Informatique et biologie : un mariage de raison

mais heureux (?)

Laurent Bloch

Institut Français
d'Analyse Stratégique

22 octobre 2013

Première révolution

Biologie moléculaire : protéines et code génétique

- Tout ce qui est vie en nous est le fait de protéines ;
- une protéine est constituée d'acides aminés ;
- le paradigme de la biologie moléculaire postule que l'information génétique est formulée par un texte, le génome, écrit dans un alphabet de quatre lettres, A, T, G, C, et que la connaissance de ce texte permet de connaître les fonctions de l'organisme considéré, sans avoir à entrer dans des considérations supplémentaires d'ordre physico-chimique.

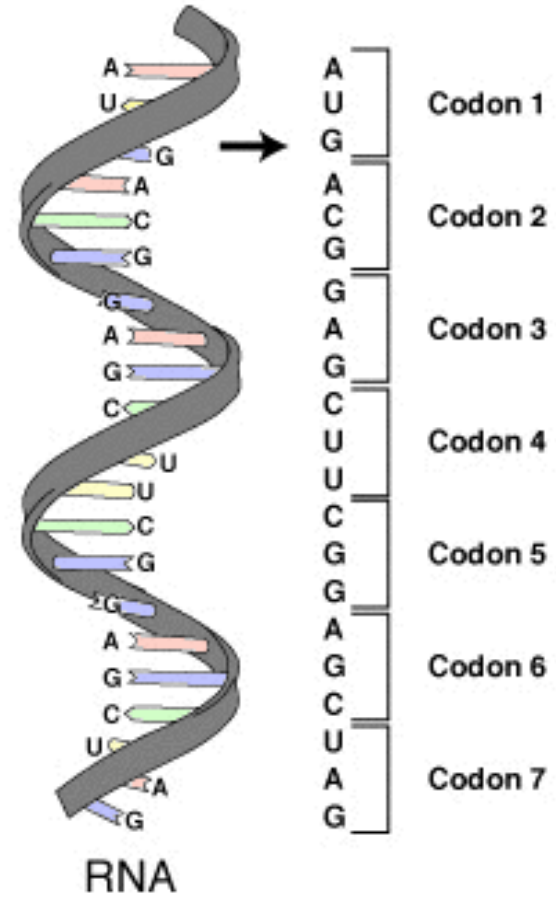
Première révolution

Biologie moléculaire : protéines et code génétique

- Le code génétique est l'ensemble des règles permettant de traduire les informations contenues dans le matériel génétique des cellules vivantes pour produire des protéines ;
- ce code établit une correspondance entre un triplet de nucléotides, appelé codon sur l'ARN messager et un acide aminé qui sera incorporé dans la protéine en cours de synthèse ;
- l'ARN messager est lui-même produit par transcription à partir de l'ADN ;
- cette correspondance codon-acide aminé permet de résumer le code génétique sous forme d'une table associant chacun des 64 codons ou triplets possibles (4^3) avec l'un des 20 acides aminés présents dans les protéines ;
- dans la cellule, ce sont les ribosomes qui effectuent l'interprétation du code génétique, un processus appelé la traduction génétique.

(Wikipedia dixit)

De la biologie moléculaire ...



ADN, adénine, thymine, cytosine, guanine :
ATGC
ARN, Uracile : U
AUGC

Au cœur de la vie, la cellule
de Dominique Morello

De la biologie moléculaire ...

- Le génome humain : 3,4 milliards de nucléotides et quelques 25 000 gènes ;
- séquençage : Frederick Sanger (deux prix Nobel de chimie) ;
- l'ARN polymérase « lit » le texte d'ADN et le *transcrit* en ARNm ;
- transcriptome ;
- l'ARN de transfert amène des acides aminés vers le ribosome ;
- la *traduction* est entamée par le ribosome ;
- PCR : Kary Mullis.

Acide aminé	**Codons**
Alanine	GCU, GCC, GCA, GCG.
Arginine	CGU, CGC, CGA, CGG ; AGA, AGG.
Asparagine	AAU, AAC.
Acide aspartique	GAU, GAC.
Cystéine	UGU, UGC.
Glutamine	CAA, CAG.
Acide glutamique	GAA, GAG.
Glycine	GGU, GGC, GGA, GGG.
Histidine	CAU, CAC.
Isoleucine	AUU, AUC, AUA.
Leucine	UUA, UUG ; CUU, CUC, CUA, CUG.
Lysine	AAA, AAG.
Méthionine	AUG.
Phénylalanine	UUU, UUC.

Acide aminé	**Codons**
Proline	CCU, CCC, CCA, CCG.
Pyrrolysine	UAG, après séquence PylIS.
Sélénocystéine	UGA, après séquence SecIS.
Sérine	UCU, UCC, UCA, UCG ; AGU, AGC.
Thréonine	ACU, ACC, ACA, ACG.
Tryptophane	UGG. (UGA)
Tyrosine	UAU, UAC.
Valine	GUU, GUC, GUA, GUG.
START	AUG. (UUG, GUG)
STOP Ambre	UAG.
STOP Ocre	UAA.
STOP Opale	UGA.

Une protéine :

```
ID   1433B_XENTR             Reviewed;         244 AA.
AC   Q5XGC8; Q28HK2;
DT   22-NOV-2005, integrated into UniProtKB/Swiss-Prot.
DT   23-NOV-2004, sequence version 1.
DT   28-NOV-2006, entry version 18.
DE   14-3-3 protein beta/alpha.
GN   Name=ywhab;
OS   Xenopus tropicalis (Western clawed frog) (Silurana tropicalis).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Amphibia; Batrachia; Anura; Mesobatrachia; Pipoidea; Pipidae;
OC   Xenopodinae; Xenopus; Silurana.
OX   NCBI_TaxID=8364;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA].
RG   Sanger Xenopus tropicalis EST/cDNA project;
RL   Submitted (MAR-2006) to the EMBL/GenBank/DDBJ databases.
RN   [2]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA].
RC   TISSUE=Embryo;
```

Une protéine :

```
RG   NIH - Xenopus Gene Collection (XGC) project;
RL   Submitted (OCT-2004) to the EMBL/GenBank/DDBJ databases.
CC   -!- FUNCTION: Adapter protein implicated in the regulation of a large
CC       spectrum of both general and specialized signaling pathway. Binds
CC       to a large number of partners, usually by recognition of a
CC       phosphoserine or phosphothreonine motif. Binding generally results
CC       in the modulation of the activity of the binding partner (By
CC       similarity).
CC   -!- SUBUNIT: Homodimer (By similarity).
CC   -!- SUBCELLULAR LOCATION: Cytoplasm (By similarity).
CC   -!- SIMILARITY: Belongs to the 14-3-3 family.
CC   ----------------------------------------------------------------------
CC   Copyrighted by the UniProt Consortium, see http://www.uniprot.org/term
CC   Distributed under the Creative Commons Attribution-NoDerivs License
CC   ----------------------------------------------------------------------
```

Une protéine :

```
DR   EMBL; CR760847; CAJ82973.1; -; mRNA.
DR   EMBL; BC084514; AAH84514.1; -; mRNA.
DR   UniGene; Str.8742; -.
DR   SMR; Q5XGC8; 1-231.
DR   Ensembl; ENSXETG00000022830; Xenopus tropicalis.
DR   InterPro; IPR000308; 14-3-3.
DR   Gene3D; G3DSA:1.20.190.20; 14-3-3; 1.
DR   PANTHER; PTHR18860; 14-3-3; 1.
DR   Pfam; PF00244; 14-3-3; 1.
DR   PRINTS; PR00305; 1433ZETA.
DR   ProDom; PD000600; 14-3-3; 1.
DR   SMART; SM00101; 14_3_3; 1.
DR   PROSITE; PS00796; 1433_1; 1.
DR   PROSITE; PS00797; 1433_2; 1.
KW   Acetylation.
FT   CHAIN         1    244       14-3-3 protein beta/alpha.
FT                                /FTId=PRO_0000058600.
FT   MOD_RES       1      1       N-acetylmethionine (By similarity).
```

Une protéine :

```
SQ   SEQUENCE   244 AA;  27721 MW;  FF766793EA1CA9E5 CRC64;
     MDKSELVQKA KLSEQAERYD DMAASMKAVT ELGAELSNEE RNLLSVAYKN VVGARRSSWR
     VISSIEQKTE GNDKRQQMAR EYREKVETEL QDICKDVLGL LDKYLVPNAT PPESKVFYLK
     MKGDYYRYLS EVASGDSKQE TVTCSQQAYQ EAFEISKSEM QPTHPIRLGL ALNFSVFYYE
     ILNSPEKACS LAKSAFDEAI AELDTLNEES YKDSTLIMQL LRDNLTLWTS ENQGEEADNA
     EADN
//
```

Protéine issue de SwissProt, banque créée par Amos Bairoch à l'Université de Genève.

De la biologie moléculaire ...

- Des investigations qui demandaient des mois de travail répétitif et entaché d'erreurs à la paillasse sont désormais résolues en quelques heures par des méthodes informatiques ;
- la consultation des banques de données qui archivent les résultats exhaustifs du séquençage et des calculs de structure des protéines donne en quelques minutes la réponse à des questions dont la solution directe aurait constitué un thème de recherche à part entière ;
- ce qui signifie que l'on peut désormais se poser des questions inenvisageables auparavant.

Que fait-on des séquences qui sont dans ces banques ?

« Imaginez que l'on place devant vous une image – par exemple, la photographie d'une scène de la vie quotidienne un jour de printemps autour du bassin du Jardin du Luxembourg – et que l'on vous demande de dire si cette image contient un petit chien, et si oui, où se situe l'animal. Supposez que l'on vous fournisse un dessin du chien afin de vous guider, et que ce dessin soit identique à celui du petit chien illustré dans l'image (s'il s'y trouve). Dans ce cas, il suffit de glisser un décalque de ce dessin sur toute l'étendue de l'image pour être à même de répondre à la question que l'on vous a posée. Si, par contre, le chien du dessin que l'on vous a donné apparaît de façon différente de celui que vous essayez de localiser (par exemple, debout de face, alors que dans l'image le chien est couché), il vous faudra repérer d'abord quelles sont les caractéristiques essentielles de l'animal (sa forme, son apparence générale) afin que le dessin puisse vous être utile dans votre examen de l'image. »

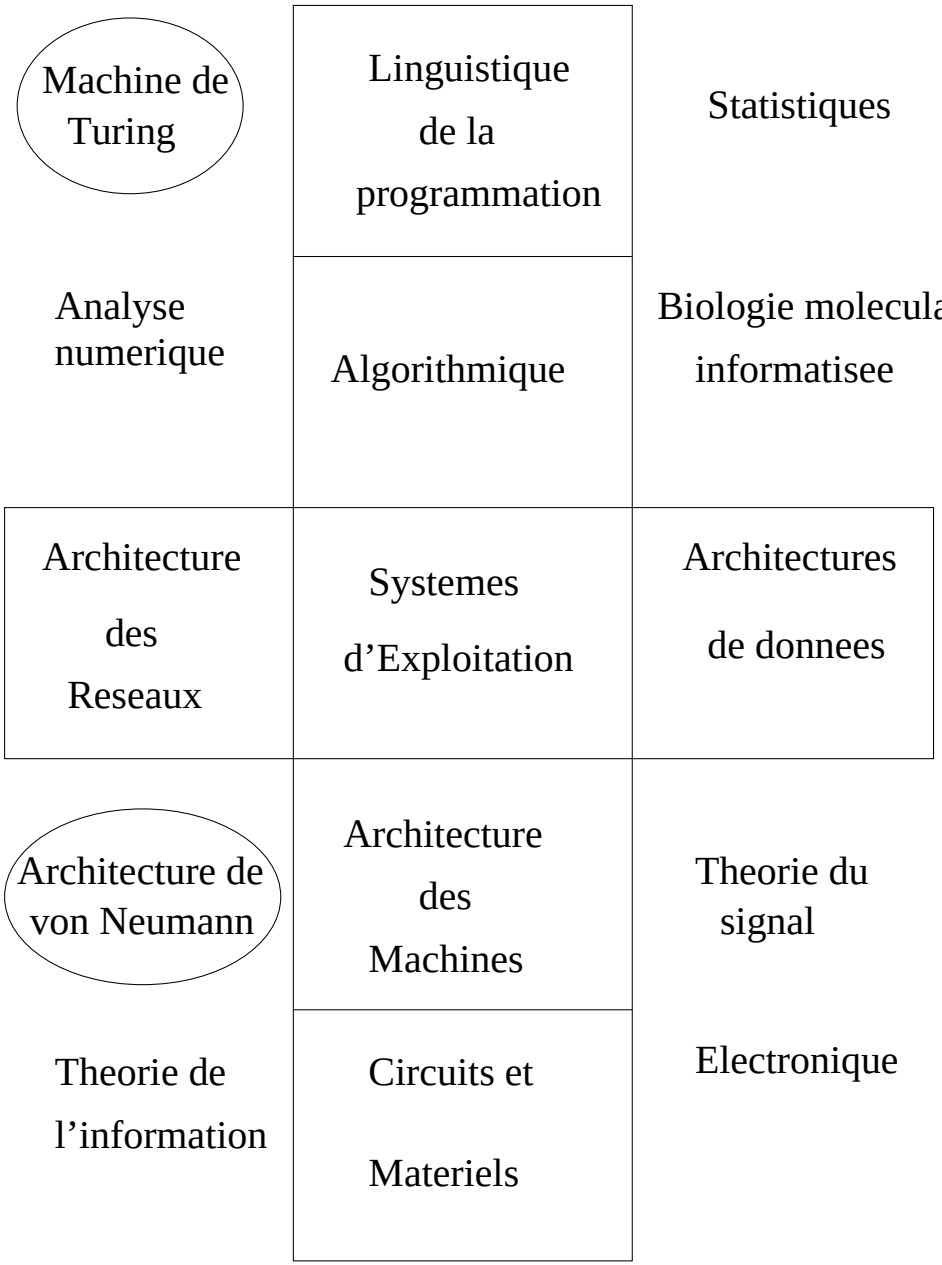
Que fait-on des séquences qui sont dans ces banques ?

« Enfin, si rien ne vous est fourni comme aide dans votre recherche, il vous faudra recréer un dessin mental de ce qui fait d'un chien un chien (et non un chat ou une souris) et c'est ce 'dessin' qui doit alors vous guider. Notez que j'ai dit recréer car si vous n'avez jamais vu un chien dans votre vie, ni lu une description de l'animal, vous serez impuissant à répondre à la question, même si l'image contient vraiment un chien. Dans tous les cas, il vous faut donc avoir une idée au moins *a priori*, concrète ou abstraite, de ce que vous cherchez. »
(Marie-France Sagot, thèse)

De la biologie moléculaire ...

- Il y a vingt ans nous étions à un tournant ;
- les gens bien informés de la recherche mondiale savaient déjà que les choses allaient dans cette direction, les mandarins résistaient pied à pied, accrochés à leurs paillasses ;
- les institutions de pointe avaient déjà lourdement investi en informatique, telles le *National Center for Biotechnology Information* (NCBI) créé en 1988 à Bethesda près de Washington, ou le Wellcome Trust Sanger Institute créé en 1992 à Hinxton près de Cambridge en Angleterre ;
- l'Institut Pasteur était en retard, et en prenait conscience.

Qu'est-ce que l'informatique ?



Il fallait expliquer que l'informatique, c'était ce qui est écrit dans les carrés et les ovales, et pas ce qui est à l'extérieur.

Dans les ovales :
les paradigmes.

De l'informatique pour la biologie moléculaire ...

Avant de pouvoir faire de l'analyse de séquences, voire de la modélisation moléculaire (à l'époque un sport de riches), il fallait déjà avoir de l'informatique : une infrastructure de calcul, un réseau, des logiciels, des ingénieurs pour faire marcher tout cela.

Mais surtout :

L'informatique n'est pas un outil.

Elle change la façon de penser.

La seule comparaison sérieuse entre informatique et automobile

« Conduire n'est pas "faire de la mécanique" pas plus que mettre en gras un titre n'est "faire de l'informatique". En revanche, déceler un bruit bizarre en roulant et l'attribuer aux pneus ou au moteur, à la direction ou au freinage, aide le diagnostic du mécanicien. Déceler si un dysfonctionnement provient de l'affichage ou du réseau, d'un disque ou de l'ordonnancement facilite, de même, le diagnostic. Encore faut-il, tout comme en mécanique, connaître les grandes fonctions et leurs relations, savoir ouvrir le capot et nommer les éléments découverts. »
(Christian Queinnec)

Enseigner l'informatique à des biologistes !

Les étudiants en biologie n'apprennent généralement pas la programmation à l'Université, ce qui en 1993 a suscité chez William Saurin, Frédéric Chauveau et Laurent Bloch l'idée d'organiser à leur intention un cours d'informatique destiné à leur procurer les bases de cette science.

Cette idée a aussitôt été soutenue par François Rougeon et Jean-Paul Aubert, alors chef du Département des Enseignements.

Enseigner l'informatique à des biologistes !

Les étudiants

Nos premiers étudiants venaient souvent des marges de la biologie, parfois après avoir découvert que les manipulations à la paillasse, passionnantes les premiers mois, devenaient vite répétitives et fastidieuses. Mais très vite la crise de l'emploi en biologie renouvela notre public. En effet, les effectifs des cursus de biologie sont très excessifs en regard des possibilités d'emploi, et nous avons vu arriver des étudiants munis d'une thèse, voire d'un postdoc dans une université étrangère prestigieuse, et qui n'avaient pas de travail.

Enseigner l'informatique à des biologistes !

Algorithmes, programmation, théorie des langages	60 h
Algorithmes et programmation T.P. (avec Scheme)	42 h
Programmation impérative et par objets (avec Java, cours)	18 h
TP de Java	21 h
Systèmes de calcul : évolution et réalisations	12 h
Introduction à Unix et X T.P.	17 h
Perspectives de l'informatique (conférences)	10 h
Algorithmes pour la Biologie	36 h
Exemples (Analyse de séquences) T.P.	36 h
Modélisation moléculaire	15 h
Installation de logiciels (Blast)	10 h
Bases de données	12 h
Architecture des machines et système d'exploitation	15 h
Logique	10 h
Réseaux	12 h
Total	326 h

Enseigner l'informatique à des biologistes !

Enseignants (cours 1998)	heures de cours
William Saurin	12 h
Laurent Bloch	40 h
Manuel Serrano	12 h
Frédérique Galisson	45 h
Christian Queinnec	6 h
Daniel Azuelos	40 h
Éric Gressier	12 h
Marie-France Sagot	15 h
Thierry Rose	16 h
Louis Jones	12 h
Bernard Caudron	36 h
Christophe Wolfhugel	40 h
Stéphane Bortzmeyer	50 h

Enseigner l'informatique à des biologistes !

Étudiants :

Année	nombre d'élèves	dont pasteuriens	dont étrangers
1994	10	6	1
1995	9	3	3
1996	7	3	2
1997	15	6	7
1998	13	4	5

Biologie et informatique : sociologie

Sans doute l'informatique est-elle au contraire perçue comme trop simple : ne suffit-il pas après tout de cliquer ici ou là avec une souris ? Alors, est-il besoin d'apprendre l'informatique ? Il n'est que de voir combien il est difficile de faire entrer l'informatique dans l'enseignement secondaire. La vraie informatique, pas les TIC, le numérique ou autres niaiseries.
Quant aux « élites », leur hostilité à l'informatique procède essentiellement de la menace qu'elles y perçoivent (avec raison) contre leurs positions.