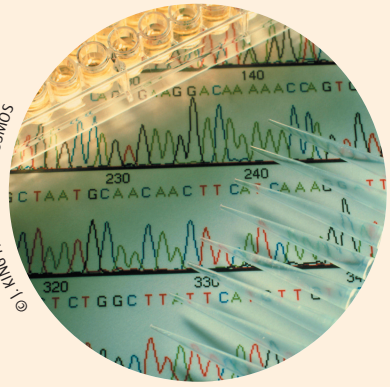


Décrypter toujours plus de génomes est une prouesse technologique. Encore faut-il parvenir à identifier les gènes qu'ils contiennent. Seuls des algorithmes efficaces peuvent éclairer ces longues listes de lettres.

Rechercher un gène dans une botte de lettres

François Rechenmann,
directeur de recherche à l'Inria Rhône-Alpes.
Francois.Rechenmann @inria.fr

Homme, poulet, chien, bœuf, rat, etc. : au total, le patrimoine génétique de plus de 250 organismes a été aujourd'hui décodé. Dans ce lot, il y a notamment plus de 200 bactéries [1]. A-t-on percé pour autant les secrets du fonctionnement des cellules ? Loin de là ! Disposer des séquences génétiques est une chose, mais trouver les fragments d'intérêt, les fameux gènes, en est une autre.

La liste brute des lettres A, C, G ou T (les initiales des quatre types de nucléotides qui composent l'ADN) n'est pas simple à lire : comment distinguer les gènes parmi les trois milliards de lettres que compte le génome humain ? En d'autres termes : comment identifier les fragments de séquences qui contiennent l'information nécessaire à la synthèse des protéines* et, ensuite, comment déterminer les fonctions de chaque protéine ?

Seuls des programmes informatiques sont à même de parcourir cette longue chaîne de caractères pour y chercher les indices suggérant la présence d'un gène. Seuls des programmes peuvent aussi recouper ces indices afin de déterminer précisément la structure de ce que l'on suppose être un gène, ainsi que de prédire la ou les protéines qu'il code.

« Start » et « stop »

Comment part-on à la recherche d'un gène ? Tout d'abord, on sait qu'il s'agit d'une succession de groupes de trois nucléotides, appelés « codons ». Chaque codon dicte la présence d'un acide aminé dans la protéine. Il existe une correspondance entre les 64 (4 × 4 × 4) codons possibles et les 20 acides aminés (c'est le code génétique). Pour identifier un gène, il faut savoir où il commence et où il termine. Or, on sait que tous les gènes commencent par un codon « start » (le triplet ATG) et se terminent par un codon « stop » (TAA, TAG ou TGA).

Suffit-il de rechercher ces différents triplets dans la séquence pour délimiter les gènes ? La solution n'est malheureusement pas aussi simple. En effet, tous ces triplets peuvent aussi se trouver à l'extérieur des gènes. De plus, le triplet « start » ATG peut, à l'inverse, apparaître au sein même d'un gène et coder alors un acide aminé.

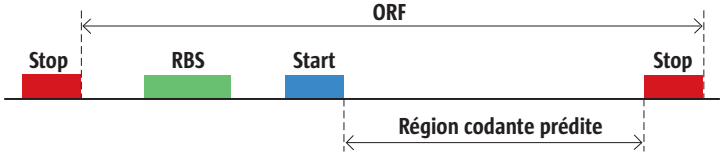
Affiner les stratégies

La stratégie des bio-informaticiens consiste tout d'abord à rechercher ce qu'ils appellent des ORF, pour « Open Reading Frames », que l'on pourrait traduire par « phases ouvertes de lecture » : il s'agit de sous-séquences encadrées par deux triplets « stop », mais qui n'en contiennent pas. L'existence d'une ORF est une condition nécessaire mais pas suffisante à la présence d'un gène. Aussi, on impose également une longueur minimale pour ces ORF : par exemple 300 lettres, soit 100 triplets. En réalité, le gène est toujours plus court puisque les algorithmes recherchent le triplet ATG le plus proche, en aval, du premier triplet « stop » et le retiennent comme le codon « start » débutant le gène hypothétique qui se termine sur le second triplet stop [fig. 1].

*** Les protéines,** formées d'acides aminés, sont, avec les glucides, les lipides et les acides nucléiques, l'un des quatre matériaux de base de tout organisme vivant.

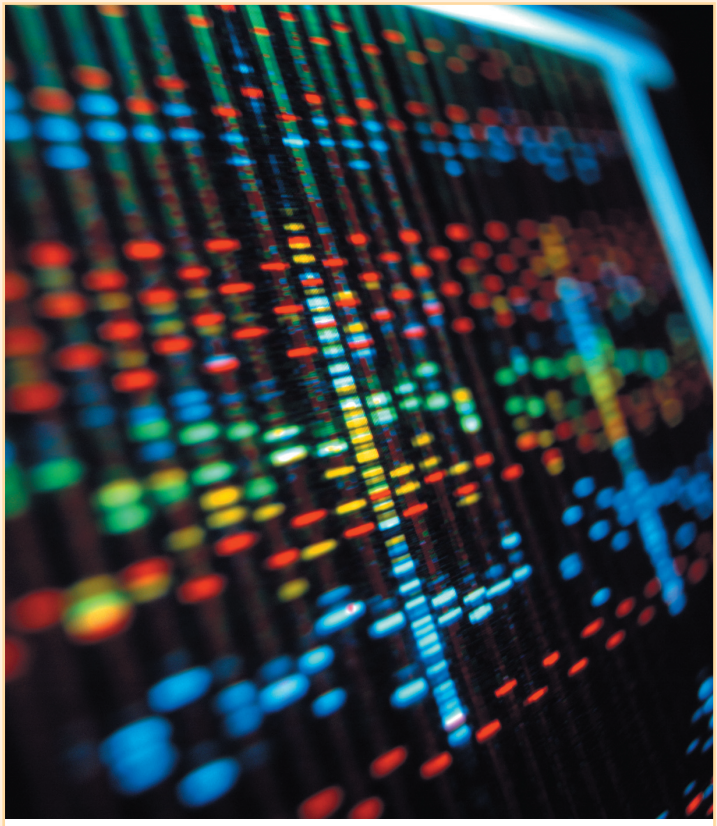
[1] http://cgg.ebi.ac.uk/services/cogent/

Fig.1 Les « phases ouvertes de lecture »



TOUT GÈNE SE SITUE OBLIGATOIREMENT DANS UNE « ORF », c'est-à-dire entre deux codons « stop ». Pour identifier un gène, on retient tout d'abord la région la plus large possible, et donc le premier codon « start ». La présence de motifs particuliers tels que RBS conforte la prédiction du gène.

Appliquée au texte du génome de la bactérie Bactilus subtilis, dont la séquence compte 4,2 millions de nucléotides, cette stratégie permet de prédire correctement l'existence de 3500 des quelques 4100 gènes connus de ce génome très étudié [2], mais en prédit plus de 1200 qui n'en sont pas. Est-il possible d'affiner cette stratégie et de réduire le nombre de « faux positifs » ? Il faut pour cela accumuler d'autres indices. Par exemple, en cherchant, en amont du codon « start », une configuration particulière de lettres qui correspond au site de fixation de la molécule d'ARN messager sur le ribosome*. La présence d'un tel « RBS » (pour « Ribosome Binding Site ») conforte la prédiction du gène. D'autres configurations de lettres, ou motifs, correspondant à la présence de sites d'interaction de l'ADN avec des molécules diverses peuvent aussi être recherchées afin de confirmer les prédictions, de les réfuter ou de les amender.



LE SÉQUENÇAGE D'UN ORGANISME PERMET DE DÉCHIFFRER LA SUCCESSION DES NUCLÉOTIDES QUI COMPOSENT L'ADN. **Sur cet écran d'ordinateur, chaque bande colorée représente une base. Cependant, cette indication ne dit rien du nombre de gènes, ni de leurs emplacements.** © TEK IMAGES/SPL/COSMOS

Six séquences

Bien que les triplets « start » et « stop » soient des motifs simples et courts, la taille des séquences et le grand nombre d'occurrences obligent à concevoir des algorithmes de recherche qui évitent les comparaisons inutiles de lettres. Et ce d'autant plus qu'il existe trois manières de grouper les lettres d'une séquence trois par trois, selon que cette opération débute à la première lettre de la séquence, à la deuxième ou à la troisième. Comme, de plus, un gène peut tout aussi bien être porté par un brin de l'ADN que par son complémentaire dans la double hélice, c'est finalement dans six séquences différentes que la recherche des gènes doit s'effectuer.

La recherche de RBS fournit un exemple d'un autre type de problèmes algorithmiques. Pour un organisme donné, il n'existe pas un motif unique qui puisse être associé à tous les sites de fixation du ribosome. Les bio-informaticiens sont donc conduits à développer des méthodes de recherche de « motifs flous », tout en minimisant le temps d'exécution.

Conforter la prédiction

Comment s'assurer qu'une séquence correspond bien à un gène ? Afin de conforter la prédiction, on parcourt des bases de séquences (telles qu'EMBL [3]) pour y rechercher des séquences similaires. Il est aussi possible de traduire la séquence du gène hypothétique en une séquence protéique, puis de regarder s'il existe des séquences protéiques similaires – par exemple dans la base Swiss-Prot [4]. Si une information sur la fonction de la protéine correspondant à ces séquences est disponible, il est alors tentant de l'attribuer à la protéine prédite. Par abus de langage, on dira que l'on a prédit la fonction du gène.

La stratégie bio-informatique de prédiction de gènes esquissée ici est rudimentaire, mais elle fournit déjà des résultats acceptables sur des génomes bactériens. D'autres algorithmes, plus complexes, possèdent de bien meilleures capacités de prédiction. C'est le cas de ceux qui, grâce à des concepts statistiques tels que les modèles de Markov, sont capables de reconnaître les agencements de nucléotides caractéristiques d'une région codante.

Le cas des eucaryotes

Ces algorithmes sont déployés pour l'analyse des génomes eucaryotes, dont celui de l'homme. En effet, outre leur taille plus importante de plusieurs ordres de grandeur, ces génomes présentent des caractéristiques qui rendent beaucoup plus difficile la prédiction des gènes. D'une part, ceux-ci y sont beaucoup plus espacés que dans un génome bactérien (il est courant que deux gènes soient séparés par plusieurs milliers de nucléotides) ; d'autre part, les gènes possèdent une structure morcelée qui, entre un codon « start » et un codon « stop », alterne régions codantes, appelées « exons », et non codantes, appelées « introns ». De ce fait, la prédiction d'un gène ne se limite plus à la recherche des bons triplets « start » et « stop », puisqu'il faut également déterminer les frontières entre exons et introns.

Une méthode consiste alors à combiner les résultats fournis par des modèles de Markov, qui estiment la probabilité pour une région de la séquence d'être codante, et la recherche des motifs flous connus pour correspondre aux frontières intron-exon. Quelles que soient la nature du génome et l'efficacité des algorithmes de recherche de gènes, leurs résultats restent toutefois des prédictions, qui ne peuvent être validées qu'à travers des démarches expérimentales. ■■

*** Un ribosome** est un assemblage moléculaire responsable de la synthèse des protéines à partir de l'information portée le long d'un gène.

[2] http://genolist.pasteur.fr/SubtiList/

[3] www.ebi.ac.uk/embl/index.html

[4] www.expasy.org/sprot/

POUR EN SAVOIR PLUS

■ Sur notre site, www.larecherche.fr, des applications pour identifier soi-même des gènes dans un génome.