Par BELHOCINE Mohamed
(Version en révision)

E-gène : un projet d'annotation de séquences biologiques

Chaque cellule humaine, par exemple, contient une molécule d'ADN dont la longueur est d'environ 3 milliards de nucléotides, que l'on peut ainsi représenter comme un texte - certes un peu ennuyeux pour le non-spécialiste, car écrit dans un alphabet de 4 lettres seulement - de 750 000 pages.

Grisés par ces capacités de séquençage quasi illimitées, les projets de recherche s'accumulent à travers le monde : le séquençage de génomes produit en une journée un nombre de séquences d'ADN équivalant à 70 annuaires (de plus de 1000 pages chacun), là où, il y a encore quelques années, on en séquençait péniblement cinq pages, les séquenceurs sont ces machines formidables qui permettent d'une part, de lire la molécule d'ADN constituée de la succession de quatre molécules fondamentales (appelées nucléotides) et représentées par l'initiale de leur base (A pour Adénine, C pour Cytosine, G pour Guanine, T pour Thymine), et d'autre part, de la représenter sous la forme d'une longue chaîne de caractères : la séquence génomique.

En parallèle, un autre pan récent de la génomique prend un essor nouveau grâce à ces technologies à haut débit : il s'agit de la méta-génomique, une nouvelle approche révolutionnaire du séquençage.

En 2004, Craig Venter, un pionnier américain de ce domaine, part avec son voilier dans la mer des Sargasses. Son idée, c'est de prélever des échantillons d'eau, et de séquencer non pas un organisme bien défini, mais le mélange d'organismes présent dans ce bouillon. Que va-t-on y trouver ? Mystère... mais on espère ainsi avoir une image de la biodiversité qui s'y trouve. Cette idée de séquencer des échantillons environnementaux a depuis été appliquée à d'autres milieux, qu'ils soient marins, terrestres (notamment des échantillons de terre prélevés dans des mines) ou biologiques, afin, à chaque fois, d'explorer la diversité des micro-organismes présents. Quels sont les micro-organismes capables de supporter le milieu acide d'une mine de fer ? Quels sont ceux présents dans les eaux côtières, ou à une profondeur de 20 mètres, ou dans une eau à 25 degrés ? Là encore, les projets sont florissants, les séquenceurs vrombissent, livrant des quantités faramineuses de courtes séquences d'ADN, correspondant chacune à l'un des organismes présents dans l'échantillon étudié. Lequel ? Impossible de le déterminer, tant que l'on n'a pas annoté le fragment par des moyens bioinformatiques. Annoter, c'est faire parler une séquence d'ADN : contient-elle un ou plusieurs gènes ? Si oui, quelle pourrait être leur fonction ? Connaît-on déjà des gènes similaires, dans quels organismes ? Peut-on ainsi avoir une idée de l'organisme, ou du moins de la famille dont provient cette séquence d'ADN ?

Craig Venter au microscope, à bord de son voilier Sorcerer II. Photo © J. Craig Venter Institute

Cela ne pose en général guère de problème technique, à condition de maîtriser quelques outils bioinformatiques disponibles en ligne, même si retracer le parcours de certains gènes baladeurs au cours de l'évolution constitue un véritable casse-tête.

Par BELHOCINE Mohamed
(Version en révision)

La principale difficulté, c'est que les séquences biologiques s'accumulent dans les banques de données à une vitesse faramineuse. Les bioinformaticiens sont bien à la peine pour traiter cet océan de données et en extraire les pépites génomiques. Ces pépites, ce sont par exemple de nouveaux gènes qui n'ont jamais été identifiés jusqu'à ce jour ; des enzymes produisant un grand nombre de composés chimiques qui ne peuvent sinon être obtenus que par de coûteux procédés de chimie synthétique ; des bactéries microscopiques produisant, tout au fond de la mer, des lipides qui nous serviront peut-être un jour de biocarburants. L'avenir de l'humanité au fond des mers, en somme...

Mais trouver une pépite demande généralement de tamiser des tonnes de terre sans intérêt. De même, c'est un travail long et ingrat que celui d'annotateur de séquences biologiques. Une tâche qui n'est pas reconnue à sa juste valeur, un véritable « *career killer* », disent certains chercheurs, et personne ne semble prêt à sacrifier sa carrière pour cela.

Alors, une idée qui fait son chemin dans la communauté des génomiciens et des bioinformaticiens est celle de distribuer ce travail entre tous, plutôt que d'attendre que quelques-uns s'attèlent à cette tâche considérable. Les outils collaboratifs existent aujourd'hui, comme les wikis ou les systèmes de gestion de contenu (*Content Management Systems* ou CMS), qui permettent à beaucoup de participer conjointement à cet effort. Récemment, une telle initiative a été proposée pour annoter les protéines (WikiProteins) ou les gènes identifiés, mais peu explorés (WikiGene). Alors, pourquoi ne pas proposer une annotation des séquences méta-génomiques sur le même principe ?

Une initiative allant dans ce sens a récemment été présentée par une équipe d'enseignants-chercheurs des Universités de Marseille qui compte parmi elle le jeune chercheur algérien *Mohamed Belhocine* où il a participé au développement de leur projet.

Avec l'aide de son collègue le développeur *Mohamed Hassairi* -un informaticien du 1er degré- ils ont reprit cette initiative. Leur idée est de proposer un environnement web, qui balise les différentes étapes de l'annotation de séquence, invitant le contributeur à remplir des cadres bien précis contenant les résultats de différentes analyses bioinformatiques : recherche d'homologues (donc de séquences semblables déjà connues dans d'autres organismes) dans les banques de séquences, recherche de signatures typiques dans la séquence d'ADN qui pourraient pointer vers la présence d'un gène codant, reconstruction phylogénétique, recherche de fonctions potentielles, etc.

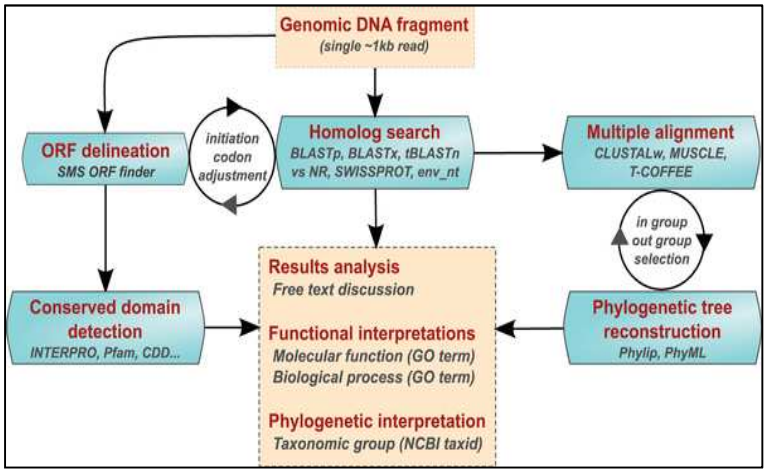


Diagramme du protocole de l'annotation

Deux projets ont vue le jour un projet français (mais aussi international) nommé ***Annotathon*** et un projet purement algérien nommé ***Electronique-gene* (*E-gene :*** http://egene.dixkey.com/**)** le principe est le même dans les deux projets : Un utilisateur peut donc se connecter au système, tirer au hasard une séquence méta-génomique, l'annoter et soumettre le résultat de son travail.

Menu de la page d'accueil du site E-gène version 2

Par BELHOCINE Mohamed
(Version en révision)

Les outils bioinformatique utilisé dans le protocole du projet E-gene *:*

Analyses	Websites	Tools	options & parameters	Results
ORF delineation	http://annotathon.org/sms2/	SMS ORF Finder	With/out start codons	Putative ORF coordinates, translation, initiation and stop codons
Molecular weight	http://annotathon.org/sms2/protein_mw.html	SMS suite	Default	Molecular weight
3D protein structure	http://cbsuapps.tc.cornell.edu/	LOOPP 4	Default	PDB File
Conserved domains	http://www.ebi.ac.uk/Tools/InterProScan/	Interpro	Default	Putative domains and associated molecular functions
Homolog search	http://blast.ncbi.nlm.nih.gov/Blast.cgi	BLASTp	Databases: SWISSPROT and NR	Summary of Blast output file Selection of putative homologous sequences (Fasta format)
Multiple alignments	http://www.ebi.ac.uk/Tools/clustalw2/index.html	CLUSTALw	Ouput format (CulstalW)	ClustalW formated multiple alignment
Phylogenetic trees	http://www.phylogeny.fr/version2_cgi/alacarte.cgi	PhyML	Neighbor-joining (NJ)	Phylogenetic tree with properly annotated branches (protein names, species acronyms)
Taxonomic classification	http://www.ncbi.nlm.nih.gov/Taxonomy/	NCBI taxid	entered the scientific name	NCBI numerical identifier or scientific name (to fill in specific boxes)
Functional annotation	http://www.geneontology.org/	GO terms	Molecular functions & Biological processes	GO terms to fill in specific boxes

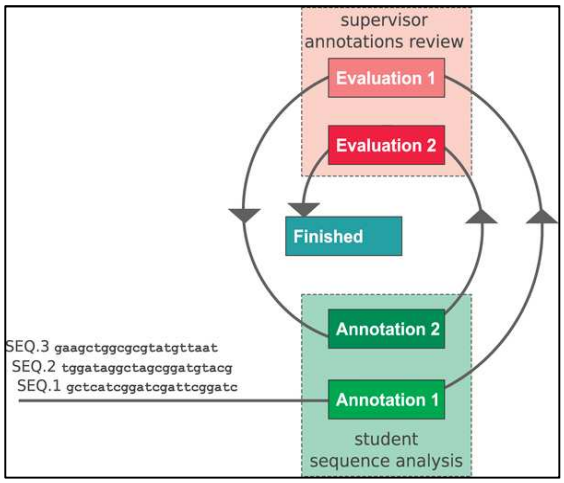
Le système étant en place, encore faut-il trouver des volontaires pour participer à cet effort collectif. Les enseignants de Marseille ont optés pour leurs étudiants et de son coté le jeune chercheur *Belhocine,* a immédiatement pensé aux étudiants de licence de son université d'origine *Blida -Saad Dahleb* -en Algérie, ***petites mains de la génomique*** comme l'a titré *Stéphane Foucart*, journaliste du Monde dans l'article consacré à ces projets (« *Les étudiants en biologie, petites mains de la génomique* », *Le Monde* du 6 décembre 2008).

Un énorme profit scientifique et pédagogique à en tirer. Quel meilleur exercice en effet que de faire utiliser de manière répétée, dans différentes conditions, les outils bioinformatiques standards ? Et quelle meilleure motivation pour les étudiants que de savoir qu'ils sont en train d'analyser des séquences que personne, avant eux, n'a étudiées ? Ils se sentent véritablement acteurs de la science, même s'ils sont parfois troublés par les analyses divergentes.

Par BELHOCINE Mohamed
(Version en révision)

Chaque étudiant annote en moyenne 3 séquences, et peut recueillir les critiques des enseignants sur son annotation, afin de les intégrer dans une seconde version, finale, de cette annotation.

Au système ensuite de trouver un moyen de valider le travail de l'annotateur, par exemple en confrontant de manière automatique les annotations de plusieurs personnes sur une même séquence : ont-ils trouvé les mêmes fonctions ? Les mêmes propriétés de structure pour les protéines ?



Système d'évaluation par l'essai et l'erreur

Concrètement, en trois ans, plus de 200 étudiants de licence (génétique, microbiologie et biologie cellulaire et moléculaire) ont participé à ce projet, annotant au total 1,4 millions de bases de séquences méta-génomiques. 1,4 millions, ce n'est encore qu'une goutte d'eau face aux quelques gigabases de séquences disponibles.

Equipes:				
	Libre	43 annotateurs	En cours	Fermeture le:2011-12-30
	BCM2011	25 annotateurs	En cours	Fermeture le:2011-01-31
	G2011	24 annotateurs	En cours	Fermeture le:2011-01-31
	Micro2011	23 annotateurs	En cours	Fermeture le:2011-01-31
	G2009	23 annotateurs	Terminé	Fermeture le:2009-10-28
	Microbiologie	23 annotateurs	Terminé	Fermeture le:2010-06-13
	BCM2010	21 annotateurs	Terminé	Fermeture le:2010-06-13
	BCM 2009	20 annotateurs	Terminé	Fermeture le:2009-11-27
	G2010	19 annotateurs	Terminé	Fermeture le:2010-02-23

Liste des équipes participant dans le projet E-gène depuis 2008

Après l'exécution de ce projet, l'impression corroborée par les réactions des étudiants, est que cette méthode d'enseignement a beaucoup plus de succès que les méthodes classiques. L'équipe BHK a été très impressionnée par la familiarisation des étudiants avec les principales boîtes à outils *in silico*. On admet l'intérêt d'exposer les étudiants à de véritables situations de recherche au début de leur formation.

A présent, on se prend à rêver d'un vaste réseau d'étudiants-annotateurs au niveau national et international qui, tel une grille de calcul distribué, alimente les bases de données de la communauté de génomiciens et contribue ainsi à faire progresser la connaissance. Un doux rêve peut-être, mais qui en tout cas est dans l'air du temps...

En fin, Nous tenons à remercier les 234 étudiants (L3 génétique, L3 Microbiologie et L3 biologie moléculaire et cellulaire) qui ont participé à ce projet, avec une mention spéciale à ceux qui nous ont donné leurs commentaires et leurs critiques constructives et ont cru en ce projet et ont fait de leur mieux pour sa réalisation.
Nous tenons à remercier Mr Ramadan Mohamed Saïd de son soutien et ses aides précieuses et ainsi son engagement irréprochable. Sans oublier le 3ème pilier du groupe Idir Kacel et toute l'équipe BHK : administrateurs et assistantes.