Sobre la naturalesa computacional de la vida

Discurs de presentació de Roderic Guigó i Serra com a membre numerari de la Secció de Ciències Biològiques, llegit el dia 11 de febrer de 2019

Institut d'Estudis Catalans

SECCIÓ DE CIÈNCIES BIOLÒGIQUES

Sobre la naturalesa computacional de la vida

Sobre la naturalesa computacional de la vida

Discurs de presentació de Roderic Guigó i Serra com a membre numerari de la Secció de Ciències Biològiques, llegit el dia 11 de febrer de 2019

Barcelona, 2022

Institut d'Estudis Catalans

SECCIÓ DE CIÈNCIES BIOLÒGIQUES

Biblioteca de Catalunya. Dades CIP

Guigó Serra, Roderic, autor
Sobre la naturalesa computacional de la vida : discurs de presentació de Roderic Guigó i Serra com a membre numerari de la Secció de Ciències Biològiques, llegit el dia 11 de febrer de 2019. — Primera edició
Bibliografia
ISBN 9788499656373
I. Institut d'Estudis Catalans. Secció de Ciències Biològiques II. Títol
1. Bioinformàtica 2. Codi genètic
575.112
575.116

Carrer del Carme, 47. 08001 Barcelona

Primera edició: gener de 2022

Text revisat lingüísticament per la Unitat d'Edició del Servei Editorial de l'IEC

Disseny de la coberta: Azcunce | Ventura

Compost per la Unitat de Producció del Servei Editorial de l'IEC
Imprès a Service Point FMI, SA

ISBN: 978-84-9965-637-3
Dipòsit Legal: B 1050-2022

DOI: 10.2436/10.1500.06.1

Ni el concepte de vida,
ni el concepte de computació són clars.[1]
Claus Emmeche, *The computational notion of life*, 1994

El caràcter fonamentalment tautològic de la vida
es deriva del simple fet que el mitjà és al mateix temps el fi.[2]
Guy Debord, *La societat de l'espectacle*, 1964

L'any 1944, a les acaballes de la Segona Guerra Mundial, Erwin Schrodinger, el físic que va contribuir al model quàntic de l'àtom, va publicar un llibre titulat *What is life?* (Schrodinger i Penrose, 2012 [1944]). En aquest llibre se suggereix per primera vegada que el material hereditari és un codi (Cobb, 2013) i que, com a tal, inclou les instruccions que determinen el desenvolupament i el funcionament dels éssers vius. Schrodinger escriu: «Són aquests cromosomes […] els que contenen en una mena de codi (*code-script*) el patró sencer del desenvolupament futur d'un individu i del seu funcionament en l'estat madur».[3] Depèn d'aquest codi que «l'ou fertilitzat es desenvolupi, en condicions adequades, en un gall de cua forcada, una gallina clapada, una mosca o una planta de blat de moro, un rododendre, un escarabat, un ratolí o una dona».[4] Tot i que en aquell moment encara no estava definitivament establert que el DNA era el material hereditari, i ni tan sols se'n coneixia l'es-

1. «Neither the concept of life, nor the concept of computation are that clear» (Emmeche, 1994). Totes les traduccions al català són de l'autor.

2. «Le caractère fondamentalement tautologique du spectacle découle du simple fait que ses moyens sont en même temps son but» (Debord, 1964). L'autor tradueix voluntàriament «l'espectacle» de l'original per «la vida».

3. «It is these chromosomes, or probably only an axial skeleton fibre of what we actually see under the microscope as the chromosome, that contain in some kind of code-script the entire pattern of the individual's future development and of its functioning in the mature state» (Schrodinger i Penrose, 2012 [1944], p. 21).

4. «In calling the structure of the chromosome fibres a code-script we mean that the all-penetrating mind, once conceived by Laplace, to which every causal connection lay immediately open, could tell from their structure whether the egg would develop, under suitable conditions, into a black cock or into a speckled hen, into a fly or a maize plant, a rhododendron, a beetle, a mouse or a woman» (Schrodinger i Penrose, 2012 [1944], p. 21).

tructura, Schrodinger suggereix que un codi capaç d'especificar l'enorme diversitat de formes i estructures que veiem en els éssers vius ha d'estar constituït per un nombre petit d'elements que s'ordenen en l'espai, i que és en l'ordre (en la seqüència), més que no pas en la naturalesa material dels elements, on resideixen les instruccions que configuren el codi. «Creiem que els gens —o potser la fibra cromosòmica sencera— són un sòlid aperiòdic».[5] «Una associació ordenada d'àtoms, dotada de suficient resistivitat per a mantenir l'ordre permanentment, sembla l'única estructura material concebible que pot oferir una varietat d'arranjaments ("isomèrics"), prou gran per tal de capturar un sistema complicat de "determinacions" dins de límits espacials molt petits. De fet, el nombre d'àtoms en una estructura així no cal que sigui gaire gran per tal de produir un nombre gairebé il·limitat de possibles arranjaments. Com a exemple, penseu en el codi Morse».[6] Schrodinger intueix, a més, allò que constitueix el tret singular d'un eventual codi de la vida: el codi és i especifica el maquinari necessari per a la seva interpretació, i escriu: «aquest codi miniatura […] ha de contenir, d'alguna manera, els mitjans per tal de posar-lo en operació».[7] Els cromosomes «són, al mateix temps, el codi de la llei i el poder executiu —o, per fer servir un altre símil, els plànols de l'arquitecte i les màquines del constructor».[8] És a dir, el codi és alhora el fi i el mitjà.

Haurien de passar deu anys encara abans que James Watson i Francis Crick publiquessin, l'any 1953, el famós article en què descriuen l'estructura del DNA (Watson i Crick, 1953*b*); el mateix any en què Frederick Sanger va publicar la primera seqüència d'aminoàcids d'una proteïna, la insulina bovina (Sanger i Thompson, 1953). Aquests descobriments confirmaven l'extraordinària intuïció de Schrodinger: el DNA (i les proteïnes) són efectivament molècules polimèriques constituïdes per la repetició moltes vegades d'un nombre limitat de molècules més elementals (quatre nucleòtids, vint aminoàcids), en la seqüència específica de les quals, més que no pas en la naturalesa material de les molècules que les conformen,

5. «We believe a gene —or perhaps the whole chromosome fibre— to be an aperiodic solid» (Schrodinger i Penrose, 2012 [1944], p. 40).

6. «A well-ordered association of atoms, endowed with sufficient resistivity to keep its order permanently, appears to be the only conceivable material structure that offers a variety of possible ("isomeric") arrangements, sufficiently large to embody a complicated system of "determinations" within a small spatial boundary. Indeed, the number of atoms in such a structure need not be very large to produce an almost unlimited number of possible arrangements. For illustration, think of the Morse code» (Schrodinger i Penrose, 2012 [1944], p. 61).

7. «What we wish to illustrate is simply that with the molecular picture of the gene it is no longer inconceivable that the miniature code should precisely correspond with a highly complicated and specified plan of development and should somehow contain the means to put it into operation» (Schrodinger i Penrose, 2012 [1944], p. 62).

8. «They are law-code and executive power —or, to use another simile, they are architect's plan and builder's craft— in one» (Schrodinger i Penrose, 2012 [1944], p. 22).

és on rau la capacitat codificant que tenen. No va ser, però, fins a un segon article publicat sis setmanes després (Watson i Crick, 1953*a*) que Watson i Crick fan referència explícitament —evocant conscientment o no les paraules de Schrodinger— a un possible codi encapsulat en la molècula de DNA: «En una molècula llarga d'aquesta mena moltes permutacions diferents són possibles i és, en conseqüència, versemblant que la seqüència precisa de bases és el codi que porta la informació genètica».[9] La relació entre aquesta seqüència de bases i la seqüència d'aminoàcids de les proteïnes va ser immediatament suggerida per George Gamow, que, el mateix any 1953, va proposar un model de síntesi de proteïnes, el qual tenia lloc directament en la seqüència de DNA. Aquest model, però, aviat va ser descartat, entre altres raons, perquè el DNA resideix en el nucli, i la síntesi de proteïnes té lloc en el citoplasma. Era clar que calia algun agent que transferís el codi resident en la seqüència del DNA des del nucli fins al citoplasma. De fet, també el mateix any 1953, Alexander Dounce ja va proposar que aquest agent era l'RNA: «Podria perfectament ser que les molècules gèniques d'àcid desoxiribonucleic actuessin com a patró per a la síntesi d'àcid ribonucleic, i que l'àcid ribonucleic sintetitzat en el patró gènic esdevingués, al seu torn, el patró per a la síntesi de proteïnes»[10] (Dounce, 1953). L'RNA missatger (mRNA), com a tal, però, no va ser descobert fins a l'any 1961 (Brenner, Jacob i Meselson, 1961; Gros *et al.*, 1961; Jacob i Monod, 1961). En la metàfora usada en aquella època, que reflectia els avenços tecnològics contemporanis, l'RNA missatger era com una «cinta magnètica» que copiava el codi del DNA i el portava fins als ribosomes, on era «interpretat» per tal de sintetitzar les proteïnes (Cobb, 2015). A partir de l'any 1961 i fins a l'any 1966, utilitzant mètodes dissenyats per Marshall W. Niremberg, Heinirch Matthaei i Philip Leder, que permetien utilitzar RNA sintètic per a dirigir la síntesi de proteïnes, es va establir la correspondència entre els seixanta-quatre triplets de nucleòtids, els codons, i els vint aminoàcids, l'anomenat *codi genètic* (Szymanski i Barciszewski, 2017).

La interrelació creixent entre biologia i computació

En paral·lel als avenços en biologia molecular, després de la Segona Guerra Mundial, se'n van produir en el camp de la informàtica. Possiblement, el primer ordinador digital, programable en memòria, un ordinador com els entenem avui en dia, va ser l'ENIAC (*electronic numerical integrator and computer*), el qual va entrar

9. «It follows that in a long molecule, many different permutations are possible, and it therefore seems likely that the precise sequence of the bases is the code which carries the genetical information.»

10. «It could conceivably happen that the deoxyribonucleic acid gene molecules would act as templates for ribonucleic acid synthesis, and that the ribonucleic acids synthesized on the gene templates would then in turn become templates for protein synthesis.»

en funcionament l'any 1945 (figura 1*a*). Tot i que l'ENIAC podia ser programat per a dur a terme càlculs diferents, la programació de l'ENIAC requeria la connexió manual dels seus components i, per tant, un coneixement profund de l'arquitectura de l'ordinador. Per tal d'obviar aquesta limitació, es van inventar els llenguatges de programació d'alt nivell, que permetien escriure en un llenguatge controlat les instruccions que conformen l'algoritme que ha d'executar l'ordinador, independentment de l'arquitectura que tingui. El primer llenguatge d'alt nivell àmpliament utilitzat va ser el FORTRAN, inventat l'any 1954 (figura 1*b*). Aquest llenguatge va permetre el desenvolupament dels programes computacionals, tal com avui els entenem. Inspirats segurament per aquesta terminologia, l'any 1961, Ernst Mayr, d'una banda (Mayr, 1961), i François Jacob i Jacques Monod, de l'altra (Jacob i Monod, 1961), utilitzen la metàfora de «programa genètic» com a anàleg de programa computacional (Peluffo, 2015). Mayr, en concret, suggereix, per primer cop, que aquest programa genètic és el resultat de l'evolució: «La selecció natural fa tot el que pot per a la producció d'un codi que garanteixi aquells comportaments que incrementen l'eficàcia biològica (*fitness*). […] L'acció voluntària d'un individu, en la mesura que es basa en les propietats del seu codi genètic, no és, per tant, més o menys propositiva que les accions d'una computadora que ha estat programada per a respondre de manera apropiada a diferents *inputs*».[11] Una idea recollida més tard per Jacob: «En el programa genètic, per tant, està escrit el resultat de totes les reproduccions passades, el recull de tots els èxits, atès que els fracassos han desaparegut. El missatge genètic, el programa dels organismes actuals, per tant, s'assembla a un text sense autor, el qual ha estat editant un corrector durant més de dos mil milions d'anys, millorant-lo contínuament, refinant-lo i completant-lo gradualment eliminant-ne totes les imperfeccions. Allò que avui és copiat i transmès per tal de garantir l'estabilitat de les espècies és aquest text, incessantment modificat pel temps».[12]

La terminologia bàsica de la biologia molecular que s'estableix durant aquells anys es nodreix de termes que provenen del camp de la informàtica: *codi genètic*,

11. «Natural selection does its best to favor the production of codes guaranteeing behavior that increases fitness. […] The purposive action of an individual, insofar as it is based on the properties of its genetic code, therefore is no more or less purposive than the actions of a computer that has been programmed to respond appropriately to various inputs.»

12. «In the genetic programme, therefore, is written the result of all past reproductions, the collection of successes, since all traces of failures have disappeared. The genetic message, the programme of the present-day organism, therefore, resembles a text without an author, that a proof-reader has been correcting for more than two billion years, continually improving, refining and completing it, gradually eliminating all imperfections. What is copied and transmitted today to ensure the stability of the species is this text, ceaselessly modified by time. Time, in this case, means the number of consecutive copies of the message, the number of successive generations leading from a remote ancestor to our present-day bacterial cell» (Jacob, 1973).

replicació, *transcripció*, *traducció*, *operó*, *programa*… Més enllà d'aquesta transferència terminològica, però, s'estableix al mateix temps una relació més substancial entre biologia i computació. És a partir d'un concepte nou, que emergeix també aquells anys. Es tracta del concepte *informació*; un concepte, per cert, tan poc clar com el de *vida* o el de *computació*. Els treballs de Claude Shanon[13] i Warren Weaber (1948) i Norbert Wiener (1948) a finals dels anys quaranta desenvolupen un marc conceptual en el qual el concepte *informació* «podia ser aplicat a qualsevol sistema, orgànic o inorgànic, viu o elèctric» (Cobb, 2013). Aquest marc conceptual va tenir un impacte enorme en moltes disciplines científiques i, en particular, en el pensament biològic. Al nostre país, per exemple, el doctor Ramon Margalef va ser un dels pioners en la introducció del concepte *informació* en ecologia (Margalef, 1957). I així, quan l'any 1958 Crick proposa el dogma central de la biologia molecular com a teoria unificada dels processos que condueixen des de la seqüència de DNA fins a la seqüència de les proteïnes, ho fa en termes explícitament informacionals. «Els dos conceptes centrals proposats […] foren els d'informació seqüencial i alfabets definits».[14] «Aleshores, el problema podia ser plantejat com la formulació de les regles generals que governen la transferència d'informació des d'un polímer amb un alfabet definit fins a un altre»[15] (Crick, 1970). La rellevància del concepte *informació* rau en el fet que ens permet fer abstracció del substrat material en què la informació és codificada. Les molècules de DNA, RNA i proteïnes són seqüències de símbols en alfabets diferents, i els processos mitjançant els quals la seqüència d'una d'aquestes molècules determina la seqüència d'una altra poden ser entesos com processos de transferència d'informació —substancialment independents de les propietats fisicoquímiques dels materials usats per a la transferència. Tot i que el concepte *informació* és absent en el seu llibre, Schrodinger potser intueix que cal un nou paradigma per a entendre les propietats emergents de la vida: «De tot allò que hem après sobre l'estructura de la matèria viva, hem d'estar preparats per a descobrir que funciona d'una manera que no pot ser reduïda a les lleis ordinàries de la física».[16] Efectivament, les lleis que governen la transferència d'informació no són contràries a les lleis de la física, però, fins a cert punt, en són independents.

13. La tesi doctoral de Shannon defensada l'any 1940 a l'Institut Tecnològic de Massachusetts es titulava *An algebra for theoretical genetics*.

14. «The two central concepts which had been produced, originally without any explicit statement of the simplification being introduced, were those of sequential information and of defined alphabets.»

15. «The principal problem could then be stated as the formulation of the general rules for information transfer from one polymer with a defined alphabet to another.»

16. «What I wish to make clear in this last chapter is, in short, that from all we have learnt about the structure of living matter, we must be prepared to find it working in a manner that cannot be reduced to the ordinary laws of physics» (Schrodinger i Penrose, 2012 [1944], p. 76).

Que la descripció de la vida com una computació sobre la informació continguda en la seqüència del DNA sigui una metàfora apropiada o útil és certament discutible i ha estat, de fet, àmpliament debatut (Emmeche, 1994). Restringida, però, als processos moleculars que condueixen des de la seqüència del DNA fins a la seqüència de les proteïnes és més que una metàfora (figura 1*c*). La replicació del DNA, la transcripció del DNA a l'RNA i la traducció de l'RNA a proteïnes (als quals cal afegir l'empalmament [*splicing*] de l'RNA en els organismes eucariotes) són computacions en el sentit més paradigmàtic del terme: processos que produeixen un *output* a partir d'un *input* (en el mateix alfabet o en un de diferent), d'acord amb un conjunt de regles que configuren un algoritme. En el cas de la traducció, per exemple, l'*input* és una seqüència d'RNA en un alfabet de quatre lletres, l'*output* és la seqüència d'una proteïna en un alfabet de vint lletres i l'algoritme d'acord amb el qual una seqüència determina l'altra és el codi genètic, que fa correspondre de manera (gairebé) unívoca grups de tres lletres en la seqüència d'RNA a una sola lletra en la seqüència d'aminoàcids (figura 2). Tret d'aquest algoritme que governa la traducció, però, els algoritmes que governen les altres computacions moleculars (la transcripció i l'empalmament, en particular) només els coneixem parcialment, i la seva caracterització és, en cert sentit, l'objectiu de la biologia molecular.

És sobretot perquè aquests processos són essencialment computacionals que la bioinformàtica —una disciplina científica, l'objectiu de la qual, en origen, era l'estudi de les molècules biològiques enteses gairebé exclusivament com a seqüències de símbols (absents de qualsevol referent fisicoquímic)— ha tingut un impacte extraordinari en la recerca biològica. Després que Sanger seqüenciés la insulina bovina, laboratoris d'arreu del món van començar a obtenir seqüències d'altres proteïnes. La disponibilitat de seqüències diferents en va fer possible la comparació i es va constatar que, tal com havia anticipat Mayr, eren efectivament portadores d'informació sobre la seva història evolutiva. Així, Emile Zuckerkandl i Linus Pauling publiquen l'any 1965, quan el nombre de seqüències d'aminoàcids conegudes era encara petit, un article amb el títol «Molecules as documents of evolutionary history» (Zuckerkandl i Pauling, 1965) en el qual proposen que la comparació de les seqüències d'una proteïna en diferents organismes pot contribuir a elucidar les seves relacions evolutives. L'acumulació creixent de seqüències de proteïnes conegudes va impulsar Margaret Dayhoff i els seus col·laboradors a compilar-les en els anomenats *Atlas of protein sequence and structure*. En la quarta edició, a finals dels anys seixanta, l'*Altas* contenia prop de tres-centes seqüències de proteïnes (figura 3*a*). Dayhoff va agrupar les proteïnes en famílies d'acord amb la similitud de la seva seqüència i en va construir l'alineament[17] (figura 3*b*). A

17. En bioinformàtica, un alineament (múltiple) és una organització matricial d'una col·lecció de seqüències, en la qual cada fila es correspon amb una seqüència diferent (per exemple, la

partir d'aquests alineaments, va derivar la matriu de les taxes amb les quals cada aminoàcid és substituït per un altre al llarg de l'evolució (Dayhoff i Schwartz, 1978) (figura 3*c*). Aquestes matrius, anomenades *matrius de substitució*, quantifiquen, per primer cop, la magnitud del canvi evolutiu i ofereixen un criteri objectiu per tal de generar alineaments de seqüències de manera automàtica. Per a obtenir aquests alineaments, es van desenvolupar durant els anys setanta els primers algoritmes computacionals, els quals estaven basats en la programació dinàmica (Needleman i Wunsch, 1970; Smith i Waterman, 1981). És amb la construcció de les matrius de substitució i el desenvolupament dels mètodes de comparació i alineament de seqüències que neix la disciplina de la bioinformàtica.[18] Per primera vegada de manera rellevant, un problema d'origen biològic es planteja en termes purament computacionals i es desenvolupen tècniques informàtiques específiques per tal de resoldre'l (Guigó, 2007). La biologia esdevé, efectivament, una ciència de la informació.

Els codis del DNA

Mentre el nombre de seqüències d'aminoàcids conegudes continuava augmentant, la seqüenciació d'àcids nucleics romania elusiva. No va ser fins a mitjan anys setanta que Sanger i Alan Coulson (Sanger i Coulson, 1975), d'una banda, i Allan Maxam i Walter Gilbert (Maxam i Gilbert, 1977), de l'altra, desenvolupen els primers mètodes eficients per a la seqüenciació dels àcids nucleics. Des d'aleshores el nombre de seqüències d'àcids nucleics no ha deixat de créixer exponencialment. A principi dels anys vuitanta, de manera similar a com va fer Dayhoff en el cas de les seqüències de proteïnes, s'estableixen, a Los Alamos National Laboratory, als Estats Units, i al Laboratori Europeu de Biologia Molecular, les bases de dades que emmagatzemaran totes les seqüències conegudes d'àcids nucleics. La compilació de seqüències de DNA i RNA (normalment seqüenciat com a DNA complementari, cDNA) permet localitzar els gens, és a dir, les regions que es transcriuen en la seqüència de DNA; comença la tasca, encara inacabada, de desxifrar el codi que regula la transcripció del DNA. Aquest codi és intrínsecament

seqüència de la mateixa proteïna en espècies diferents) i cada columna correspon a una posició «equivalent» en les seqüències. Sovint *equivalent* significa 'equivalent des del punt de vista evolutiu', és a dir, les posicions alineades corresponen a la mateixa posició en una seqüència ancestral, de la qual provenen les seqüències alineades.

18. El terme *bionformàtica* el van proposar per primer cop Paulien Hogeweg i Ben Hesper per referir-se a l'estudi de processos d'informació en sistemes biòtics, un significat diferent del que té avui en dia. A Pubmed, el primer article en el qual apareix el terme *bionformatics* és de l'any 1989 (D. R. Masys, 1989, «New directions in bioinformatics», *Journal of Research of the National Institute of Standards and Technology*, 94, p. 59-63).

diferent del codi que regula la traducció de l'mRNA, almenys en els organismes multicel·lulars, en el sentit que la traducció és essencialment robusta a l'entorn cellular; és a dir, un mateix mRNA missatger es tradueix, en general, a la mateixa proteïna en entorns (tipus) cel·lulars diferents. El codi que governa la transcripció del DNA, en canvi, depèn fortament de l'entorn cel·lular. Usant terminologia computacional, la interpretació del codi (transcripcional) depèn del maquinari que l'interpreta. De fet, és la interpretació diferencial de les instruccions codificades en la seqüència del DNA, és a dir, la transcripció diferencial de gens, la qual, durant el desenvolupament, dona lloc a tipus cel·lulars diferents, i és la responsable del manteniment d'aquests tipus en l'organisme desenvolupat. En les regions del genoma que es troben entre els gens, les regions intergèniques, les quals en alguns organismes eucariotes representen la majoria, fins i tot més del 90 %, del DNA, resideixen els elements que controlen la transcripció. En particular, precedint els gens, hi ha les anomenades *regions promotores*. Aquestes regions contenen uns motius específics, als quals s'uneixen els factors de transcripció, les proteïnes que activen (o reprimeixen) la transcripció dels gens. L'arranjament combinatorial d'aquests motius en la regió promotora és característic de cada gen, cosa que li confereix un patró transcripcional específic i propi, el qual està modulat per la presència o absència dels factors de transcripció corresponents en un entorn cel·lular determinat (figura 4*a*).

David Pribnow va ser un dels primers a postular explícitament que l'activació de la transcripció d'un gen estava governada per la seqüència específica de la seva regió promotora: «La transcripció d'un gen comença a un lloc específic anomenat *promotor*. En aquest lloc, l'RNA polimerasa interacciona amb el DNA per tal d'iniciar la síntesi de la molècula d'RNA. Com reconeix la polimerasa el promotor? La resposta potser es troba en la seqüència de nucleòtids del DNA».[19] Pribnow va comparar la seqüència dels sis promotors coneguts aleshores en el bacteri *Escherichia coli*, va observar que el motiu TATA o similar era comú a tots i va suggerir que aquest motiu estava implicat genèricament en la unió de l'RNA polimerasa al DNA (Pribnow, 1975; Schaller, Gray i Herrmann, 1975) (figura 4*b*). L'homòleg eucariota d'aquest motiu és la caixa TATA (TATA box), amb la seqüència TATA(A/T)A(A/T)(A/G) que va ser descoberta l'any 1978, també a través de l'anàlisi i la comparació de regions promotores en diversos metazous (Lifton *et al.*, 1978). Amb l'increment de les seqüències de DNA conegudes, un nombre creixent de motius corresponents a la unió de factors de transcripció di-

19. «The transcription of a gene begins at a specific site called a promoter. At such a site the RNA polymerase interacts with the DNA in order to initiate the synthesis of an RNA molecule. How does the polymerase recognize a promoter? The answer might be found in the nucleotide sequence of the DNA.»

ferents en tipus cel·lulars i espècies diferents van ser descoberts i caracteritzats, i els primers catàlegs de motius van ser publicats i organitzats en bases de dades (Wingender, 1988). Aquestes col·leccions, en les quals cada motiu de DNA és assignat a un factor de transcripció, són, en cert sentit, l'anàleg al codi genètic per a la transcripció. La complexitat d'aquest codi, molt més gran que la del codi traduccional, es va fer evident de seguida i ja des del principi es va recórrer a models lingüístics i gramaticals per a investigar-lo (Collado-Vides, 1989). En primer lloc, és un codi molt més degenerat que el codi genètic, en el sentit que motius diferents (tot i que, en general, exhibeixen un cert grau de similitud) són funcionalment equivalents, és a dir, uneixen el mateix factor de transcripció. La fortalesa d'aquesta unió, però, depèn de la seqüència específica de nucleòtids en el motiu, de manera que seqüències lleugerament diferents poden contribuir de manera diferent a l'activació de la transcripció de gens diferents. En segon lloc, la interpretació de la seqüència de DNA no és directa, sinó que està mediada per modificacions (anomenades *epigenètiques*) en aquesta seqüència (la metilació dels nucleòtids), i en les histones, les proteïnes que compacten el DNA en la cromatina. Aquestes modificacions, que en part són específiques del tipus cel·lular, afavoreixen o desfavoreixen l'obertura de la cromatina, la qual ha de permetre la unió dels factors de transcripció al DNA i l'inici de la transcripció. En tercer lloc, i potser més important encara, perquè separa fonamentalment el codi transcripcional dels codis computacionals, el codi transcripcional no és pas adimensional, ni tan sols linear, sinó tridimensional. És cada cop més evident que, a més de les regions promotores, altres regions en la seqüència del genoma estan involucrades en la regulació de la transcripció dels gens. Aquestes regions, anomenades *estimuladors* (*enhancers*), poden estar linealment molt allunyades dels gens que regulen, fins i tot, en organismes eucariotes, poden residir en cromosomes diferents. A través del plegament dels cromosomes en l'espai nuclear, però, els estimuladors entren físicament en contacte amb les regions promotores dels gens, de manera que contribueixen a regular-ne l'activitat transcripcional (figura 4*a*). És a dir, el codi que regula la transcripció d'un gen no està totalment encapsulat en la regió proximal promotora, sinó que la seva interpretació requereix interaccions físiques amb regions distals del genoma. Aquestes interaccions són diferents en diferents tipus cel·lulars i contribueixen a definir el patró transcripcional dels gens, que és específic de cada tipus cellular. És com si el programari dels ordinadors es reconfigurés per tal de produir un *output* diferent a partir d'un mateix *input*. Dit d'una altra manera, *outputs* diferents s'obtenen no com a conseqüència del processament de programes diferents amb el mateix ordinador, sinó com a conseqüència del processament del mateix programa amb ordinadors diferents.

Entre la transcripció i la traducció, en la majoria d'organismes eucariotes, hi ha un altre procés important, el qual està també fortament governat per la infor-

mació continguda en la seqüència d'àcids nucleics. Es tracta de l'empalmament, el procés nuclear mitjançant el qual regions internes del transcrit d'RNA, anomenades *introns*, són eliminades, i les regions que romanen, els exons, són reenganxades per a formar la seqüència de l'RNA missatger, la qual és la que eventualment serà traduïda a proteïna. Els punts de tall dels introns (els llocs d'empalmament) estan definits per motius de seqüència extraordinàriament conservats: el dinucleòtid GT apareix sempre al principi (extrem 5′) dels introns i el dinucleòtid AG, al final (extrem 3′) (figura 5*a*). L'any 1981, Stephen Mount va alinear la seqüència genòmica al voltant dels cent trenta llocs d'empalmament que es coneixien en aquell moment, i va observar que la conservació s'estenia, tot i que més dèbilment, més enllà d'aquests dos dinucleòtids. Va deduir-ne les seqüències de consens (és a dir, les seqüències que inclouen els nucleòtids que apareixen amb més freqüència a cada posició al voltant dels llocs d'empalmament) (figura 5*b*). En el cas del lloc d'empalmament de l'extrem 5′ de l'intró, la seqüència consens és CAG|GTAAGT (en què el símbol | denota la frontera entre l'exó i l'intró). Aquesta seqüència és exactament la complementària a la seqüència d'un molècula d'RNA (un membre de la família dels snRNA, de l'anglès *small nuclear RNA*) que forma part d'un complex d'RNA i proteïnes (una ribonucleoproteïna) que reconeix el lloc d'empalmament 5′ (figura 5*a*). Tot i que aquesta és, aparentment, la seqüència òptima per a definir els llocs d'empalmament 5′, només apareix exactament en poc més de l'1 % de tots els llocs d'empalmament 5′ en el genoma humà, i ni tan sols és la seqüència més freqüent. En aquest sentit, podríem dir que el codi de l'empalmament és subòptim, és a dir, entre seqüències sinònimes (funcionalment equivalents) no s'utilitza preferentment aquella que pot dur a terme la funció de la manera més eficient. Aquesta suboptimalitat té conseqüències biològiques importants, en fer possible el fenomen de l'empalmament alternatiu, mitjançant el qual combinacions diferents d'exons dins el mateix gen donen lloc a mRNA diferents (figura 6*a*). L'empalmament alternatiu incrementa de manera substancial la capacitat codificant del genoma, en fer possible la generació d'un gran nombre d'mRNA missatgers (i, potencialment, de proteïnes) a partir d'un nombre més reduït de gens. La selecció de llocs d'empalmament alternatius, subòptims, està regulada per motius auxiliars, normalment situats en les regions intròniques dels gens (Fairbrother *et al*., 2002), que són reconeguts per unes proteïnes que s'uneixen a l'RNA i s'anomenen *factors d'empalmament* (figura 6*b*). Aquests motius, juntament amb els llocs canònics d'empalmament, constitueixen el nucli del codi de l'empalmament. L'abundància dels diferents factors d'empalmament en un tipus cel·lular concret determina el patró d'empalmament alternatiu dels gens en aquell tipus cel·lular. Tot i que s'han desenvolupat mètodes molt sofisticats per a caracteritzar el codi de l'empalmament (Barash *et al*., 2010), l'àlgebra mitjançant la qual combinacions de motius reguladors

de l'empalmament en la seqüència de l'RNA primari d'un gen, en cooperació amb els llocs d'empalmament pròpiament, produeixen mRNA diferents en tipus cel·lulars diferents ens és majoritàriament desconeguda.

Els ordinadors basats en DNA

Com hem vist, en el desxiframent del codi de la transcripció i de l'empalmament, com ara també en el desxiframent d'altres codis que regulen altres processos en la ruta que condueix de la seqüència del DNA del genoma a la seqüència d'aminoàcids de les proteïnes, hi ha tingut un paper essencial la concepció d'aquests processos com a computacions de seqüències d'àcids nucleics. Aquesta concepció no és només una metàfora útil, com podria semblar, sinó que el DNA té efectivament la capacitat de calcular. Això ho va demostrar per primer cop Leonard Adleman en l'article «Molecular computation of solutions to combinatorial problems» (Adleman, 1994). Adleman va abordar un problema clàssic en ciències de la computació: el problema del viatjant de comerç. En la versió que Adleman va resoldre, el problema es pot formular de la manera següent: «donada una llista de ciutats i de les connexions entre elles (carreteres, vies de tren, rutes d'avió…), i donada una ciutat d'inici i una de finalització, hi ha algun camí que passi per totes i cadascuna de les ciutats i ho faci només una vegada?». En teoria de grafs, aquest problema es representa com un graf dirigit,[20] en el qual les ciutats són els vèrtexs i les connexions són les arestes. Un camí que comença en un vèrtex d'inici i acaba en un de finalització, passant per tots i cadascun dels nodes, es diu que és *hamiltonià*. El problema que va abordar Adleman era el de determinar si, donat un graf dirigit (i vèrtex d'inici i finalització), hi ha algun camí hamiltonià (figura 7*a*). Tot i que aparentment és un problema molt simple, pertany a la classe de problemes anomenats *NP-complets*, és a dir, per als quals no es coneix un algoritme que, per a resoldre'l, el nombre de càlculs no creixi de manera exponencial amb la mida de l'*input* (en aquest cas, el nombre de ciutats). En la pràctica, això significa que per a un nombre molt gran de ciutats no hi ha prou capacitat computacional per a resoldre el problema, ni tan sols fent servir, si això fos possible, tots els recursos disponibles actualment al món. Adleman va proposar l'algoritme següent per resoldre aquest problema (figura 7*b*):

— Pas 1: cal generar (molts) camins aleatoris en el graf.

— Pas 2: cal mantenir només aquells camins que comencen pel vèrtex d'inici i acaben al vèrtex de finalització.

20. És a dir, les connexions entre les ciutats tenen una direcció; potser es pot anar de la ciutat A a la B, pero no de la B a la A perquè, per exemple, no hi ha vols disponibles.

— Pas 3: cal mantenir només els camins que contenen exactament un nombre de vèrtexs exactament igual al nombre de vèrtexs del graf.

— Pas 4: cal mantenir només els camins que passen per tots els vèrtexs del graf almenys una vegada.

— Pas 5: si queda algun camí, aleshores el graf té un camí hamiltonià.

Certament, no es tracta d'un algoritme particularment intuïtiu, però Adleman el va dissenyar així perquè es pogués implementar utilitzant molècules de DNA i els enzims i les proteïnes que les processen. Va aplicar aquesta implementació a un graf constituït per set ciutats i un conjunt de trajectes entre elles. Per tal d'implementar el pas 1, Adleman va codificar les ciutats (vèrtexs) mitjançant seqüències de DNA de vint nucleòtids. Si dues ciutats estaven connectades en una direcció determinada, va codificar el trajecte (aresta) entre la ciutat d'origen i la ciutat de destinació mitjançant una seqüència de nucleòtids també de vint nucleòtids, en la qual els deu primers nucleòtids eren idèntics als deu darrers de la primera ciutat i els deu darrers idèntics als deu primers de la segona ciutat (figura 7*c*). Va sintetitzar aproximadament 3×10^{13} còpies de cadascuna de les seqüències complementàries a les seqüències que representen les ciutats i de cadascuna de les seqüències corresponents als trajectes i les va barrejar en una reacció de lligació. Les propietats de complementarietat de DNA fan que aquesta reacció generi molècules més llargues, resultat de la concatenació de (seqüències corresponents a) ciutats connectades per trajectes (figura 7*c*, pas 1, de 1 a 4). Per a implementar el segon pas de l'algoritme, va utilitzar la reacció en cadena per la polimerasa (PCR, de l'anglès *polymerase chain reaction*) amb encebadors dissenyats contra la ciutat inicial i la ciutat final, de manera que, del resultat de la reacció de lligament, seleccionava només aquells trajectes que començaven per la ciutat inicial i acabaven a la ciutat final (figura 7*d*, pas 2). Per implementar el tercer pas de l'algoritme, va córrer un gel d'electroforesi amb les seqüències resultants del pas anterior i va seleccionar la banda de DNA d'exactament 140 parells de bases —que correspon als trajectes que passen exactament per set ciutats (figura 7*d*, pas 3). Per implementar el quart pas de l'algoritme, va generar les seqüències corresponents a cadascuna de les ciutats conjugades a microesferes magnètiques. Fent servir el conjugat corresponent a la primera ciutat, va seleccionar entre les seqüències resultants del pas anterior les que contenien la primera ciutat. A continuació, entre aquestes seqüències, va seleccionar les que contenien la segona ciutat fent servir el conjugat corresponent, i així successivament (figura 7*d*, pas 4). Si al final del procés quedava alguna molècula, aleshores el graf contenia un camí hamiltonià. Òbviament, l'objectiu d'Adleman no era dissenyar un ordinador molecular capaç de ser utilitzat en la pràctica, sinó demostrar que el DNA té capacitat de calcular, i que els ordinadors basats en DNA podrien tenir alguns avantatges, comparats amb els ordinadors

actuals basats en silici, com ara la gran capacitat de paral·lelisme, i l'eficiència energètica.

L'arquitectura de l'ordinador molecular dissenyat per Adleman, però, permetia resoldre només un problema, en contraposició als ordinadors digitals, els quals poden ser programats per a resoldre una gran varietat de problemes. L'any 2001, l'equip d'Elud Shapiro va dissenyar, per primer cop, un ordinador programable basat en DNA i en enzims que el manipulen (Benenson *et al.*, 2001). En principi, utilitzant aquest ordinador seria possible resoldre qualsevol problema resoluble en un ordinador digital. Tot un camp de recerca, la computació molecular, que explora les possibilitats d'ordinadors basats en DNA, ha emergit des d'aleshores. Potser no com a ordinadors genèrics, però, per a algunes aplicacions específiques, els ordinadors basats en DNA podrien constituir eventualment una alternativa als ordinadors basats en silici. Algunes aplicacions potencialment importants serien en el camp de la medicina, en què ordinadors moleculars podrien funcionar dins de les cèl·lules i, d'acord amb l'estat d'aquestes cèl·lules, prendre decisions terapèutiques, que es podrien implementar mitjançant la síntesi d'RNA. El disseny de xarxes neuronals basades en DNA és una altra àrea que està sent explorada actualment (Cherry i Qian, 2018).

És possible dissenyar sistemes de computació basats en sistemes biològics diferents del DNA. Tot i que la discussió d'aquests sistemes no és l'objectiu d'aquest discurs, voldria només mencionar el treball dels nostres col·legues Francesc Posas, Ricard Solé i col·laboradors, els quals van dissenyar cèl·lules de llevat que implementaven funcions lògiques simples, les quals, en ser combinades en xarxes multicel·lulars, permetien implementar circuits lògics més complexos (Regot *et al.*, 2011).

Potser, però, més que com a mitjà per a computar, el DNA podria esdevenir una alternativa pràctica com a medi per a gravar, emmagatzemar i recuperar informació. Aquesta idea, proposada per primer cop per Mikhail Neiman ja l'any 1964,[21] no va ser implementada en la pràctica fins a l'any 2012, quan els grups de George Church i col·laboradors (Church, Gao i Kosuri, 2012) i de Nick Goldman i collaboradors (Goldman *et al.*, 2013) van ser capaços de gravar textos i fotografies sintetitzant molècules de DNA, i de recuperar-los després seqüenciant aquestes molècules. Amb el disseny de sistemes de codificació tolerants a la taxa d'error inherent a la síntesi i còpia del DNA va ser possible recuperar la informació amb un 100 % de precisió. Aquest estudi va demostrar, a més, que el DNA té una gran capacitat de compressió, ja que aconsegueix emmagatzemar 2 petabytes (PB) d'informació per gram de DNA, la qual cosa és un ordre de magnitud més gran que la capacitat d'emmagatzematge de les cintes magnètiques, els dispositius d'emmagat-

21. *https://en.wikipedia.org/wiki/DNA_digital_data_storage* (consulta: juny 2020).

zematge d'informació disponibles avui.[22] Recentment, Yaniv Erlich i Dina Zielinski han aconseguit incrementar encara més aquesta capacitat i han dissenyat un mètode de codificació, inspirat en els que s'utilitzen en la radiodifusió mòbil, capaç d'emmagatzemar 215 PB per gram de DNA (Erlich i Zielinski, 2017). Aquesta densitat s'aproxima a la capacitat de Shannon[23] d'emmagatzematge amb DNA, que assoleix el 85 % del límit teòric (el qual és de 2 bits per nucleòtid).

Una limitació important d'aquests mètodes d'emmagatzematge basats en DNA és que, per a accedir a la informació, cal fer-ho de manera seqüencial (és a dir, per a recuperar un segment d'informació concret, per exemple, un arxiu determinat dins un conjunt d'arxius, cal llegir de manera seqüencial tota la informació emmagatzemada fins a arribar al segment que es vol recuperar). Això significa, en la pràctica, la seqüenciació de tot el DNA que ha calgut sintetitzar per a emmagatzemar la informació. Això és impracticable, sobretot si hom vol emmagatzemar grans quantitats d'informació. Els ordinadors digitals fan servir, en general, un mode d'accés a la informació diferent, que s'anomena *aleatori*, de manera que es pot accedir a qualsevol element emmagatzemat directament. Això requereix que els elements siguin adreçables, és a dir, que disposin d'una adreça que pugui ser reconeguda pel dispositiu que llegeix o que recupera la informació emmagatzemada. Recentment, Organick i col·laboradors han dissenyat un dispositiu d'emmagatzematge en DNA d'accés aleatori, el qual utilitza encebadors de PCR que reconeixen adreces escrites en DNA per a accedir directament a arxius específics (Organick *et al.*, 2018). D'aquesta manera, han pogut emmagatzemar (i recuperar) 200 MB d'informació, la quantitat més gran d'informació emmagatzemada fins ara en un dispositiu molecular. Aquesta informació inclou un vídeo d'alta definició, imatges, àudio i text. Entre els textos emmagatzemats hi ha la Declaració Universal dels Drets Humans en més de cent llengües diferents, incloent-hi el català.

El principal problema per a la implementació pràctica de dispositius moleculars d'emmagatzematge de la informació és el cost actual de la síntesi de DNA. Per exemple, el cost del mètode d'Erlic i Zielinski per a emmagatzemar 2 MB d'informació va ser de 7.000 dòlars.[24] El DNA, però, és una molècula molt estable i, en conseqüència, l'emmagatzematge en DNA pot esdevenir un mètode cost-efectiu quan hom vol mantenir la informació durant llargs períodes de temps. Goldman i col·legues van estimar que, amb els costos actuals de la síntesi de DNA, aquest se-

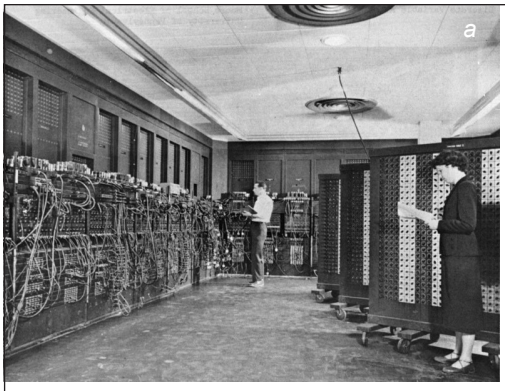
22. L'any 2018, els dispositius més eficients són capaços d'emmagatzemar aproximadament 1 PB per kg.

23. Capacitat d'un canal d'informació: la capacitat màxima d'informació que pot ser transmesa de manera fiable per un canal.

24. Avui el cost per TB en un disc dur comercial és de 30 dòlars (és a dir, cent milions de vegades més econòmic que en el dispositiu d'Erlich i Zelinski).

ria el cas si es vol emmagatzemar la informació durant més de mil anys. Tot i així, atès el ràpid progrés que es produeix avui en dia en les tecnologies que fan possible la manipulació del DNA (incloent-hi la seqüenciació i la síntesi), l'emmagatzematge d'informació en DNA té el potencial de convertir-se en una alternativa real, o, almenys, un complement, a l'emmagatzematge en cintes magnètiques.

Si, més enllà dels processos moleculars bàsics en els quals es fonamenta, la metàfora de vida com a computació sobre la seqüència del DNA és apropiada o útil, requereix una reflexió molt més profunda. Restringida, però, a aquests processos, és a dir, al moment en el qual els nucleòtids s'organitzen en la seqüència de DNA —el moment en què emergeix la vida— la frontera entre vida i computació certament s'esborra.



a

```
C       THE TPK ALGORITHM                                    b
C       FORTRAN IV STYLE
        DIMENSION A(11)
        FUN(T) = SQRT(ABS(T)) + 5.)*T**3
        READ (5,1) A
  1     FORMAT(5F10.2)
        DO 10 J = 1, 11
           I = 11 - J
           Y = FUN(A(I+1))
           IF (400.0-Y) 4, 8, 8
  4            WRITE (6,5) I
  5            FORMAT(I10, 10H TOO LARGE)
               GO TO 10
  8            WRITE(6,9) I, Y
               FORMAT(I10, F12.6)
 10     CONTINUE
        STOP
        END
```

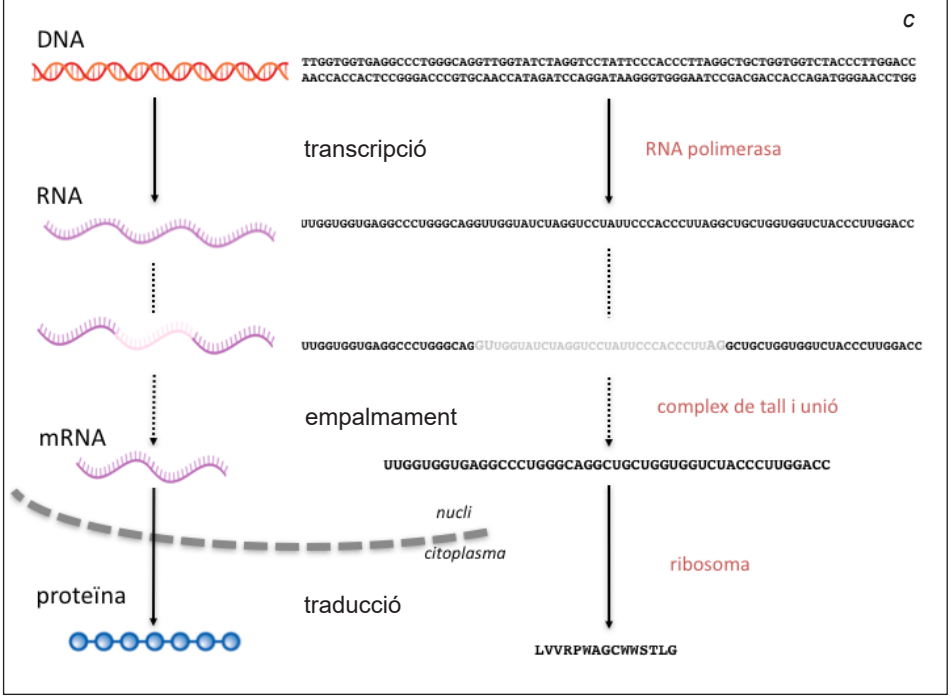


Figura 1. Informàtica i biologia molecular. *a*) Hom considera l'ENIAC (*electronic numerical integrator and computer*) el primer ordinador electrònic, digital, programable i d'ús general —el primer dels ordinadors actuals. Va començar a funcionar l'any 1945 a la Universitat de Pennsilvània i va continuar funcionant fins a l'any 1955. Pesava unes 30 tones i feia aproximadament 2,4 m × 0,9 m × 30 m. En la fotografia veiem Glen Beck (al fons) i Betty Snyder programant l'ENIAC (fotografia de l'Exèrcit dels Estats Units d'autor desconegut, domini públic, *https://commons.wikimedia.org/w/index.php?curid=55124*). L'ENIAC havia de ser, efectivament, programat de manera manual, connectant físicament els diferents components de l'ordinador, connexions que eren diferents segons el programa que es volia executar. La programació de l'ENIAC requeria, en conseqüència, un bon coneixement de la seva arquitectura. *b*) Amb la invenció dels llenguatges de programació d'alt nivell se separava l'ús dels ordinadors del coneixement de la seva arquitectura. Hom podia proporcionar les instruccions a un ordinador (programari) sense conèixer com estava construït (maquinari). En la imatge, un programa escrit en FORTRAN, un dels primers llenguatges de programació d'alt nivell. Desenvolupat per IBM durant els anys cinquanta, FORTRAN va esdevenir ràpidament el llenguatge de programació dominant en l'àmbit de la recerca científica. La figura il·lustra la implementació (treta del Fortran Specialist Group, *https://fortran.bcs.org*) de l'algoritme TPK. Aquest algoritme va ser introduït per Donald Kuth i Luis Trabb Pardo per a il·lustrar l'evolució dels llenguatges de programació pel que fa a les estructures de dades i les instruccions lògiques i aritmètiques. *c*) La descodificació de la informació en el genoma comença amb la transcripció del DNA a RNA. Només les regions del genoma que corresponen als gens es transcriuen a molècules RNA. En les cèl·lules eucariotes, aquests transcrits d'RNA experimenten una sèrie de modificacions posttranscripcionals. Entre aquestes modificacions hi ha l'empalmament, mitjançant el qual fragments de la seqüència dels transcrits inicials d'RNA, els introns, són eliminats, i els fragments que romanen, els exons, són enganxats per a donar lloc a l'RNA missatger (mRNA). Els mRNA s'exporten al citoplasma, on són traduïts a proteïna.[25] Cada etapa en aquest procés de transferència d'informació des de la seqüència de DNA fins a la seqüència de les proteïnes es pot considerar una computació, en la qual l'*input* és una seqüència en un alfabet determinat, l'*output* una seqüència en el mateix alfabet o en un altre i hi ha un maquinari que interpreta el codi que determina l'*output* en funció de l'*input*. En el cas de la transcripció, l'*input* és una seqüència en l'alfabet del DNA (A,C,G,T) i l'*output* una seqüència en l'alfabet de l'RNA (A,C,G,U); el maquinari és constituït per l'RNA polimerasa i el conjunt de factors i proteïnes responsables de la transcripció d'un gen en unes condicions determinades (els factors de transcripció). Aquests factors s'uneixen a motius específics en la seqüència del DNA, els quals constitueixen una mena de codi transcripcional. En el cas de l'empalmament, l'*input* i l'*output* són seqüències en el mateix alfabet de l'RNA, però en l'*output* s'han eliminat les seqüències que corresponen als introns. El maquinari és constituït per un conglomerat de proteïnes i RNA anomenat *complex de tall i unió* (*splicesoma*), que reconeix els llocs de tall dels introns mitjançant motius específics en la seqüència de l'RNA, els quals constitueixen una mena de codi de l'empalmament. Finalment, en el cas de la traducció, l'*input* és una seqüència en l'alfabet de l'RNA i l'*output* una seqüència en l'alfabet de les proteïnes, i el maquinari és el ribosoma que reconeix les instruccions codificades en la seqüència de l'mRNA d'acord amb l'anomenat *codi genètic* (vegeu la figura 2).

25. No tots els transcrits processats després de l'empalmament donen lloc a mRNA. Alguns no són traduïts a proteïna i s'anomenen *RNA llargs no codificants* (lncRNA, de l'anglès *long non coding RNA*), els quals poden quedar-se al nucli o ser també exportats al citoplasma.

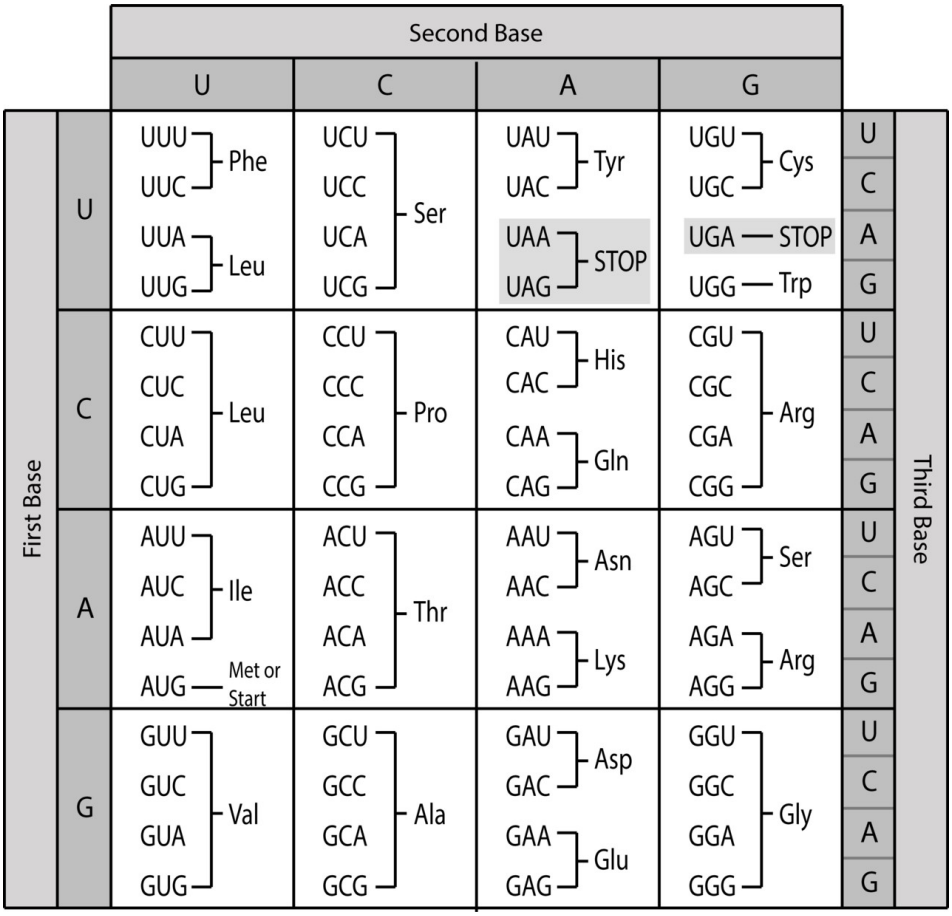

Second Base
U
C
A
G
First Base
Third Base
U
C
A
G
UUU Phe
UUC
UUA Leu
UUG
UCU
UCC
UCA
UCG Ser
UAU Tyr
UAC
UAA STOP
UAG
UGU Cys
UGC
UGA STOP
UGG Trp
CUU
CUC
CUA
CUG Leu
CCU
CCC
CCA
CCG Pro
CAU His
CAC
CAA Gln
CAG
CGU
CGC
CGA
CGG Arg
AUU
AUC
AUA Ile
AUG Met or Start
ACU
ACC
ACA
ACG Thr
AAU Asn
AAC
AAA Lys
AAG
AGU Ser
AGC
AGA
AGG Arg
GUU
GUC
GUA
GUG Val
GCU
GCC
GCA
GCG Ala
GAU Asp
GAC
GAA Glu
GAG
GGU
GGC
GGA
GGG Gly

FIGURA 2. El codi genètic. Equivalència entre els triplets a la seqüència dels mRNA i els aminoàcids a la seqüència de les proteïnes. Tret d'*https://commons.wikimedia.org/wiki/File:Genetic_Code.png*.

a

HUMAN GROWTH HORMONE

```
                5         10        15        20        25        30
  1 F P T I P L S R L F D N A M L R A H R L H Q L A F D T Y Q E
 31 F E E A Y I P K E Q K Y S F L Q N P Q T S L C F S E S I P T
 61 P S N R E E T Q Q K S N L Q L L R I S L L L I Q S W L E P V
 91 Q F L R S V F A N S L V Y G A S N S D V Y D L L K D L E E G
121 I Q T L M G R L E D G S P R T G Q I F K Q T Y S K F D T N S
151 H N D D A L L K N Y G L L Y C F R K D M D K V E T F L R I V
181 Q C R S V E G S C G F
```

COMPOSITION

7	ALA A	14	GLN Q	26	LEU L	18	SER S
11	ARG R	13	GLU E	9	LYS K	10	THR T
9	ASN N	8	GLY G	3	MET M	1	TRP W
11	ASP D	3	HIS H	13	PHE F	8	TYR Y
4	CYS C	8	ILE I	8	PRO P	7	VAL V

b

```
mosca          GAKKVIISAP SAD.APM..F VCGVNLDAYK PDMKVVSNAS CTTNCLAPLA
humà           GAKRVIISAP SAD.APM..F VMGVNHEKYD NSLKIISNAS CTTNCLAPLA
planta         GAKKVIISAP SAD.APM..F VVGVNEHTYQ PNMDIVSNAS CTTNCLAPLA
bacteri        GAKKVVMTGP SKDNTPM..F VKGANFDKY. AGQDIVSNAS CTTNCLAPLA
llevat         GAKKVVITAP SS.TAPM..F VMGVNEEKYT SDLKIVSNAS CTTNCLAPLA
arqueobacteri  GADKVLISAP PKGDEPVKQL VYGVNHDEYD GE.DVVSNAS CTTNSITPVA
```

c

	C Cys	S Ser	T Thr	P Pro	A Ala	G Gly	N Asn	D Asp	E Glu	Q Gln	H His	R Arg	K Lys	M Met	I Ile	L Leu	V Val	F Phe	Y Tyr	W Trp
C Cys	12																			
S Ser	0	2																		
T Thr	-2	1	3																	
P Pro	-3	1	0	6																
A Ala	-2	1	1	1	2															
G Gly	-3	1	0	-1	1	5														
N Asn	-4	1	0	-1	0	0	2													
D Asp	-5	0	0	-1	0	1	2	4												
E Glu	-5	0	0	-1	0	0	1	3	4											
Q Gln	-5	-1	-1	0	0	-1	1	2	2	4										
H His	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F Phe	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
	C Cys	S Ser	T Thr	P Pro	A Ala	G Gly	N Asn	D Asp	E Glu	Q Gln	H His	R Arg	K Lys	M Met	I Ile	L Leu	V Val	F Phe	Y Tyr	W Trp

Figura 3. Alineament de seqüències. *a*) Pàgina de l'*Atlas of protein sequence and structure*, de Margaret Dayhoff i col·laboradors, corresponent a l'hormona de creixement humana. *b*) Alineament d'una regió de les seqüències d'aminoàcids de l'enzim gliceraldehid 3-fosfat deshidrogenasa en diferents espècies. Algunes posicions d'aquest alineament (columnes) estan totalment conservades (és a dir, hi ha el mateix aminoàcid en totes les espècies), d'altres posicions són variables (estan ocupades per aminoàcids diferents en diferents espècies) i algunes posicions només són presents en algunes espècies. En general, les posicions conservades en un alineament corresponen als aminoàcids més importants per al manteniment de la funció comuna a les proteïnes alineades. D'altra banda, com més properes filogenèticament són dues espècies, més semblant és la seqüència d'aminoàcids de les seves proteïnes. Per exemple, la seqüència humana de la gliceraldehid 3-fosfat deshidrogenasa s'assembla més a la de la mosca (un altre animal) que a la d'un arqueobacteri. La similitud de seqüència constitueix, de fet, una bona indicació de proximitat filogenètica. *c*) A partir d'alineaments de proteïnes diferents, Margaret Dayhoff va construir les anomenades *matrius de substitució*. El valor dels coeficients d'aquestes matrius està relacionat amb la probabilitat d'observar dins una posició d'un alineament la substitució d'un aminoàcid determinat per un altre. El valor zero és el valor neutral; és a dir, el valor que indica que la substitució d'un aminoàcid per l'altre en els alineaments de proteïnes ocorre amb la freqüència que esperaríem a l'atzar donades les freqüències globals d'aquests aminoàcids. Els valors positius (com ara, per exemple, el valor +2 entre arginina, Arg, i histidina, His) indiquen que l'intercanvi entre aquests dos aminoàcids és observat, en els alineaments de proteïnes, amb una freqüència més elevada que l'esperada, mentre que els valors negatius (com, per exemple, el valor −7 entre glicina, Gly, i triptòfan, Trp) indiquen que l'intercanvi entre els dos aminoàcids s'observa amb menys freqüència que l'esperada.

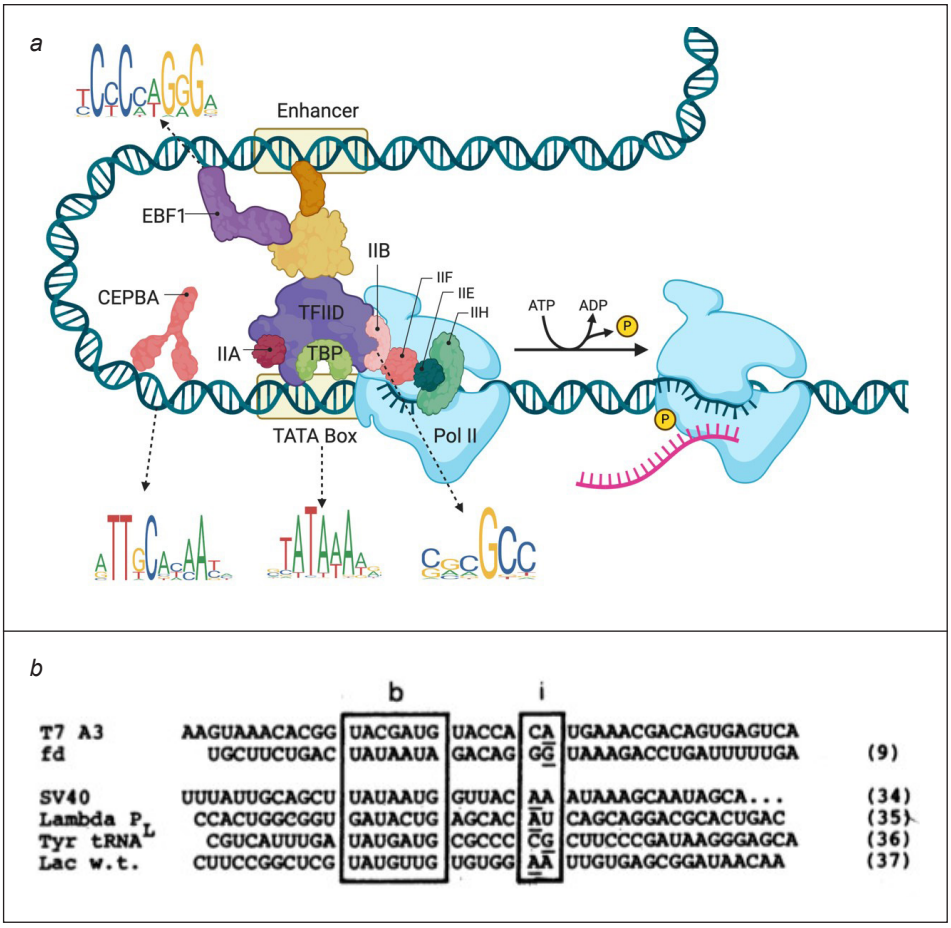
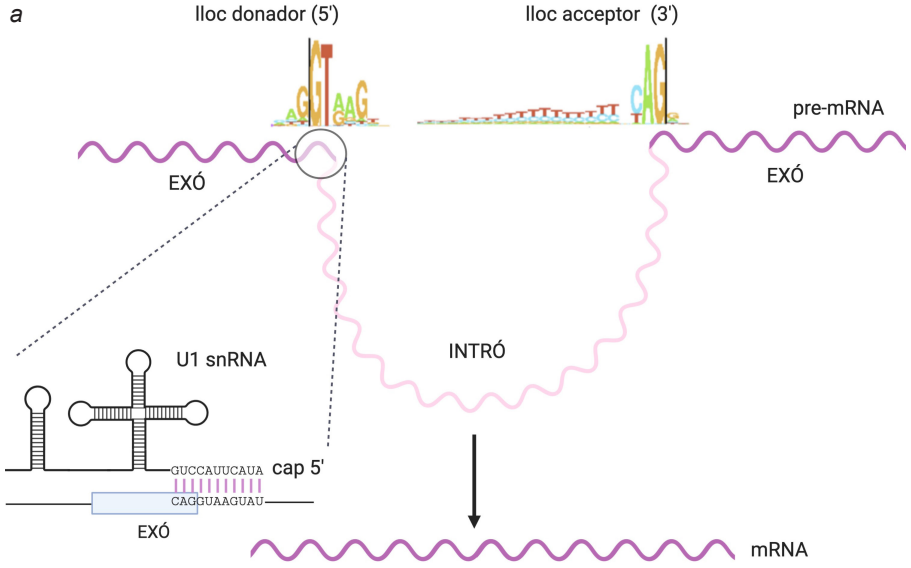

a
Enhancer
EBF1
IIB
IIF
IIE
IIH
CEPBA
TFIID
IIA
TBP
ATP
ADP
P
TATA Box
Pol II
b
b
i
T7 A3 AAGUAAACACGG UACGAUG UACCA CA UGAAACGACAGUGAGUCA
fd UGCUUCUGAC UAUAAUA GACAG GG UAAAGACCUGAUUUUUGA (9)
SV40 UUUAUUGCAGCU UAUAAUG GUUAC AA AUAAAGCAAUAGCA... (34)
Lambda P_L CCACUGGCGGU GAUACUG AGCAC AU CAGCAGGACGCACUGAC (35)
Tyr tRNA^L CGUCAUUUGA UAUGAUG CGCCC CG CUUCCCGAUAAGGGAGCA (36)
Lac w.t. CUUCCGGCUCG UAUGUUG UGUGG AA UUGUGAGCGGAUAACAA (37)

Figura 4. El codi transcripcional. *a*) Activació de la transcripció d'un gen. La descodificació de la informació en el genoma comença amb la transcripció a RNA (representat de color rosa en la figura) de les regions del DNA que corresponen als gens. És en la transcripció diferencial dels gens codificats en la seqüència del genoma, la qual és, essencialment, idèntica en totes les cèl·lules de l'organisme, on rau l'especificitat cel·lular, és a dir, el fet que les cèl·lules de l'organisme, tot i compartir el mateix DNA, exhibeixen morfologies i funcions molt diferents. L'inici de la transcripció del DNA per l'RNA polimerasa (Pol II a la figura) depèn de l'acció coordinada d'unes proteïnes que s'anomenen *factors de transcripció*. Aquests factors reconeixen determinats motius en el DNA, majoritàriament localitzats en la regió que precedeix el gen, la qual s'anomena *regió promotora*. En unir-se específicament a aquests motius, els factors de transcripció fan que s'engegui l'RNA polimerasa. Alguns d'aquests factors de transcripció són genèrics i participen en l'activació de la transcripció de molts gens. Aquest és el cas, per exemple, de la proteïna TBP (de l'anglès, *tata binding protein*). Aquesta proteïna reconeix un motiu que s'anomena *caixa TATA*. En la figura aquest motiu està representat per un logo de seqüències (*sequence logo*). Un logo de seqüències és una representació gràfica d'un alineament de seqüències en la qual l'alçada del símbol en cada posició correspon a la freqüència en la qual aquell símbol apareix en aquella posició en l'alineament. El logo per a la caixa TATA ha estat derivat a partir d'un conjunt de seqüències, les quals se sap que són reconegudes pel factor de transcripció TBP. Els nucleòtids més freqüents en aquestes seqüències són TATA(A/T)A(A/T) (en què (A/T) indica que, en aquella posició, A i T apareixen amb freqüències similars). L'alçada del símbol a cada posició indica el contingut informatiu d'aquella posició. Aquest és màxim quan en aquella posició sempre ocorre el mateix símbol (el mateix nucleòtid en el cas de seqüències de DNA) i mínim quan tots els símbols de l'alfabet (els quatres nucleòtids en el cas del DNA) apareixen amb la mateixa freqüència. A part de factors de transcripció genèrics, hi ha factors de transcripció específics per als diferents tipus cel·lulars i que són els responsables de la transcripció dels gens que defineixen aquests tipus. En el cas de l'exemple de la figura (que no correspon a cap gen real), aquests factors específics són CEPBA i EBF1, ambdós involucrats en la diferenciació de certs tipus cel·lulars en la sang. Aquests motius no sempre ocorren en la regió promotora, sinó que, de vegades, estan localitzats en regions allunyades linealment del gens, fins i tot, en cromosomes diferents. A través del plegament dels cromosomes en l'espai nuclear, aquestes regions que s'anomenen *estimuladors* entren físicament en contacte amb les regions promotores dels gens, i participen d'aquesta manera en la regulació de la seva activitat transcripcional. En el cas de l'exemple, el factor de transcripció EBF1 es troba en un estimulador. En la figura, cada factor de transcripció està associat al motiu de seqüència que reconeix. Aquests motius configuren el codi transcripcional. L'arranjament combinatorial d'aquests motius és característic de cada gen, i li confereix un patró transcripcional específic i propi. Figura adaptada a partir d'un model de BioRender (*http://biorender.com*). *b*) Alineament obtingut per Pribnow (1975) en comparar la seqüència dels promotors de sis gens al bacteri *Escherichia coli*.

a



b

Gene and organism	line#	pair#	D#	A#	Donor Sequence	Acceptor Sequence	D fit	match	AG–AG	ref. #
Soybean leghaemoglobin	1	1	1	1	ATTCGTAAGT	AAATAGGAT	6	1	>5	1
Soybean leghaemoglobin	2	2	2	2	ATTGGTAAGT	TTGTAGGTG	7	3	>5	1
Soybean leghaemoglobin	3	3	3	3	CGTGGTAAGT	TGTAGGTG	7	3	>4	1
D. discoideum M4	4	4	4	4	TTAAGTTTGTAT	TTTATTTATTTCAAGATT	5	0	112	2
D. discoideum M4	5	5	5	5	TAAAGTATGTTT	TAAATGATATATCAGGGA	7	1	92	2
French bean phaseolin	6	6	6	6	TCATGTACTGCC	ATGTTTGTCCTGTAGGAA	5	1	94	3
French bean phaseolin	7	7	7	7	CAATGTAAGAAA	GCATGATTTTTATAGAGC	7	0	73	3
French bean phaseolin	8	8	8	8	AGAGGTAAATAC	TGTTGGCGGATTTAGGGA	7	3	40	3
Yeast actin	9	9	9	9	TCTGGTATGTTC	ATATTATAATGTTTAGAGG	7	1	58	4
Sea urchin actin	10		10		ACAGGTAAGAAC		8			5
D. melanogaster actin	11			10		TTCTTTCGATTGCAGCTT			>15	6
D. melanogaster actin	12	10	11	11	AATGGTGGGTGG	TGTGTTATCCTGCAGGCT	7	2	>14	6
D. melanogaster actin	13	11	12	12	CCAGGTGCGTAG	CTGTCCTGTTCAGGTA	8	5	>12	6
D.m. heat shock 83K	14	12	13	13	CAAGGTGAGTAA	GTAAATCCATTGCAGATG	9	2	53	7
D.m. alcohol DH	15	13	14	14	GAAGGTAACTAT	TGTATTCAATCCTAGAAC	8	1	>14	8
D.m. alcohol DH	16	14	15	15	GGCGGTAAGTTG	TTATAACACCTTTAGAAA	7	1	>15	8
B. mori fibroin	17	15	16	16	GCAGGTGAGTTA	AACATTTTGTTTCAGTAT	9	3	47	9
Silkmoth chorion	18	16	17	17	TCAGGTAAGTTT	ATATCCAAAACACAGTCC	9	3	23	10
Silkmoth chorion	19	17	18	18	TCAGGTAAGGTA	GTATGTCTTTTATAGTCT	8	2	35	10
Silkmoth chorion	20	18	19	19	CCAAGTGAGTTT	GTTTTTTTTTCTCAGTCT	8	0	14	10
Silkmoth chorion	21	19	20	20	CCAGGTAAGTGA	CGCTTTTCCTTTAAGAAT	9	2	147	10
Chick type 1 α2 collagen	22			21		TGTATTTAACAGGGT			>11	11
Chick type 1 α2 collagen	23	20	21	22	AGATGTAAGTCA	TGAATTTTTTAGGGT	7	1	>11	11

Nucleic Acids Research

A. Donor Sequences

position	−4	−3	−2	−1	+1	+2	+3	+4	+5	+6	+7	+8
total	139	139	139	139	139	139	139	139	139	139	136	136
A	42	56	89	12	0	0	86	94	12	23	53	33
T	28	10	18	17	0	139	9	16	7	87	30	36
C	42	60	16	8	0	0	3	13	3	17	28	40
G	27	13	16	102	139	0	41	16	117	12	25	27

35

Figura 5. El codi de l'empalmament. *a*) L'empalmament és el procés mitjançant el qual regions internes del transcrit d'RNA, anomenades *introns*, són eliminades, i les regions que queden, els exons, són reenganxades per a formar la seqüència de l'mRNA. Els punts de tall dels introns (els llocs d'empalmament) estan definits per motius de seqüència. En la figura, aquests motius estan representats per logos (vegeu la figura 4 per a una explicació). Tot i que l'empalmament ocorre en la molècula d'RNA, sovint s'utilitza el símbol T per a timina, en lloc de U per a uracil, en els logos que representen els llocs d'empalmament i, en general, els motius de seqüència d'RNA. En el cas del lloc d'empalmament de l'extrem 5′ de l'intró, la seqüència consens és CAG|GTAAGT (en què el símbol | denota la frontera entre l'exó i l'intró). La seqüència consens simplement inclou el símbol (nucleòtid) més freqüent en cada posició. Aquesta seqüència és exactament la complementària a la seqüència d'una molècula d'RNA (un membre de la família dels snRNA) que forma part d'un complex d'RNA i proteïnes (una ribonucleoproteïna) que reconeix el lloc d'empalmament 5′ (en la cartel·la [*inset*] en la figura). Figura creada amb BioRender (*http://biorender.com*). *b*) Alineament obtingut per Mount (1982) de la seqüència genòmica al voltant de cent trenta llocs d'empalmament. Mount va derivar una matriu, els coeficients de la qual indiquen la freqüència amb la qual cada nucleòtid apareix en cada posició de l'alineament. Aquestes matrius s'anomenen *matrius de pesos posicionals* (i és a partir d'aquestes matrius que es deriven els logos de seqüència). En aquesta matriu, la posició +1 correspon a la primera posició dins l'intró, que és sempre ocupada pel nucleòtid G, de la mateixa manera que la posició +2 és sempre ocupada pel nucleòtid U. Al voltant d'aquests nucleòtids, hi ha altres posicions que també estan conservades, tot i que exhibeixen més variabilitat.

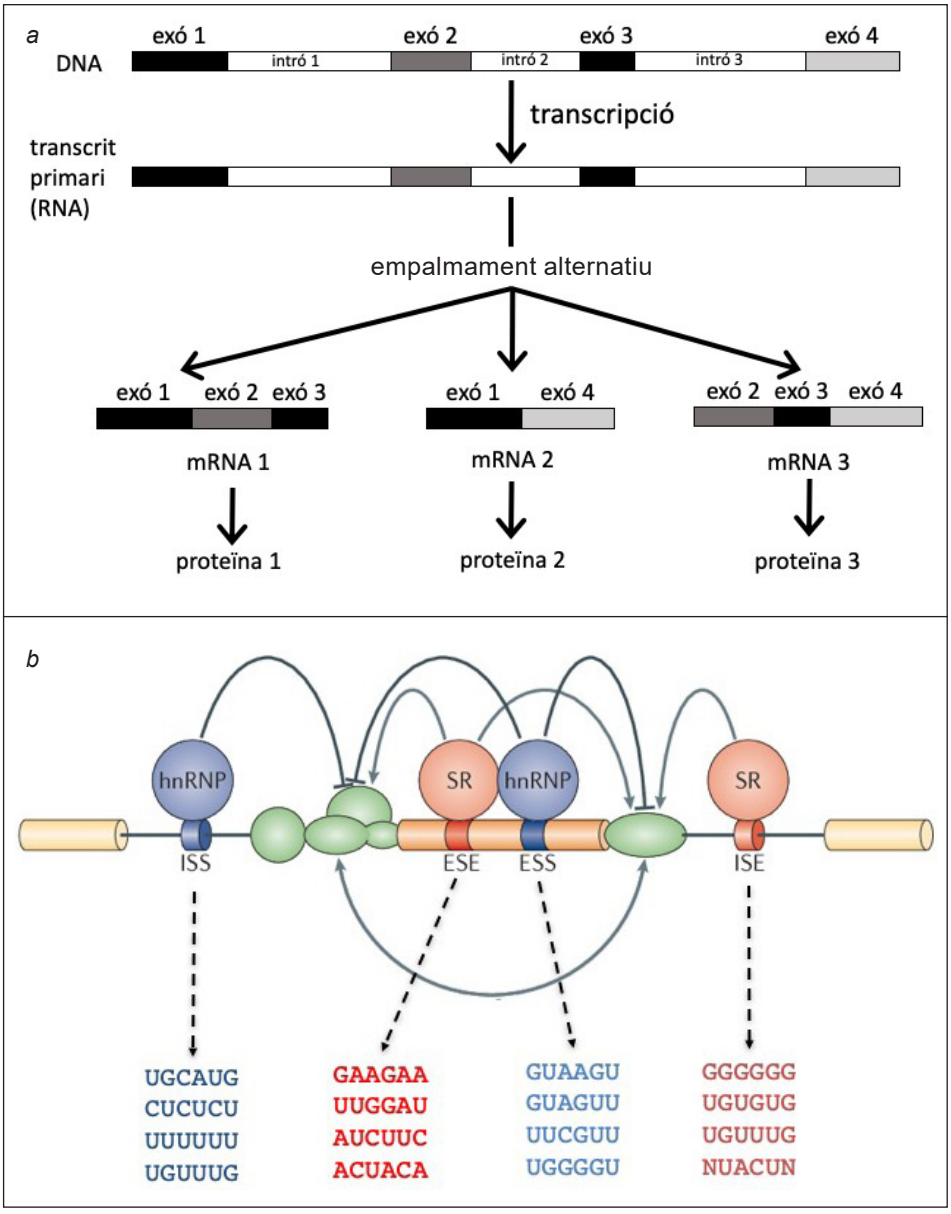
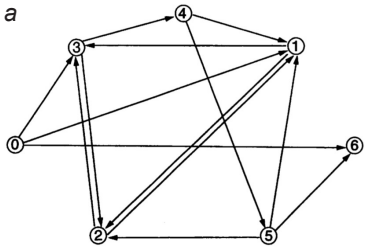

a
exó 1
exó 2
exó 3
exó 4
DNA
intró 1
intró 2
intró 3
transcripció
transcrit primari (RNA)
empalmament alternatiu
exó 1 exó 2 exó 3
exó 1 exó 4
exó 2 exó 3 exó 4
mRNA 1
mRNA 2
mRNA 3
proteïna 1
proteïna 2
proteïna 3
b
hnRNP
SR
hnRNP
SR
ISS
ESE
ESS
ISE
UGCAUG
CUCUCU
UUUUUU
UGUUUG
GAAGAA
UUGGAU
AUCUUC
ACUACA
GUAAGU
GUAGUU
UUCGUU
UGGGGU
GGGGGG
UGUGUG
UGUUUG
NUACUN

Figura 6. El codi de l'empalmament. *a*) L'empalmament alternatiu és el procés mitjançant el qual combinacions diferents d'exons del mateix gen donen lloc a transcrits diferents i, eventualment, a proteïnes distintes. En la figura s'esquematitza un gen amb quatre exons, el qual forma tres transcrits, cadascun dels quals és definit per una combinació d'exons diferents. *b*) Motius auxiliars en la seqüència dels introns i dels exons contribueixen a la selecció de llocs d'empalmament, la seqüència dels quals és aparentment subòptima. Aquests motius són reconeguts per unes proteïnes que s'uneixen a l'RNA i que s'anomenen *factors d'empalmament*. Aquests factors pertanyen majoritàriament a dues grans famílies: SR, els quals tenen sobretot un paper potenciador dels llocs d'empalmament veïns, i hnRNP, que tenen un paper debilitador. Depenent de la ubicació i de la funció, aquests motius s'anomenen *estimulador exònic d'empalmament* (*exonic splicing enhancers*, ESE), *silenciador exònic d'empalmament* (*exonic splicing silencers*, ESS), *estimulador intrònic d'empalmament* (*intronic splicing enhancers*, ISE) o *silenciador intrònic d'empalmament* (*intronic splicing silencers*, ISS). La figura mostra alguns d'aquests motius, els quals, juntament amb els llocs canònics d'empalmament, configuren una mena de codi de l'empalmament. La figura és una adaptació de la figura 1 a Kornblihtt *et al*. (2013). Les seqüències associades als factors d'empalmament provenen de Holste i Ohler (2008).

a

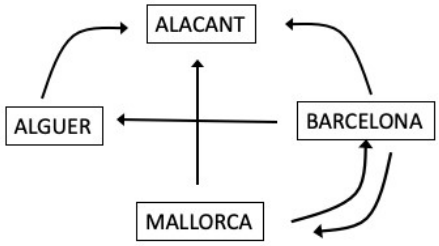


b

Pas 1: Generar (molts) camins aleatoris en el graf.
Pas 2: Mantenir només els camins que comencen pel vèrtex d'inici i acaben al vèrtex de finalització.
Pas 3: Mantenir només els camins que contenen exactament un nombre de vèrtexs exactament igual al nombre de vèrtexs del graf.
Pas 4: Mantenir només els camins que passen per tots els vèrtexs del graf almenys una vegada.

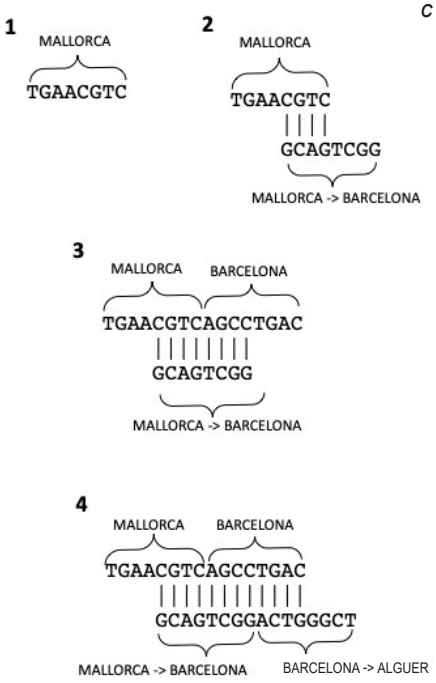
c

Pas 1



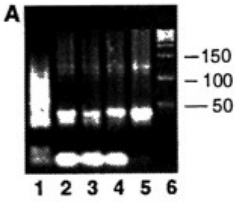
CIUTAT	CODI DNA	COMPLEMENT
MALLORCA	ACTTGCAG	TGAACGTC
BARCELONA	TCGGACTG	AGCCTGAC
ALGUER	GGCTATGT	CCGATACA
ALACANT	CCGAGCAA	GGCTCGTT

VOL	CODI DNA
MALLORCA-BARCELONA	GCAGTCGG
MALLORCA-ALACANT	GCAGCCGA
BARCELONA-ALGUER	ACTGGGCT
BARCELONA-ALACANT	ACTGCCGA
BARCELONA-MALLORCA	ACTGACTT
ALGUER-ALACANT	ATGTCCGA

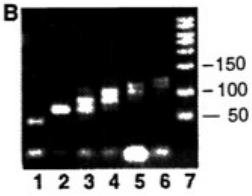


d

Pas 2: selecció per PCR amb encebadors complementaris a la ciutat inicial i a la ciutat final.



Pas 3: gel d'electroforesi per a identificar els trajectes que contenien exactament set ciutats (140 parells de bases).



Pas 4: selecció per afinitat utilitzant seqüències complementàries a cada ciutat conjugades a microesferes magnètiques.

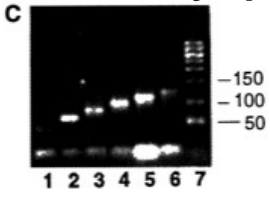


Figura 7. Els ordinadors basats en DNA. *a*) El problema del viatjant de comerç és un problema típic en computació que es pot formular de la manera següent: donada una llista de ciutats i de les connexions disponibles entre elles (per exemple, per via aèria) i una ciutat d'inici i una de finalització, hi ha algun camí que passi per totes i cadascuna de les ciutats i ho faci només una vegada? En termes més generals, el problema es pot plantejar en qualsevol graf dirigit. En el cas específic del problema del viatjant de comerç, els vèrtexs representarien ciutats i les arestes, connexions entre elles. (Que un graf és dirigit significa que les arestes tenen una direcció; no és el mateix A → B que B → A.) Donat un vèrtex d'inici i un de finalització, un camí que els connecta, passant per tots i cadascun dels vèrtexs només una vegada, es diu que és *hamiltonià*. L'any 1994, Leonard M. Adleman va implementar un algoritme per determinar si en un graf hi ha un camí hamiltonià usant molècules de DNA (Adleman, 1994). En concret, Adleman va resoldre una instància específica del problema en el graf que es representa en la figura. Aquest graf està constituït per set vèrtexs i les arestes que els connecten entre ells. En aquest graf, si comencem al vèrtex 0 i acabem al vèrtex 6, hi ha un camí que és hamiltonià: 0 → 1 → 2 → 3 → 4 → 5 → 6. *b*) Algoritme dissenyat per Adleman per a determinar si, donats un graf, i un vèrtex d'inici i un de finalització, existeix un camí hamiltonià. *c*) Adleman va implementar aquest algoritme fent servir DNA. Va codificar els vèrtexs del graf mitjançant seqüències de DNA de vint nucleòtids i les arestes que connectaven dos vèrtexs, mitjançant una seqüència de nucleòtids també de vint nucleòtids, en la qual els deu primers nucleòtids eren idèntics als deu darrers del vèrtex d'origen i els deu darrers idèntics als deu primers del vèrtex de destinació. Per simplificar, en la figura hem representat el graf com el mapa de connexions aèries un dia determinat, només entre quatre aeroports dels Països Catalans. Si el viatjant comença a Mallorca i acaba a Alacant, pot visitar les quatre ciutats només una vegada en un sol dia (és a dir, hi ha un camí hamiltonià), però, si volgués començar a Barcelona i acabar a Mallorca, no ho podria fer. També per simplificar, en la figura hem utilitzat només deu nucleòtids (en lloc dels vint que va utilitzar Adleman) per a codificar les ciutats i els trajectes entre elles. Per implementar el pas 1 de l'algoritme, Adleman va sintetitzar aproximadament 3×10^{13} còpies de cadascuna de les seqüències complementàries a les seqüències que representen les ciutats i de cadascuna de les seqüències corresponents als trajectes i les va barrejar en una reacció de lligació. Les propietats de complementarietat de DNA fan que aquesta reacció generi molècules més llargues, com a resultat de la concatenació de (seqüències corresponents a) ciutats connectades per trajectes. Així, la seqüència TGAACGTC (que representa Mallorca) s'aparella amb la seqüència GCAGTCGG (que representa el trajecte Mallorca → Barcelona), gràcies al fet que la segona meitat de la primera molècula (CGTC) és complementària a la primera part de la segona molècula (GCAG). La doble hèlix resultant serveix de motlle perquè s'incorpori la seqüència AGCCTGAC (que representa Barcelona), ja que la primera meitat d'aquesta seqüència és complementària a la segona meitat de la seqüència que representa el trajecte Mallorca → Barcelona). I així successivament. Donat el gran nombre de molècules generades inicialment, aquest procés dona lloc, amb tota certesa, a seqüències de DNA que representen tots els camins possibles dins el graf. *d*) Metodologia emprada per Adleman per a implementar els passos 2, 3 i 4 del seu algoritme. Vegeu el text per a més detalls.

Referències

Adleman, L. M. (1994). «Molecular computation of solutions to combinatorial problems». *Science*, 266 (5187), p. 1021-1024.

Barash, Y.; Calarco, J. A.; Gao, W.; Pan, Q.; Wang, X.; Shai, O.; Blencowe, B. J.; Frey, B. J. (2010). «Deciphering the splicing code». *Nature*, 465 (7294), p. 53-59.

Benenson, Y.; Paz-Elizur, T.; Adar, R.; Keinan, E.; Livneh, Z.; Shapiro, E. (2001). «Programmable and autonomous computing machine made of biomolecules». *Nature*, 414 (6862), p. 430-434.

Brenner, S.; Jacob, F.; Meselson, M. (1961). «An unstable intermediate carrying information from genes to ribosomes for protein synthesis». *Nature*, 190, p. 576-581.

Cherry, K. M.; Qian, L. (2018). «Scaling up molecular pattern recognition with DNA-based winner-take-all neural networks». *Nature*, 559 (7714), p. 370-376.

Church, G. M.; Gao, Y.; Kosuri, S. (2012). «Next-generation digital information storage in DNA». *Science*, 337 (6102), p. 1628.

Cobb, M. (2013). «1953: When genes became "information"». *Cell*, 153 (3), p. 503-506.

— (2015). «Who discovered messenger RNA?». *Current Biology*, 25 (13), p. R526-532.

Collado-Vides, J. (1989). «A transformational-grammar approach to the study of the regulation of gene expression». *Journal of Theoretical Biology*, 136 (4), p. 403-425.

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

Dayhoff, M. O.; Schwartz, R. M. (1978). «A model of evolutionary change in proteins». A: *Atlas of protein sequence and structure*. Washington, DC: National Biomedical Research Foundation, cap. 22.

Debord, G. (1964). *La societat de l'espectacle*. Apple Books, p. 9.

Dounce, A. L. (1953). «Nucleic acid template hypotheses». *Nature*, 172 (4377), p. 541.

Emmeche, C. (1994). «The computational notion of life». *Theoria*, Segona època, 9 (21), p. 1-30.

Erlich, Y.; Zielinski, D. (2017). «DNA fountain enables a robust and efficient storage architecture». *Science*, 355 (6328), p. 950-954.

Fairbrother, W. G.; Yeh, R. F.; Sharp, P. A.; Burge, C. B. (2002). «Predictive identification of exonic splicing enhancers in human genes». *Science*, 297 (5583), p. 1007-1013.

Goldman, N.; Bertone, P.; Chen, S.; Dessimoz, C.; LeProust, E. M.; Sipos, B.; Birney, E. (2013). «Towards practical, high-capacity, low-maintenance information storage in synthesized DNA». *Nature*, 494 (7435), p. 77-80.

Gros, F.; Hiatt, H.; Gilbert, W.; Kurland, C. G.; Risebrough, R. W.; Watson, J. D. (1961). «Unstable ribonucleic acid revealed by pulse labelling of *Escherichia coli*». *Nature*, 190, p. 581-585.

Guigó, R. (2007). «Bioinformàtica». *Treballs de la Societat Catalana de Biologia*, 58.

Holste, D.; Ohler, U. (2008). «Strategies for identifying RNA splicing regulatory motifs and predicting alternative splicing events». *PLoS Computational Biology*, 4 (1), e21. DOI: 10.1371/journal.pcbi.0040021.

Jacob, F. (1973). *The logic of life*: *A history of heredity*. Traducció de B. E. Spillmann. Nova York: Pantheon Books.

Jacob, F.; Monod, J. (1961). «Genetic regulatory mechanisms in the synthesis of proteins». *Journal of Molecular Biology*, 3, p. 318-356.
Kornblihtt, A.; Schor, I. E.; Alló, M.; Dujardin, G.; Petrillo, E.; Muñoz, M. J. (2013). «Alternative splicing: A pivotal step between eukaryotic transcription and translation». *Nature Reviews Molecular Cell Biology*, 14, p. 153-165. DOI: 10.1038/nrm3525.
Lifton, R. P.; Goldberg, M. L.; Karp, R. W.; Hogness, D. S. (1978). «The organization of the histone genes in *Drosophila melanogaster*: Functional and evolutionary implications». *Cold Spring Harbor Symposia on Quantitative Biology*, 42, Pt 2, p. 1047-1051.
Margalef, R. (1957). «La teoría de la información en ecología». *Memorias de la Real Academia de Ciencias y Artes de Barcelona*, 32 (13), p. 373-449.
Maxam, A. M.; Gilbert, W. (1977). «A new method for sequencing DNA». *Proceedings of the National Academy of Sciences of the United States of America*, 74 (2), p. 560-564.
Mayr, E. (1961). «Cause and effect in biology». *Science*, 134 (3489), p. 1501-1506.
Mount, S. M. (1982). «A catalogue of splice junction sequences» [en línia]. *Nucleic Acids Research*, 10 (2), p. 459-472. <https://doi.org/10.1093/nar/10.2.459>.
Needleman, S. B.; Wunsch, C. D. (1970). «A general method applicable to the search for similarities in the amino acid sequence of two proteins». *Journal of Molecular Biology*, 48 (3), p. 443-453.
Organick, L.; Ang, S. D.; Chen, Y. J.; Lopez, R.; Yekhanin, S.; Makarychev, K.; Racz, M. Z.; Kamath, G.; Gopalan, P.; Nguyen, B.; Takahashi, C. N.; Newman, S.; Parker, H. Y.; Rashtchian, C.; Stewart, K.; Gupta, G.; Carlson, R.; Mulligan, J.; Carmean, D.; Seelig, G.; Strauss, K. (2018). «Random access in large-scale DNA data storage». *Nature Biotechnology*, 36 (3), p. 242-248.
Peluffo, A. E. (2015). «The "genetic program": Behind the genesis of an influential metaphor». *Genetics*, 200 (3), p. 685-696.
Pribnow, D. (1975). «Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter». *Proceedings of the National Academy of Sciences of the United States of America*, 72 (3), p. 784-788.
Regot, S.; Macia, J.; Conde, N.; Furukawa, K.; Kjellén, J.; Peeters, T.; Hohmann, S.; Nadal, E. de; Posas, F.; Solé, R. (2011). «Distributed biological computation with multicellular engineered networks». *Nature*, 469 (7329), p. 207-211. DOI: 10.1038/nature09679.
Sanger, F.; Coulson, A. R. (1975). «A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase». *Journal of Molecular Biology*, 94 (3), p. 441-448.
Sanger, F.; Thompson, E. O. P. (1953). «The amino-acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from partial hydrolysates». *The Biochemical Journal*, 53 (3), p. 353-366.
Schaller, H.; Gray, C.; Herrmann, K. (1975). «Nucleotide sequence of an RNA polymerase binding site from the DNA of bacteriophage fd». *Proceedings of the National Academy of Sciences of the United States of America*, 72 (2), p. 737-741.
Schrodinger, E.; Penrose, R. (2012 [1944]). *What is life?: With mind and matter and autobiographical sketches*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9781107295629.

Shannon, C.; Weaber, W. (1949). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press. ISBN 0-252-72548-4.
Smith, T. F.; Waterman, M. S. (1981). «Identification of common molecular subsequences». *Journal of Molecular Biology*, 147 (1), p. 195-197.
Szymanski, M.; Barciszewski, J. (2017). «The path to the genetic code». *Biochimica et Biophysica Acta*: *General Subjects*, 1861 (11 Pt A), p. 2674-2679.
Watson, J. D.; Crick, F. H. (1953*a*). «Genetical implications of the structure of deoxyribonucleic acid». *Nature*, 171 (4361), p. 964-967.
— (1953*b*). «Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid». *Nature*, 171 (4356), p. 737-738.
Wiener, N. (1948). *Cybernetics*: *Or control and communication in the animal and the machine.* París: Hermann & Cie; Cambridge (Mass.): MIT Press. ISBN 978-0-262-73009-9.
Wingender, E. (1988). «Compilation of transcription regulating proteins». *Nucleic Acids Research*, 16 (5), p. 1879-1902.
Zuckerkandl, E.; Pauling, L. (1965). «Molecules as documents of evolutionary history». *Journal of Theoretical Biology*, 8 (2), p. 357-366.


INSTITVT D'ESTVDIS CATALANS
MCMVII


9 788499 656373


INSTITVT D'ESTVDIS CATALANS
MCMVII


9 788499 656373