Présentation de l'*applet* de recherche de gènes

Régis Monte et François Rechenmann
INRIA Rhône-Alpes
Juin 2005

Cette *applet* vous permet de chercher, dans une sous-séquence du génome de la bactérie *B. subtilis*, les gènes, plus précisément les régions codantes, qui s'y trouvent et de caractériser la protéine codée par chacun de ces gènes en exploitant la base de protéines Swiss-Prot. Ce processus de recherche et de caractérisation est appelé « annotation ».

Les termes « gène » et « région codante » ne désignent pas exactement les mêmes entités biologiques. Au sein d'un gène, la région codante (souvent désignée par le sigle anglais CDS pour *CoDing Sequence*) porte la séquence de nucléotides qui, à travers les processus de transcription et de traduction, dicte la séquence en acides aminés de la protéine. Outre la région codante, un gène comporte de plus courtes portions d'ADN qui jouent un rôle dans les processus de transcription et de traduction : promoteur, opérateur, terminateur, RBS, etc.

NB : Dans un organisme eucaryote, la région codante d'un gène s'obtient en éliminant les introns et en concaténant les exons.

Ce processus d'annotation est décomposé ici en quatre étapes, listées dans le cadre « Etapes » en haut à gauche de la fenêtre de l'*applet*.

Première étape : Sélection d'une CDS candidate

La portion de séquence du génome de *B. subtilis* s'étend de la position 285 000 à la position 291 000 de la séquence génomique complète, telle qu'elle peut être trouvée dans la base de séquences EMBL (http://www.ebi.ac.uk/embl/) ; elle est longue de 6001 nucléotides.

La « carte génomique » affichée par l'*applet* fait apparaître les triplets *start* sous la forme de traits verts verticaux et les triplets *stop* sous la forme de traits rouges, dans les trois phases numérotées +1, +2 et +3 sur le brin direct (orienté de la gauche vers la droite) et dans les trois phases numérotées –1, -2 et –3 sur le brin complémentaire (orienté de la droite vers la gauche). Au sein d'un gène, un RBS ne fait pas partie de la région codante, il n'apparaît donc pas obligatoirement dans la même phase que les *start* et *stop*. C'est pourquoi les motifs RBS apparaissent sur deux lignes distinctes, l'une associée au brin direct, l'autre au brin complémentaire.

Une région codante est une succession non chevauchante de groupes de trois nucléotides, les codons. Elle débute par un codon *start* et se finit par un codon *stop*. Or, il existe trois manières différentes de grouper les éléments d'une séquence trois par trois, selon que l'on commence au premier, au deuxième ou au troisième élément de la séquence. Ces trois manières de grouper les éléments trois par trois déterminent trois phases sur la séquence. Les régions codantes doivent être recherchées dans chacune de ces trois phases, comme si ces dernières déterminaient trois séquences différentes. De plus, un gène peut être porté par n'importe lequel des deux brins complémentaires de cette fameuse double hélice d'ADN. Chacun des brins pouvant être lu selon trois phases, c'est donc en fait dans six séquences différentes, dont cinq virtuelles, que la recherche doit s'effectuer (Figure 1).

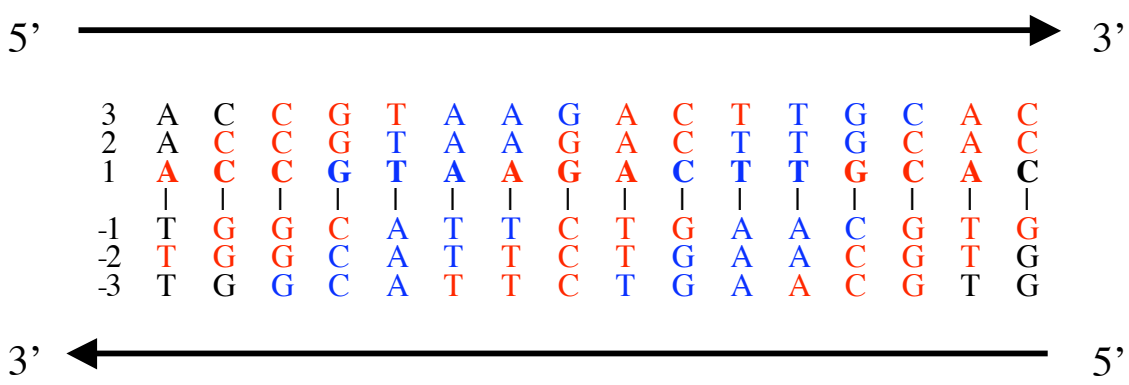

Figure 1 – La séquence « ACCGTAAGACTTGCAC » peut être lue, de gauche à droite, de trois manières différentes selon que l’on commence à considérer les lettres par groupes de trois à la première, à la deuxième ou à la troisième lettre. Ces groupes apparaissent sur la figure alternativement en bleu et en rouge. Il en est de même de la séquence complémentaire, qui est lue de droite à gauche. Une séquence de nucléotides est lue dans le sens 5’ → 3’. Les notations 5’ et 3’ font référence aux atomes de carbone qui apparaissent aux extrémités libres d’un brin d’ADN.

En cliquant sur les boutons étiquetés « START», « STOP » et « RBS », vous faites disparaître et apparaître respectivement les triplets *start* et *stop* et les motifs RBS. Ainsi, en ne faisant apparaître que les triplets *stop*, plusieurs ORF suffisamment longues se distinguent assez nettement comme des « trous » sur chaque ligne associée à une phase. C’est au sein de ces ORF que vous sélectionnerez une première CDS candidate (Figure 2).

Les RBS apparaissent sur la carte génomique, mais il vous est conseillé de ne pas en tenir compte pour la sélection de CDS et de vous en tenir aux critères énoncés dans l’article : choix sur une phase d’une ORF de plus de 300 nucléotides et du *start* immédiatement en aval du codon *stop* qui précède l’ORF ; le codon *stop* de la CDS est le codon *stop* qui termine l’ORF.

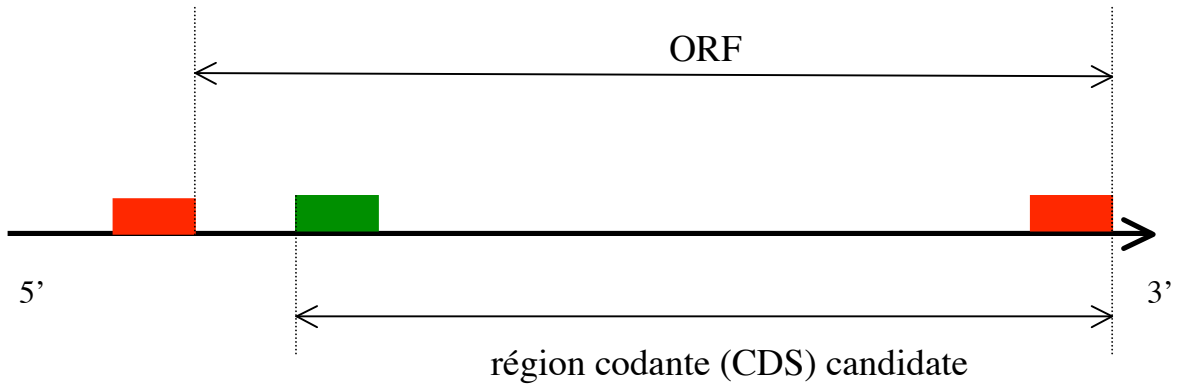

Figure 2 – Une ORF est une portion de séquence délimitée, dans une même phase, par deux triplets *stop* et ne contenant pas de triplet *stop*. Une région codante ne peut se trouver qu’au sein d’une ORF, mais l’existence d’une ORF n’implique pas l’existence d’une région codante en son sein. La stratégie de prédiction des CDS proposée ici consiste à chercher des ORF suffisamment longues et à retenir le premier triplet *start* de l’ORF comme codon *start* de la CDS.

Vous devez donc choisir un *start* et un *stop* dans une même phase. Le choix se fait en positionnant la flèche de la souris sur le trait correspondant au triplet *start* ou *stop*. Si le positionnement est correct, la flèche se transforme en une main à l'index pointé. En cliquant, vous sélectionnez le triplet. La sélection d'un triplet *start* (respectivement *stop*) annule l'éventuelle sélection précédente d'un autre triplet *start* (respectivement un autre triplet *stop*).

La position d'une CDS est déterminée par le rang, compté à partir de 1, du premier nucléotide de son codon *start* et par le rang du dernier nucléotide de son codon *stop*. Les positions sont toutes calculées de la même manière par référence à la séquence de phase 1, que les CDS soient portées par le brin direct ou par le brin complémentaire. Ainsi, sur le brin complémentaire, les rangs sont comptés à partir du nucléotide le plus « à gauche », en 3' donc.

La fenêtre « Messages » fait apparaître des messages qui sont susceptibles de vous aider en cas de choix erronés. Par exemple, un message vous rappellera que, sur les phases négatives, le triplet *start* doit se situer « à gauche » du triplet *stop*.

Enfin, il est souvent nécessaire d'effectuer une opération de zoom pour mieux distinguer les triplets et les sélectionner. À un niveau de zoom élevé, vous verrez apparaître les lettres qui composent la séquence de chaque brin.

Deuxième étape : Traduction de la CDS candidate

Dans cette deuxième étape, la CDS candidate est traduite en une séquence polypeptidique, c'est-à-dire un enchaînement d'acides aminés. À chaque codon est associé, *via* le code génétique, un acide aminé. La séquence de la CDS est ainsi parcourue séquentiellement du premier codon, un *start*, au dernier codon, un *stop*.

Ici, le code génétique est figuré par des disques concentriques découpés en secteurs : le premier acide nucléique du codon détermine l'un des quadrants du disque le plus intérieur, le deuxième l'un des 16 secteurs figurant dans le disque immédiatement supérieur et le troisième l'un des 64 secteurs en périphérie du secteur précédent. Ces quadrants et secteurs prennent la même couleur rosée pour un triplet lu sur la séquence. Sur le dernier disque figurent les acides aminés. En cliquant sur le secteur associé à un acide aminé, vous obtenez sa structure chimique, son nom complet et son code en une et trois lettres. Par exemple, à la lecture du codon CTG, c'est le secteur correspondant à l'acide aminé Leucine qui est sélectionné ; cet acide aminé peut être désigné par la seule lettre L ou par les trois lettres Leu.

Le résultat de l'application de l'algorithme de traduction de la séquence de la CDS que vous avez sélectionnée est une séquence de lettres écrite dans l'alphabet des 20 lettres associées aux 20 acides aminés. Si N est la longueur de la CDS candidate, la longueur de la séquence polypeptidique est N / 3 – 1 ; en effet, le *stop* n'est pas traduit.

La traduction formelle de la séquence de la CDS candidate en une séquence polypeptidique est bien entendu une abstraction et une simplification de la réalité biologique. Au sein de la cellule, la CDS est une région d'un des deux brins de la molécule d'ADN. Cette région est tout d'abord copiée en une molécule d'ARN, identique à la région d'ADN originelle, à la seule différence que l'uracile (désignée par la lettre U) y remplace la thymine (T). Une séquence d'ARN se représente donc par une chaîne de caractères écrite dans l'alphabet A, U, C et G. Ce processus de copie préserve l'archive qu'est la molécule d'ADN ; il est appelé « transcription ». La transcription fait intervenir une molécule, l'ARN polymérase, qui parcours la région d'ADN et synthétise progressivement la copie.

La molécule d'ARN ainsi obtenue est ensuite traduite en une chaîne polypeptidique. Cette traduction fait intervenir des molécules diverses, telles que les ribosomes et les ARN de transfert. Un ARN de transfert est une molécule qui réalise la correspondance codon – acide aminé : elle porte d'un côté un anti-codon, qui peut s'apparier à un codon lu par le ribosome, et de l'autre l'acide aminé correspondant à ce codon. Autrement dit, l'ensemble des différents ARN de transfert constitue la matérialisation du code génétique.

Troisième étape : Recherche dans la base Swiss-Prot

La séquence polypeptidique qui résulte de l'étape précédente peut être comparée aux séquences des 185 000 protéines répertoriées dans la base de données Swiss-Prot (http://www.expasy.org/sprot/). Le principe est de trouver des séquences qui ressemblent à la vôtre. L'existence de séquences similaires signifie que votre protéine ressemble à une protéine répertoriée, qui appartient à un organisme plus ou moins apparenté.

Ici, l'organisme est la bactérie *B. subtilis* dont le protéome, c'est-à-dire l'ensemble des protéines codées par les gènes, est bien connu. Si vous avez sélectionné une vraie CDS, votre protéine doit ressembler à une des protéines connues du protéome de *B. subtilis*. Inversement, si vous avez sélectionné une région qui n'est pas une CDS, la recherche dans Swiss-Prot ne ramènera pas de protéine dont la séquence ressemble à votre séquence.

Bien entendu, lors de l'identification de régions codantes dans un génome nouvellement séquencé, les recherches dans des bases de données renvoient des séquences similaires exclusivement associées à d'autres organismes que celui qui est étudié.

Pour lancer la recherche, il vous faut tout d'abord « copier » la séquence traduite qui apparaît dans le cadre intitulé « Séquence polypeptidique », puis appeler le programme de recherche dans la base Swiss-Prot en cliquant sur l'un des deux boutons « Serveur genevois» ou « Serveur australien». La base Swiss-Prot est en effet disponible sur plusieurs sites différents, si bien que si l'un est indisponible, la recherche peut malgré tout être lancée sur un autre.

Une nouvelle fenêtre de votre navigateur s'ouvre. Il vous suffit 1) de « coller » la séquence polypeptidique dans le cadre situé juste en dessous de l'instruction « Enter a Swiss-Prot/TrEMBL accession number or a PROTEIN sequence in RAW format », 2) de sélectionner « NiceBlast » dans la liste « Output format », 3) de cliquer sur le bouton « Run Blast » pour lancer la recherche (qui met en œuvre un programme appelé « Blast »).

La recherche prend normalement quelques secondes, mais peut être plus longue si le serveur est très sollicité (vous n'êtes probablement pas le seul à effectuer une recherche dans Swiss-Prot au même moment). Si l'attente est vraiment trop longue, relancez la recherche sur l'autre serveur.

Le programme d'interrogation de Swiss-Prot renvoie une liste de protéines dont la séquence ressemble à la vôtre, même si les deux séquences n'ont pas la même longueur. Il affiche un degré de similarité qu'il n'est pas nécessaire ici d'interpréter de façon précise. Il vous suffit de regarder le code couleur dans la colonne « Score » : une protéine listée avec le code de couleur vert est très semblable à la séquence donnée en entrée ; un code rouge indique une protéine trop peu semblable.

Si le code de la première ligne est vert, vous pouvez considérer que la recherche confirme votre CDS candidate. Faites alors un « Copier » du nom de la protéine qui apparaît dans la

colonne « Entry name » (ou encore du libellé de la protéine qui apparaît en gras sur la ligne en dessous), retournez dans la fenêtre de l'*applet* et collez le nom ou le libellé dans le cadre en bas à droite de la fenêtre, puis cliquez sur le bouton « Valider l'annotation ». Vous passez alors à la quatrième et dernière étape.

Dans le cas contraire, cliquez sur le bouton « Annuler l'annotation », vous revenez alors à la première étape pour choisir une autre CDS candidate.

Quatrième étape : Affichage de la CDS confirmée

La carte fait maintenant apparaître une épaisse flèche bleue en coïncidence avec la zone que vous aviez initialement sélectionnée comme une CDS candidate. Au-dessus de cette flèche apparaît le nom ou le libellé de la protéine retrouvée similaire dans Swiss-Prot.

Étapes suivantes

Vous pouvez recommencer ce processus d'identification et de validation d'une région codante dans cette portion du génome de la bactérie *B. subtilis*.

Notez bien que des régions codantes ne peuvent se chevaucher, qu'elles soient sur un même brin ou non, qu'elles soient sur une même phase ou non.

Rappelez-vous que la séquence génomique complète de cette bactérie comporte 4,2 millions de nucléotides et contient 4106 gènes.

Remarques

Il convient de bien prendre conscience des limites de la démarche suivie. Les critères simples de sélection d'une CDS candidate adoptés ici ne sont pas très sélectifs et discriminants. Le fait de trouver des séquences similaires dans Swiss-Prot conforte l'hypothèse de l'existence d'une région codante, mais des erreurs importantes peuvent subsister sur les bornes de cette région. Ainsi, le choix du premier codon *start* apparaissant dans l'ORF doit fréquemment être remis en faveur d'un triplet *start* plus en aval, conduisant à une CDS plus courte.

À l'inverse, le critère de taille (la sélection de CDS de plus de 300 nucléotides) peut s'avérer trop restrictif et interdire l'identification de CDS courtes.

Enfin, il est important de rappeler que, si l'utilisation d'autres méthodes plus évoluées permettent de les affiner, seules les techniques expérimentales sont à même de confirmer les prédictions issues des programmes bioinformatiques.