MacLeod i Maclyn McCarty (Avery *et al.* 1944), primer, i Alfred Hershey i Martha Chase (Hershey i Chase, 1952), després, van demostrar que es tractava d'àcid desoxiribonucleic (DNA). El pas final en la caracterització d'aquesta molècula va ser el descobriment de la seva naturalesa polimèrica fet per Watson i Crick l'any 1953 (Watson i Crick, 1953). Al mateix temps, Frederick Sanger desenvolupava metodologies de seqüenciació que van permetre caracteritzar la seqüència de proteïnes i, dues dècades més tard, la seqüència del DNA (Sanger i Tuppy, 1951; Sanger *et al.*, 1977). Sanger va ser guardonat amb un Premi Nobel per cadascun d'aquests descobriments.

Paral·lelament a aquests avenços en l'estructura molecular del material hereditari, George Beadle i Edward Tatum van establir l'equivalència entre gens i proteïnes el 1941 (Beadle i Tatum, 1941). Poc després, George Gamow va establir la relació entre les seqüències d'aquestes dues biomolècules (Gamow *et al.*, 1956) i el 1961 Sidney Brenner, François Jacob i Matthew Melenson van descobrir l'àcid desoxiribonucleic (RNA) (Brenner *et al.*, 1961). En concret, van descobrir l'RNA missatger (mRNA), que era la biomolècula que va permetre explicar com la informació codificada en el DNA, que es trobava dins el nucli, podia arribar al citoplasma, on es produeixen les proteïnes. Se sabia que hi havia d'intervenir una tercera molècula, i aquests tres científics la van identificar. Un mes després, Jacob i Jaques Monod van descriure l'mRNA en profunditat i en van destacar la possible funció reguladora (Jacob i Monod, 1961); i poc després d'aquest descobriment, van definir la correspondència entre els triplets de nucleòtids de les seqüències de DNA i mRNA (codons), i els aminoàcids, que formen la seqüència de les proteïnes. El 1970, recopilant els avenços fets des de principis de segle, es va establir el *dogma central de la biologia molecular* (Crick, 1970), el qual estableix que «el DNA que resideix al nucli es transcriu a mRNA, que es desplaça al citoplasma, on successivament es tradueix a proteïna».

A partir d'aquests treballs i d'altres, i un cop establerta la naturalesa molecular dels gens, es va anar configurant una visió de com la informació genètica és configurada en la seqüència del DNA dels cromosomes d'acord amb la qual els gens correspondrien a regions (*loci*) d'aquesta seqüència, separades les unes de les altres, cada una de les quals especificaria la seqüència d'un mRNA, el qual, al seu torn, especificaria la seqüència d'aminoàcids d'una proteïna.

Anotació gènica

Entenem per *anotació gènica* el procés mitjançant el qual s'identifiquen els gens codificats en la seqüència genòmica. A part de la seva ubicació al genoma, cada gen té una estructura interna. Aquesta estructura es compon d'exons i introns i és determinada pels proces-

sos d'empalmament (*splicing*, en anglès) alternatiu i maduració als quals se sotmet l'mRNA després de produir-se amb la transcripció. El procés d'empalmament consisteix a eliminar les regions intròniques de l'mRNA i concatenar les regions restants conegudes com a *exons*. Aquest procés està mediat pel complex de tall i unió (*spliceosome*, en anglès), i és responsable que hi hagi una diversitat més gran d'RNA missatgers que de gens. Utilitzant les estadístiques de l'anotació de GENCODEv39 (Frankish *et al*., 2021) per al genoma humà, hi ha 3,98 vegades més transcrits que gens (244.939 i 61.533, respectivament).

L'objectiu del procés d'anotació d'un genoma és identificar tots els gens presents en aquell genoma, determinar les coordenades exactes de tots els exons i introns (figura 1) i assignar la funció o funcions corresponents a cada gen (Brent, 2005; Harrow *et al*., 2009). La identificació de la posició de tots els elements d'un gen en el genoma l'anomenem *anotació estructural*, mentre que el procés d'assignar la funció corresponent a cada gen és l'*anotació funcional*.

Centrant-nos en l'anotació estructural, tenint en compte la naturalesa de l'estructura interna dels gens, es poden utilitzar diferents enfocaments per arribar a una anotació gènica precisa. L'obtenció experimental de la seqüència d'mRNA és la via més directa, i es va començar a implementar a finals del segle anterior amb el desenvolupament dels primers mètodes de seqüenciació. Una manera alternativa d'identificar gens és fer prediccions mitjançant eines computacionals que poden calcular mètriques sobre la seqüència del genoma i determinar quines regions tenen més probabilitat de codificar proteïnes. Aquest segon grup de mètodes es van poder començar a desenvolupar i aplicar un cop es van veure les característiques distintives de les regions amb gens respecte de les regions en què no n'hi havia.

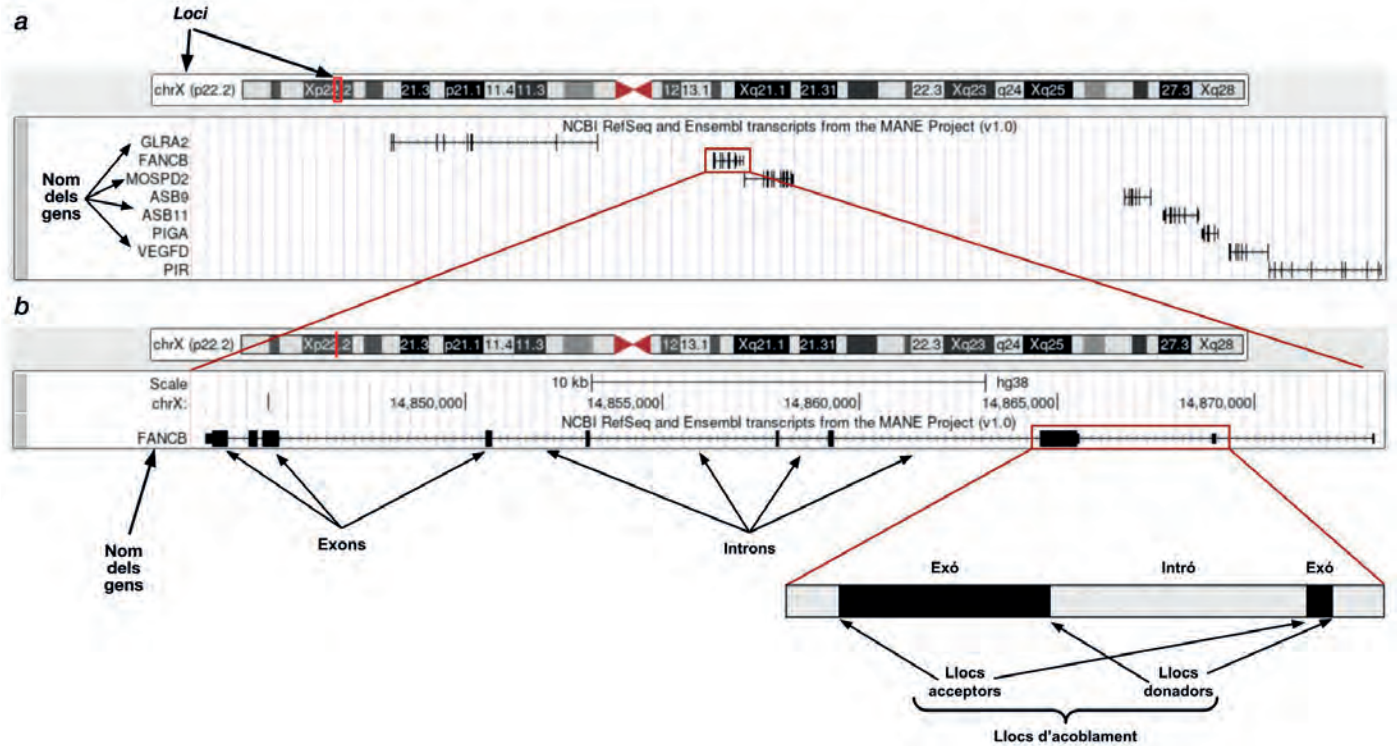
Tots dos enfocaments han evolucionat des de les primeres implementacions, i en la majoria de protocols d'anotació es complementen amb l'objectiu d'identificar l'estructura exacta de tots els gens en el genoma de qualsevol espècie seqüenciada.

Les següents seccions d'aquesta revisió se centraran en els avenços fets en les eines experimentals i computacionals per obtenir l'estructura dels gens, i en els passos seguits per la majoria de mètodes d'anotació gènica (*pipelines*, en anglès).

Obtenció experimental de l'estructura gènica

Amb els grans avenços en el camp de la biologia molecular durant les dècades de 1970 i 1980, els investigadors van poder estudiar regions específiques del genoma amb un detall sense precedents. L'interès principal se centrava en aquells *loci* relacionats amb malalties que havien estat prèviament cartografiats en el genoma. Tanmateix, un cop disponible la seqüència de nucleòtids de la regió, la seqüència exacta del gen codificat en aquella regió no era fàcil d'aconseguir, ja que les seqüències dels exons i dels introns no es poden distingir fàcilment, i això és imprescindible per poder trobar marcs de lectura oberta (ORF, de l'anglès *open reading frame*). Aquestes són les regions dels gens que contenen la seqüència que codifica la proteïna final.

Com que la seqüència genòmica no era suficient, es van desenvolupar estratègies per obtenir la seqüència de l'mRNA, on es poden buscar directament els codons d'inici i aturada que defineixen l'ORF. La seqüenciació directa de l'mRNA no era possible, i la millor alternativa era utilitzar un pas intermedi per retrotranscriure la molècula d'mRNA a DNA complementari (cDNA), ja que aquesta sí que es



↑ Figura 1. Visualització de la posició i l'estructura interna d'un gen. *a*) Posició, també anomenada *loci*, de 8 gens a la regió p22.2 del cromosoma X d'humà. Representació dels gens reduïda a un únic transcrit per gen. *b*) Descripció de l'estructura interna del gen FANCB. S'observen tots els exons i els introns d'aquest gen i també els llocs d'acoblament. El canvi exó-intró l'anomenem *lloc donador* i l'intró-exó, *lloc acceptor*. Figura extreta d'*http://genome.ucsc.edu* (Lee *et al*., 2022), amb les anotacions de Morales *et al*. (2022) [consulta: 15 març 2022].

podia seqüenciar amb els protocols de seqüenciació de DNA ja desenvolupats. Aquesta encara és actualment una de les tècniques més utilitzades per obtenir la seqüència d'mRNA. Havent trobat aquesta alternativa, hi havia un altre obstacle. Les lectures de seqüenciació més llargues produïdes per les diverses tècniques de seqüenciació disponibles fins al moment no eren suficients per cobrir tota la molècula d'mRNA. A causa d'això, els mRNA s'havien de retrotranscriure a cDNA i fragmentar-se en segments prou petits que sí que es podien seqüenciar de manera completa. Finalment, es necessitava un procés d'assemblatge per reconstruir les molècules d'mRNA a partir de tots els fragments seqüenciats. Aquest procés era especialment complex perquè no es podien aïllar només les molècules d'mRNA d'interès, tret que es tingués coneixement *a priori* de la seqüència.

Els mètodes experimentals per determinar l'estructura de l'mRNA s'han actualitzat des de les primeres versions (Wang *et al*., 2009). El volum de dades generat és més gran i les dades són de millor qualitat, i els mètodes per analitzar-les també són millors i més eficients. Aquests canvis es tradueixen en la capacitat de determinar l'estructura dels gens de manera molt més precisa.

Entre les millores més rellevants, en l'àmbit experimental, que faciliten el procés d'anotació gènica podem trobar el desenvolupament de tècniques de seqüenciació d'RNA (RNA-seq) de lectura llarga i tècniques de seqüenciació directa. En aquests moments, hi ha dues tecnologies principals per RNA-seq de lectura llarga. La primera es basa en la capacitat de detectar la incorporació d'un únic nucleòtid en una cadena de DNA; aquesta incorporació està catalitzada per una DNA-polimerasa unida a la molècula de DNA que es vol seqüenciar (Schadt *et al*., 2010). La segona fa ús de la capacitat de detectar canvis en un camp elèctric que s'indueixen quan cada nucleòtid d'una molècula de DNA passa per un porus (Clarke *et al.*, 2009). Un avantatge d'aquesta segona tecnologia és que fa possible seqüenciar directament les molècules d'RNA, sense necessitat de copiar-les a cDNA.

Pipelines d'anotació gènica

El procés d'anotació no té uns requeriments estrictes i els investigadors que volen produir una anotació ho fan seguint els mètodes que els semblin més adequats en cada cas. Tot i això, hi ha un conjunt de passos que són els més habituals i els utilitzats per les *pipelines* d'anotació dels principals centres que generen i publiquen més anotacions (Aken *et al*., 2016; Thibaud-Nissen *et al*., 2013).

El primer pas abans d'iniciar el procediment d'anotació ha de ser comprovar la qualitat de la seqüència del genoma. Idealment, hauríem de disposar d'una única seqüència per a cada cromosoma, però, ateses les limitacions tecnològiques, això sovint és difícil, i molts cops les seqüències dels cromosomes estan fragmentades. Mirant les estadístiques del genoma, centrant-nos en el nombre de seqüències i la seva llargada, i comparant-ho amb el nombre de cromosomes esperats, es pot veure com és de fragmentat el genoma. Això influirà en l'anotació, ja que alguns gens poden caure a les unions entre seqüències i es dividiran en dos o més gens diferents. Un altre procés més elaborat consisteix a utilitzar el mètode BUSCO (de l'anglès *benchmarking universal single-copy orthologs*) (Simão *et al.*, 2015) per avaluar la integritat del genoma. Aquest mètode es basa en la recerca d'un conjunt de gens conservats en totes les espècies d'un grup taxonòmic concret, i que haurien d'existir en el genoma que es vol anotar; informa, aleshores, de la presència o absència d'aquests gens, així com de possibles duplicacions o fragmentacions. Manni *et al.* (2021) indica els protocols que cal seguir per avaluar la qualitat de les seqüències genòmiques.

Un cop s'ha comprovat que la qualitat del genoma és suficient per poder iniciar el procés d'anotació, la disponibilitat de dades experimentals determinarà quins protocols es poden seguir. Quan hi ha dades transcriptòmiques de l'espècie que es vol anotar o d'una espècie propera, s'utilitzen per aconseguir una identificació més precisa dels gens en el genoma. En cas contrari, quan no hi ha dades transcriptòmiques disponibles de l'espècie en estudi ni de cap espècie propera, es poden usar procediments alternatius, però amb l'inconvenient que la qualitat de l'anotació final és pitjor.

Mètodes dependents de la disponibilitat de dades transcriptòmiques

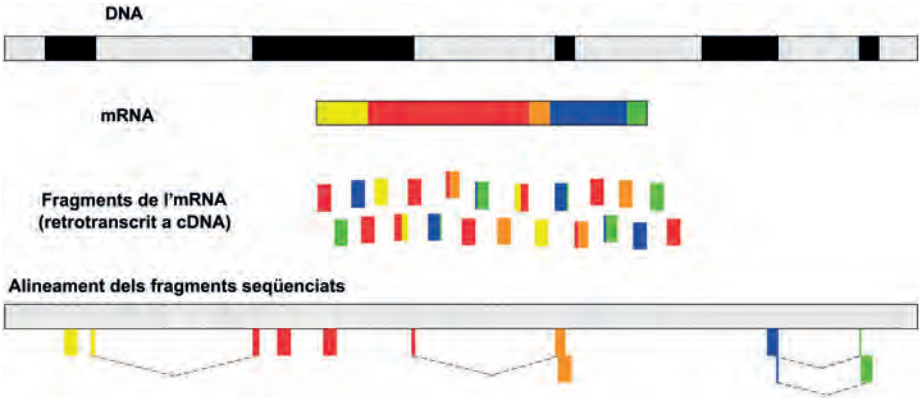
El millor enfocament per determinar la ubicació i l'estructura dels gens és alinear l'evidència transcriptòmica amb el genoma en estudi. Aquesta evidència transcriptòmica pot tenir la forma de dades d'RNA-seq de lectura curta o de lectura llarga procedents de la mateixa espècie o d'una espècie evolutivament propera. El cas ideal és utilitzar dades transcriptòmiques de lectura llarga de la mateixa espècie de la qual es vol anotar el genoma. Si és possible, és encara millor tenir les dades transcriptòmiques del mateix individu del qual es va obtenir el genoma.

Després de la comprovació de la qualitat inicial de les dades d'RNA-seq, s'han d'alinear amb el genoma. Aquest pas és clau per poder construir després els models dels gens i els seus mRNA. L'alineació de seqüències d'RNA-seq amb el genoma permet identificar els llocs d'acoblament (*splice sites*, en anglès) i les posicions dels introns, les quals són indicades per interrupcions en alineament (figura 2). També permet identificar parcialment les regions exòniques en què s'alineen lectures senceres. Aquesta informació pot ser utilitzada per diversos programes que construeixen els models de gens i transcrits més probables. Si fem servir RNA-seq de lectura llarga, el procés d'assemblatge del model de transcripció és més senzill, ja que les cadenes contínues d'exons i introns són capturades de manera completa per les seqüències. Les dades generades amb les primeres versions d'RNA-seq de lectura llarga eren propenses a errors i el procés d'alineació tampoc no era trivial. Actualment, amb els avenços de les tecnologies de seqüenciació, s'obtenen dades de més qualitat i aquest problema s'ha minimitzat.

Mètodes independents de la disponibilitat de dades transcriptòmiques

Després d'utilitzar les dades experimentals per generar el primer conjunt de transcrits altament fiable, es pot combinar l'ús d'altres mètodes per afegir nous models de gens. Aquests es basen en l'alineació de la informació disponible en bases de dades públiques o analitzen les periodicitats i els biaixos de les regions codificants del genoma per suggerir gens potencials que codifiquin proteïnes, alguns dels quals podrien no haver estat detectats en el procés d'alineació de les dades d'RNA-seq. La majoria d'aquests mètodes són sensibles a les regions repetitives o de baixa complexitat i, per aquest motiu, necessiten un fitxer amb la seqüència completa on aquestes regions que se sap que no acostumen a tenir gens estiguin senyalades d'una manera diferent de la resta. Els genomes que estan en aquest format s'anomenen *genomes emmascarats* (*masked genomes*, en anglès).

El procés d'emmascarament del genoma consisteix a identificar regions de DNA de baixa complexitat i regions amb elements repetitius (Hancock, 2002). Es pot fer *ab initio* o basant-se en biblioteques de regions repetitives.



↑ Figura 2. Representació simplificada del procés d'RNA-seq. En primer lloc, s'obtenen els mRNA cel·lulars que s'han produït en el procés de transcripció i la maduració posterior, durant la qual els introns són eliminats. Al laboratori, aquests mRNA es retrotranscriuen a cDNA. Les molècules resultants es fragmenten en trossos més petits, els quals poden ser seqüenciats de manera completa. Un cop s'obtenen les seqüències d'aquests fragments, s'alineen amb el genoma de l'espècie que es vol anotar. En funció de si els fragments seqüenciats provenien d'una regió que corresponia a un mateix exó o es trobava a la intersecció entre dos exons, ens donarà informació de la localització dels introns i dels llocs d'acoblament o només de la localització dels exons. Elaboració pròpia.

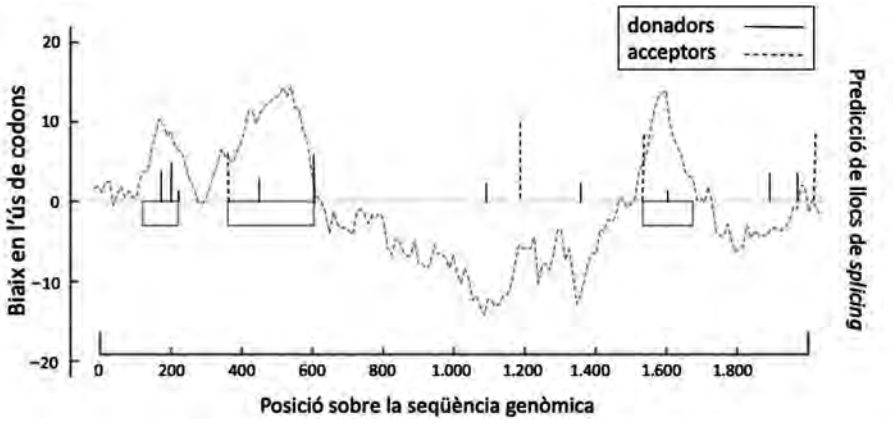
Si no s'ha emmascarat prèviament el genoma de cap espècie estretament relacionada, el millor enfocament és utilitzar un mètode *ab initio*. Això requereix més temps, però analitza la seqüència en detall i troba regions que es repeteixen en diferents regions del genoma, així com aquelles amb una composició esbiaixada a causa de les repeticions de seqüències curtes. El segon enfocament, basat en una biblioteca de regions repetitives de referència, és més ràpid, ja que busca regions del genoma que tinguin la mateixa seqüència que les proporcionades d'entrada, però també pot ser inexacte en cas que les regions repetitives del genoma que es vol emmascarar siguin molt diferents de les descrites per la biblioteca de regions repetitives utilitzada (Storer *et al.*, 2021).

Entre els mètodes que requereixen el genoma emmascarat i que aprofiten les bases de dades públiques, es troben els mètodes que alineen seqüències de proteïnes directament a la seqüència de DNA (Iwata i Gotoh, 2012). De manera similar a les dades d'RNA-seq, aquests alineaments són informatius per identificar les regions exòniques, però també els llocs d'acoblament potencials. Aquest és un mètode basat en l'homologia, i els models d'mRNA construïts amb aquestes dades poden ser fiables, encara que menys que els generats a partir de dades d'RNA-seq.

L'altra font de models de gens i d'mRNA són els predictors de gens *ab initio* (Scalzitti *et al.*, 2020). Es tracta de programes que, utilitzant només el conjunt del genoma emmascarat i un fitxer de paràmetres, prediuen les estructures gèniques d'aquesta seqüència. Aquests mètodes es basen en el biaix en l'ús de codons que exhibeixen els exons, és a dir, en el fet que en les regions que codifiquen proteïnes, a diferència de les no codificants, els diferents codons no apareixen amb freqüència idèntica (figura 3). El biaix en l'ús de codons en les seqüències codificants és conseqüència, en primer lloc, de l'ús desigual dels aminoàcids en les proteïnes. És a dir, en les proteïnes hi ha aminoàcids que apareixen amb molta més freqüència que d'altres (per exemple, en les proteïnes dels vertebrats, més del 8 % dels aminoàcids són lisines, mentre que només l'1 % són triptòfans). En segon lloc, el biaix és conseqüència de l'ús desigual de codons sinònims per un mateix aminoàcid. S'han desenvolupat nombrosos mètodes per mesurar aquests biaixos (Guigó, 1997).

Cadascun dels programes té les seves particularitats, però els fitxers de paràmetres requerits contenen informació similar per modelar, principalment, les regions de codificants, els llocs d'acoblament i els introns (figura 3). Idealment, aquests fitxers de paràmetres es generen amb un procés d'entrenament basat en l'ús d'mRNA reals, obtinguts de la mateixa espècie que es vol anotar, i ajustant els paràmetres progressivament fins a arribar a una bona predicció d'aquests gens coneguts. Un inconvenient d'aquest tipus de mètodes és que aquest entrenament només és possible quan hi ha dades experimentals disponibles per a aquella espècie. Hi ha diverses maneres d'abordar aquest problema. Un és desenvolupar mètodes d'autoentrenament que no requereixin cap conjunt de gens coneguts per generar el fitxer de paràmetres, però la solució més habitual és utilitzar fitxers de paràmetres entrenats en espècies taxonòmicament properes.

Durant els pròxims deu anys, gràcies a projectes com l'Earth BioGenome Project (EBP) (Lewin *et al.*, 2022) i els seus projectes afiliats, com la iniciativa catalana per a l'EBP (CBP, de



↑ Figura 3. Representació del biaix en l'ús dels codons i dels llocs d'acoblament en una seqüència. Utilitzant la seqüència d'un gen humà, veiem com el biaix en l'ús dels codons és més gran a les regions que es corresponen als exons del gen. Les línies verticals indiquen les puntuacions dels llocs d'acoblament, calculades utilitzant matrius de pesos posicionals (PWM, de l'anglès *position weight matrix*). Aquestes es calculen a partir de seqüències reals dels donadors i dels acceptors i després s'utilitzen per puntuar la probabilitat que una posició en concret sigui un donador o un acceptor, respectivament. Figura adaptada de la memòria *La codificació de la informació biològica en el genoma: Memòria llegida per l'acadèmic electe Dr. Roderic Guigó i Serra a l'acte de la seva recepció del dia 28 d'octubre de 2021*, núm. 1064 de les *Memòries de la Reial Acadèmia de Ciències i Arts de Barcelona* [en línia], <https://www.racab.cat/publicacions/memories/1064> [consulta: 15 març 2022].

l'anglès Catalan Initiative for the Earth BioGenome Project), seran desxifrats els genomes de centenars de milers d'espècies eucariotes. La seqüenciació, catalogació i caracterització d'aquests genomes proporcionarà per primera vegada una visió global de la vida a la Terra, i contribuirà a la identificació dels esdeveniments genòmics subjacents a les principals transicions durant la història de la vida: l'emergència dels eucariotes, de la pluricel·lularitat, de l'especialització de cèl·lules i d'òrgans, de la reproducció sexual, etc. També contribuirà a la comprensió de fenòmens biològics fonamentals, com ara el desenvolupament, la diferenciació (inclosa la regeneració), o aquells que estan involucrats en malalties i altres condicions humanes, com ara el càncer i l'envelliment.

Aquest coneixement, però, no pot ser inferit directament de la seqüència dels genomes, si abans els gens codificats en aquestes seqüències no han estat identificats. En aquest sentit, els mètodes eficients per a la identificació de gens són prerequisits per a projectes com l'EBP; tanmateix, alhora, són també la conseqüència d'aquests projectes. Efectivament, la continuïtat històrica que lliga totes les espècies que habiten el planeta queda reflectida en la semblança de la seva seqüència genòmica, una semblança que és molt més acusada en les regions que codifiquen proteïnes, és a dir, en els gens, els quals estan sotmesos a processos de selecció més forts que no pas les regions no gèniques. La conservació de regions funcionals ens proporciona una manera general d'identificar gens en seqüències de DNA: mitjançant la comparació de seqüències genòmiques d'espècies diferents, fins i tot en el cas que les seqüències que comparem siguin completament anònimes (és a dir, que en totes les seqüències que comparem desconeixem els gens que eventualment puguin estar-hi codificats). La identificació de regions conservades entre seqüències genòmiques de dues o més espècies ens pot revelar la presència de gens homòlegs en aquestes seqüències (Alexandersson *et al*., 2003; Korf *et al*., 2001; Wiehe *et al*., 2001). En conseqüència, a mesura que el nombre de genomes coneguts creixi, també ho farà la nostra capacitat per identificar-ne els gens en cadascun i per entendre, en conseqüència, com la biologia de les espècies està codificada en els seus genomes.

Bibliografia

Aken, B. L. [*et al.*] (2016). «The Ensembl gene annotation system». *Database: J. Biol. Databases Curation* (2016).

Alexandersson, M. [*et al.*] (2003). «SLAM: Cross-species gene finding and alignment with a generalized pair hidden Markov model». *Genome Res.*, 13 (3): 496-502.

Avery, O. T. [*et al.*] (1944). «Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Inductions of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III». *J. Exp. Med.*, 79 (2): 137-158.

Beadle, G. W.; Tatum, E. L. (1941). «Genetic control of biochemical reactions in neurospora». *Proc. Natl. Acad. Sci. USA*, 27 (11): 499-506.

Brenner, S. [*et al.*] (1961). «An unstable intermediate carrying information from genes to ribosomes for protein synthesis». *Nature*, 190 (4776): 576-581.

Brent, M. R. (2005). «Genome annotation past, present, and future: How to define an ORF at each locus». *Genome Res.*, 15 (12): 1777-1786.

Clarke, J. [*et al.*] (2009). «Continuous base identification for single-molecule nanopore DNA sequencing». *Nat. Nanotechnol.*, 4 (4): 265-270.

Crick, F. (1970). «Central dogma of molecular biology». *Nature*, 227 (5258): 561-563.

Frankish, A. [*et al.*] (2021). «GENCODE 2021». *Nucleic Acids Res.*, 49 (D1): D916-923.

Gamow, G. [*et al.*] (1956). «The problem of information transfer from the nucleic acids to proteins». *Adv. Biol. Med. Phys.*, 4: 23-68.

Guigó, R. (1997). «DNA composition, codon usage and exon prediction». A: *Genetic databases*. Cambridge (RU): Elsevier, 53-80.

Hancock, J. M. (2002). «Genome size and the accumulation of simple sequence repeats: Implications of new data from genome sequencing projects». *Genetica*, 115 (1): 93-103.

Harrow, J. [*et al.*] (2009). «Identifying protein-coding genes in genomic sequences». *Genome Biol.*, 10 (1): 201.

Hershey, A. D.; Chase, M. (1952). «Independent functions of viral protein and nucleic acid in growth of bacteriophage». *J. Gen. Physiol.*, 36 (1): 39-56.

Iwata, H.; Gotoh, O. (2012). «Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features». *Nucleic Acids Res.*, 40 (20): e161.

Jacob, F.; Monod, J. (1961). «Genetic regulatory mechanisms in the synthesis of proteins». *J. Mol. Biol.*, 3: 318-356.

Korf, I. [*et al.*] (2001). «Integrating genomic homology into gene structure prediction». *Bioinformatics*, 17, supl. 1: S140-148.

Lee, B. T. [*et al.*] (2022). «The UCSC Genome Browser database: 2022 update». *Nucleic Acids Res.* [en línia], 50, D1: D1115-D1122. <https://doi.org/10.1093/nar/gkab959>.

Lewin, H. A. [*et al.*] (2022). «The Earth BioGenome Project 2020: Starting the clock». *Proc. Natl. Acad. Sci. USA*, 119 (4): e2115635118.

Manni, M. [*et al.*] (2021). «BUSCO: Assessing genomic data quality and beyond». *Curr. Protoc.*, 1 (12): e323.

Morales, J. [*et al.*] (2022). «A joint NCBI and EMBL-EBI transcript set for clinical genomics and research». *Nature*, 604: 310-315.

Sanger, F. [*et al.*] (1977). «DNA sequencing with chain-terminating inhibitors». *Proc. Natl. Acad. Sci. USA*, 74 (12): 5463-5467.

Sanger, F.; Tuppy, H. (1951). «The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates». *Biochem. J.*, 49 (4): 481-490.

Scalzitti, N. [*et al.*] (2020). «A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms». *BMC Genomics*, 21 (1): 293.

Schadt, E. E. [*et al.*] (2010). «A window into third-generation sequencing». *Hum. Mol. Genet.*, 19 (R2): R227-240.

Simão, F. A. [*et al.*] (2015). «BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs». *Bioinformatics*, 31 (19): 3210-3212.

Storer, J. [*et al.*] (2021). «The Dfam community resource of transposable element families, sequence models, and genome annotations». *Mob. DNA*, 12 (1): 2.

Thibaud-Nissen, F. [*et al.*] (2013). «Eukaryotic genome annotation pipeline» A: *The NCBI Handbook* [en línia]. <https://www.ncbi.nlm.nih.gov/books/NBK169439/> [Consulta: 30 abril 2022].

Wang, Z. [*et al.*] (2009). «RNA-seq: A revolutionary tool for transcriptomics». *Nat. Rev. Genet.*, 10 (1): 57-63.

Watson, J. D.; Crick, F. H. (1953). «Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid». *Nature*, 171 (4356): 737-738.

Wiehe, T. [*et al.*] (2001). «SGP-1: Prediction and validation of homologous genes based on sequence alignments». *Genome Res.*, 11 (9): 1574-1583.