RAPPORT DE RECHERCHE

L R I

STRATEGIES D'ANNOTATION FONCTIONNELLE

TOFFANO-NIOCHE C

Unité Mixte de Recherche 8623
CNRS-Université Paris Sud – LRI

12/2009

Rapport de Recherche N° 1532

CNRS – Université de Paris Sud
Centre d'Orsay
LABORATOIRE DE RECHERCHE EN INFORMATIQUE
Bâtiment 490
91405 ORSAY Cedex (France)

Stratégies d'annotation fonctionnelle

Claire Toffano-Nioche*
IBP Institut de Biothechnologie des Plantes
et LRI Laboratoire de Recherche en Informatique,
Université Paris-Sud XI

Abstract : The production capacities of functional protein annotation don't follow the sequencing rate. Despite the development of annotation softwares, manual annotations are still preferred because of their better quality, even if they are not error free. With the aim to improve annotations, we build up a compendium of the annotation strategies used by annotators of bacterial genomes. We suggest a graphical representation to outline the collected strategies that identify the different steps leading to the annotation. Compendium analysis gives us many leads as the existence of a flow of annotation hypothesis that guides the analysis step choice, the use of variable criteria and thresholds that are associated to a set of references built at each step. Finally, as for the quality evaluation we propose to distinguish the quality of annotation from the quality of the used annotation process.

Résumé : Les capacités de production d'annotation fonctionnelle des protéines ne suivent pas celles de production des séquences. Malgré le développement des logiciels d'annotation automatique, les annotations manuelles leur sont encore aujourd'hui préférées par les biologistes car jugées de meilleure qualité, même si elles ne sont pas dénuées d'erreur. Avec cet objectif d'amélioration des annotations, nous avons constitué un recueil des stratégies mises en oeuvre par des annotateurs de génomes bactériens. Nous proposons une représentation graphique pour schématiser les stratégies recueillies en identifiant les différentes étapes qui conduisent à l'annotation. L'analyse du recueil révèle différentes pistes telles que l'existence d'un flux de l'hypothèse d'annotation qui guide le choix de l'enchaînement des étapes, l'usage de critères et de seuils de décisions variables et associés à un ensemble de référence qui est construit à chaque étape. Enfin, pour évaluer la qualité des annotations, nous proposons de dissocier l'évaluation la qualité des annotations de celle de la qualité du processus d'annotation mis en oeuvre.

* actuellement à l'IGM, Institut de Génétique et de Microbiologie

Contexte et remerciements

Ce rapport décrit une partie des recherches que j'ai effectuées au sein de l'équipe Bioinformatique (Christine Froidevaux) du LRI[1] (UMR[2] 8623, Michel Beaudouin-Lafon) qui m'a accueillie de juin 2007 à janvier 2009.

Ce rapport n'aurait vu le jour sans mes interactions avec les personnes suivantes que je tiens ici à remercier.
Je remercie Michel Beaudouin-Lafon, directeur du LRI de l'écoute et du soutien qu'il a manifesté lors de nos rencontres. Mes remerciements vont ensuite à Christine Froidevaux, responsable de l'équipe bioinformatique du LRI, pour m'avoir accueillie au sein de son équipe, proposé cette thématique de recherche, accompagnée et encouragée dans sa réalisation, et surtout, aidée à en formaliser chaque étape de maturation.

Je remercie aussi les annotateurs de l'INRA que j'ai interviewés pour leur accueil lors des moments que j'ai passées à leurs côtés, pour leur usage d'annotation qu'ils ont bien voulu partager avec moi, et pour la relecture des interviews qu'ils ont assuré par la suite. Je remercie Eric Duchaud, Hélène Falentin, et Christophe Monnet de la confiance qu'ils m'ont chacun accordée alors qu'ils ne me connaissaient pas.

Je remercie mes collègues dont les remarques précieuses ont directement enrichi la réalisation de cette étude : Jérôme Azé, Sarah Cohen-Boulakia, Lucie Gentils, Jean-François Gibrat, Valentin Loux et Anne Poupon. Collègues, réunis pour la plupart grâce au projet RAFALE (ACI-IMP Bio) et qui a été le point de départ de cette étude concernant les stratégies d'annotation fonctionnelle.

Enfin, pendant toute ou une partie de ces années, beaucoup de personnes ont soutenu mes projets et je gâche un peu de forêt afin de les remercier aussi ici : Nicolas Beaume, Gaëlle Claisse, Hervé Delacroix, Alain Denise, Martin Kreis, Alain Lecharny, Christine Martin, Florence Mougel-Imbert, Cédric Saule, Michel Termier, Vincent Thareau, Bill Turner.

1 LRI : Laboratoire de Recherche en Informatique
2 UMR : Unité Mixte de Recherche ; Université Paris-Sud et CNRS dans ce cas.

Table des matières

1 - Introduction

Depuis les publications des premiers génomes entièrement séquencés, les plates-formes de séquençage mettent de plus en plus de séquences à disposition dans les bases de données. Ces séquences acquièrent une valeur auprès des biologistes après deux étapes d'« annotation ». La première étape, l'annotation structurale, permet d'identifier les zones de la séquence génomique qui déterminent les séquences protéiques. Une seconde étape cherche à associer une information aux zones identifiées et en particulier la ou les fonctions de la protéine dans l'organisme. Cette deuxième étape, l'annotation fonctionnelle, se réalise en précisant les réactions biochimiques auxquelles la protéine participe ou ses rôles dans les processus biologiques. La première source d'information de l'annotation fonctionnelle provient des études réalisées en paillasse par les biologistes. Cependant, la caractérisation biologique d'une protéine étant très longue, peu de protéines de l'ensemble des séquences disponibles sont décrites de cette façon. Ainsi, la majorité des annotations fonctionnelles actuelles reposent sur l'hypothèse qu'une similarité entre deux séquences correspond à une similarité de fonction. Cette hypothèse permet ainsi de transférer les annotations fonctionnelles d'une protéine déjà annotée vers une protéine à annoter lorsque la similarité entre les deux protéines est jugée suffisante. Beaucoup de logiciels proposant une annotation fonctionnelle automatique ont été développés afin de suivre le rythme du séquençage.

Cependant, de transferts en transferts des erreurs apparaissent et se propagent [Brenner *et al.*, 1999 ; Devos *et al.*, 2001 ; Gilks *et al.*, 2002 ; Wieser *et al.,* 2004]. Et cela d'autant plus que les transferts se font sans garder ni la trace du degré de similarité entre les protéines qui a été retenu, ni les informations qui étaient disponibles dans les ressources au moment du transfert. Ainsi, aujourd'hui encore, les annotations fonctionnelles de protéines réalisées manuellement par des annotateurs sont jugées plus fiables que les annotations automatiques. Leur différence par rapport aux annotations automatiques réside principalement dans le réglage des paramètres et l'interprétation pas à pas des résultats, ainsi que dans le fait qu'elles utilisent les ressources à jour.

L'annotation fonctionnelle est une tâche essentiellement dépendante des informations que l'on peut associer à la protéine à annoter, et réalisée de façon manuelle, elle est donc directement dépendante du temps que l'annotateur passe à explorer les différentes ressources, constituées des différentes sources d'information et de l'exécution d'outils bioinformatiques de prédictions. Un des désavantages des annotations manuelles est qu'elles ne permettent pas de suivre l'évolution des techniques de production des séquences. Ainsi, l'écart entre la proportion de nouvelles séquences par rapport à celles qui sont annotées augmente-t-il continuellement. Ceci est indiqué par exemple par la comparaison de la quantité de protéines référencées dans la section TrEMBL[3] de la base de données UniProtKB [The UniProt Consortium, 2008] qui rassemble l'ensemble des séquences identifiées (7001017 protéines, version 39.7 du 20-Jan-2009) à celle de la section Swiss-Prot, base de connaissances au sein de laquelle les protéines sont enrichies d'annotations contrôlées manuellement (408099 protéines, version 56.7 du 20-Jan-2009). Ces dernières années, les techniques de séquençage abordent une nouvelle phase dans leurs capacités de production de séquences. On parle de techniques à très haut débit ou de séquençage profond [Shendure and Ji, 2008 ; Lister *et al.*, 2009]. Il s'agit d'un véritable changement d'échelle, alors que le précédent n'est toujours pas relevé par la phase d'annotation.

3 TrEMBL : Translated European Molecular Biology Laboratory ; base européenne de données des séquences protéiques

Diverses solutions ont été proposées pour réduire l'ensemble de ces séquences protéiques non annotées. Il s'agit par exemple de projets de comparaisons de toutes les séquences de protéines contre toutes, mais les résultats de ces projets sont peu utilisés des biologistes, probablement suite à un défaut de compréhension des choix techniques sous-jacents [Bastien *et al.*, 2005] ou aussi peut-être parce que ces calculs, qui concernent l'ensemble des protéines disponibles, ne correspondent pas aux attentes et usages des biologistes qui ne s'intéressent en général qu'à un faible nombre de protéines. Une autre solution est la mise en place d'un site wiki dédié à l'annotation fonctionnelle des protéines mais ce projet, dont la phase de l'appel à participation est parue en 2008 [Mons *et al.*, 2008], n'en est qu'à ses débuts et n'est donc pas encore utilisable. Cette solution suppose aussi que les biologistes acceptent de jouer le jeu en rendant accessible les résultats de leurs travaux au fur et à mesure de leur avancée.

Toutes ces remarques aboutissent au constat actuel : un biologiste ré-annote aujourd'hui manuellement les protéines qui l'intéressent. Il y passe beaucoup de temps car d'une part, ce n'est pas son activité principale s'il n'est pas annotateur puisqu'il ne possède pas la rapidité apportée par l'expérience et d'autre part, cette activité lui demande d'analyser les résultats de logiciels bioinformatiques sans toujours en maîtriser les modes de fonctionnement sous-jacents ni les bornes de validité des résultats obtenus.

Plutôt que de proposer un système d'annotation automatique qu'il faudra relancer à chaque mise à jour des annotations, l'objectif poursuivi ici est alors de concevoir les premières bases d'une méthode accessible aux biologistes pour qu'ils puissent ré-annoter efficacement les quelques protéines d'intérêt qui les concernent.
Les propriétés que devra satisfaire cette méthode pour être utilisée par les biologistes sont les suivantes : être accessible à la communauté des biologistes, exploiter les versions en cours des bases de données et des outils bioinformatique, et non les versions d'une mise à jour antérieure, et enfin, nécessiter peu de compétences en programmation. Pour se substituer à l'annotation manuelle, elle devra permettre de lancer plus efficacement différents outils de prédiction et faciliter la consultation de bases de données variées.

Depuis quelques années, des logiciels proposent au sein d'un même environnement une aide à la conception de workflows scientifiques [Stevens *et al.*, 2003, Oinn *et al.*, 2004 ; Shah *et al.* 2004 ; Altintas *et al.*, 2004]. A l'aide d'une interface dédiée à la conception, ces environnements réunissent :

- une liste d'outils distants et accessibles par le web, les web services [Romano *et al.*, 2005],
- un formalisme décrivant les objets manipulés : les analyses, les données en entrées et sorties, les résultats pour chaque étape de l'analyse, et un langage permettant d'enchaîner les analyses,
- un moteur de contrôle du flux de données actif lors de l'exécution du workflow.

La composition d'une analyse sous un éditeur de workflow à l'aide des listes de web services ne nécessite que peu de connaissance en programmation et le caractère fortement typé des objets guide l'enchaînement des analyses. Plusieurs environnements dédiés aux analyses bioinformatiques ont vu le jour [Hull *et al.*, 2006 ; Ludäscher et al. 2006]. Les chercheurs de la communauté Taverna s'appuient sur un espace ouvert et accessible à tous par le site internet « myexperiment », qui permet à chacun de rechercher, utiliser et partager les workflows scientifiques développés [De Roure *et al.*, 2007].
Ainsi, les environnements de conception de workflows scientifiques représentent une solution qui pourrait répondre aux attentes précédemment évoquées pour une méthode d'annotation.

L'objectif à terme étant de concevoir un workflow d'annotation fonctionnelle qui soient adaptés aux utilisateurs finals, nous proposons ici une analyse du processus manuel d'annotation fonctionnelle tel qu'il est pratiqué par les annotateurs humains. Nous nous sommes intéressés à l'activité d'annotation fonctionnelle de génomes bactériens réalisés grâce à la plate-forme AGMIAL [Bryson *et al.*, 2006]. Cette activité comprend plusieurs tâches et de façon à les identifier, nous avons pris appui sur un recueil de plusieurs annotations.

Le plan du présent rapport est le suivant : après une présentation de la tâche d'annotation fonctionnelle et du recueil des stratégies d'annotation de protéines bactériennes que j'ai mené auprès de trois annotateurs, je décris les règles de schématisations des annotations que j'ai appliquées afin d'identifier les différentes tâches relevées dans le recueil. Les parties suivantes présentent différentes analyses initiées à partir de ces schématisations et abordent plusieurs points : la multiplicité sémantique d'une tâche, l'enchaînement temporel de tâches, leur hiérarchisation et enfin, l'établissement d'une stratégie globale d'annotation fonctionnelle. En conclusion, sont discutés certains des aspects soulevés lors de cette étude : la description d'une méthodologie semblable qui est retrouvée dans plusieurs des tâches réalisées, l'intervention de seuils de décisions sur un ensemble de paramètres choisis par chaque annotateur et à chaque étape, l'influence de la confiance des annotateurs dans les bases de données explorées et enfin, comment l'attribution d'une note de qualité aux annotations réalisées les enrichiraient en permettant d'estimer non seulement la valeur de l'annotation mais aussi le processus d'annotation.

2 - Annotation fonctionnelle de protéine

Pour associer une annotation à une séquence protéique, l'annotateur exploite différents critères et met en œuvre plusieurs étapes. Lorsque l'on suppose l'annotation structurale résolue, annoter fonctionnellement une protéine consiste à identifier i) ses caractéristiques intrinsèques directement calculables à partir de la séquence protéique, ii) les caractéristiques issues de prédictions apportées par l'analyse des résultats des logiciels bioinformatiques et iii) une protéine déjà annotée dans le but de transférer les annotations à la protéine à annoter.

Les paramètres intrinsèques directement calculables concernent par exemple la longueur en acides aminés, la séquence, le point isoélectrique, leur début et fin de traduction.
Les paramètres prédits nécessitent le lancement d'un logiciel bioinformatique et l'analyse des résultats obtenus. Pour réaliser cette analyse, l'annotateur s'appuie sur les valeurs des scores proposés par les logiciels si elles sont suffisamment discriminantes ou observe les résultats plus en détails dans le cas contraire avant de conclure. Les prédictions concernent par exemple, la liste et l'organisation de motifs protéiques déjà identifiés dans d'autres séquences, des signaux d'adressage ou révélant des segments transmembranaires (présence et nombre), la localisation sub-cellulaire de la protéine dans les différents compartiments cellulaires, ou encore des informations issues du contexte génique comme par exemple, les annotations fonctionnelles associées aux gènes présents en amont et en aval du gène codant pour la protéine en cours d'annotation.

Ces caractéristiques obtenues à partir de la séquence de la protéine à annoter constituent une partie de l'annotation fonctionnelle à laquelle s'ajoute, lorsque c'est possible, une hypothèse quant à sa fonction biochimique et/ou biologique dans l'organisme.
Dans très peu de cas, la fonction de la protéine a été étudiée par les biologistes et est décrite dans les publications. La fonction reprise dans l'annotation est cependant résumée par quelques mots qui

synthétisent les observations mises en évidence. Dans la majorité des cas, l'annotation est plutôt une hypothèse car aucune vérification expérimentale n'y est associée. Le transfert d'annotations d'une séquence à une autre repose sur leur similarité de séquences. On suppose que les protéines de séquences proches n'ont que peu ou pas divergé dans le temps et partagent alors une même fonction, même si cette définition est très large et que certains contre-exemples existent [Audit *et al.*, 2007].
L'annotateur choisit les annotations qu'il transfère sur la base des similarités de séquences qui existent entre la protéine à annoter et les protéines déjà annotées et répertoriées dans les bases de données de séquences. La similarité entre deux séquences est évaluée à partir de leur alignement et peut parfois être réduite au dépassement ou non d'un seuil sur le score des alignements calculés par les outils bioinformatiques d'alignements de séquences (blast [Altschul *et al.*, 1990]) ou d'identification de motifs (interproscan [Mulder and Apweiler, 2007]).

3 - Recueil d'annotations fonctionnelles

3.1 - Cadre du recueil

3.1.1 Contexte : annotateurs, et environnement d'annotation

Nous nous sommes appuyés sur différentes annotations réalisées par plusieurs annotateurs de l'INRA[4] que nous avons rencontrés par le biais d'une collaboration entre le LRI et l'unité MIG[5] (INRA, Jouy-en-Josas) initiée dans le cadre du projet RAFALE, Règles pour l'Annotation Fonctionnelle semi-Automatisée de la Levure (projet ACI-IMP Bio).
Nous avons interviewés des biologistes qui annotent les protéines de génomes bactériens à l'aide de la plate-forme d'annotation AGMIAL [Bryson *et al.*, 2006] développée par l'unité. Le recueil rassemble les annotations de 29 protéines réalisées par 3 annotateurs en charge de l'annotation d'un génome bactérien différent (*Arthrobacter arilaitensis*, 12 annotations ; *Flavobacterium branchrophilum,* 6 annotations ; *Propionibacterium freudenreichii,* 11 annotations).

Au moment du recueil, certaines des annotations de *F.branchrophilum* constituaient une deuxième passe d'annotation, la première passe n'ayant pas abouti suite à un problème particulier nécessitant plus de temps et donc d'y revenir ultérieurement (cas des protéines fbr3271 et fbr1542).

3.1.2 Impacts de la distance évolutive dans la stratégie d'annotation

La distance évolutive de chacun des 3 génomes avec un génome déjà annoté et proche n'est pas identique. En particulier, *F.branchrophilum* est très proche de *Flavobacterium psychrophilum* qui a été annoté juste auparavant par l'annotateur [Duchaud *et al.*, 2007] tandis que les génomes annotés les plus proches des deux autres eubactéries sont plus éloignés. Pour aider à l'annotation d'*A.arilaitensis,* l'annotation des protéines de deux autres arthrobactéries existe mais elles ne partagent pas la même niche écologique, *A.arilaitensis* se développant sur du fromage alors que les deux autres vivent dans le sol. Dans le cas de *P.freudenreichii,* le génome d'une autre propionibactérie est, là aussi séquencé, mais les annotations associées ne conviennent pas à l'annotateur. Les annotateurs d'*A.arilaitensis* et de *P.freudenreichii* s'appuient donc aussi sur les annotations de génomes d'actinomycètes plus éloignés tels que *Corynebacterium glutamicum, Streptomyces coelicolor,* ou *Mycobacterium tuberculosis*. Cette notion de proximité évolutive et

4 INRA : Institut National de la Recherche Agronomique
5 MIG : Mathématique, Informatique et Génome

l'accès à un génome « bien » annoté influent sur la stratégie mise en oeuvre car les annotations en provenance de ces génomes constituent le point de départ de la plupart des informations utilisées lors de la mise en place de l'annotation fonctionnelle.

3.1.3 Spécificité du recueil

Le recueil de stratégies d'annotation mené présente au moins deux caractéristiques qui ont un impact sur les stratégies d'annotation récoltées : i) il couvre l'annotation de génomes bactériens et ii) à l'aide de la plate-forme d'annotation AGMIAL.

Les annotateurs interviewés annotent des génomes bactériens, la description des stratégies obtenues est spécifique de ces génomes. En effet, chez les bactéries procaryotes, l'étape d'annotation structurale pose peu de problème car ces génomes ne contiennent pas ou peu d'intron. Dans le recueil, elle n'a concerné que le positionnement du début de traduction et n'a pas été systématiquement recherchée. L'annotation structurale de génomes eucaryotes dont les gènes contiennent généralement des introns impliquerait alors l'usage d'outils de prédiction de modèle des gènes. L'étape d'annotation structurale est déterminante pour accéder à la séquence protéique et pourrait impliquer plusieurs allers-retours de correction des modèles de gènes lors de l'annotation fonctionnelle, imbriquant alors les deux étapes d'annotation.

Outre la caractéristique bactérienne des stratégies d'annotation récoltées, une deuxième spécificité du recueil concerne l'usage de la plate-forme d'annotation AGMIAL exploitée par les annotateurs. Si certaines bases de données ont parfois été redéfinies et modifiées selon le contexte d'annotation (par exemple, ajout de données supplémentaires) et s'il est arrivé que les annotateurs utilisent des outils autres que ceux présents dans la plate-forme, son usage homogénéise les logiciels et les bases de données exploités et donc les stratégies récoltées.

3.2 – Expertise des annotateurs et stratégies du choix des gènes

L'expertise des trois annotateurs interviewés était différente : il s'agissait de deux novices qui avaient annoté chacun les protéines d'environ 400 gènes, et d'un expert ayant déjà annoté les protéines de plus d'un génome complet.
Le critère de choix de l'ordre des gènes à annoter était lui aussi différent en fonction de l'intérêt ou de l'objectif de l'annotation pour l'annotateur :

- *F.branchrophilum* : protéines choisies au fur et à mesure de l'ordre des gènes dans la séquence génomique après annotation d'une liste de protéines bien connues de l'annotateur,
- *P.freudenreichii* : protéines choisies selon le degré de similarité décroissant avec une protéine de la base de données Swissprot,
- *A.arilaitensis* : choix des protéines de façon à compléter des voies métaboliques.

Notre objectif concernant l'annotation d'une protéine et non la stratégie d'annotation de plusieurs protéines ou d'un génome entier, nous n'avons pas tenu compte de ces différences dans le choix de l'ordre des protéines à annoter.

3.3 - Interviews

L'accès à l'annotation se fait par la séquence protéique grâce à la plate-forme d'annotation AGMIAL. Une grande partie des étapes effectuées par les annotateurs consiste en une navigation au sein de la fiche PAM – AGMIAL (Protein Analysis Manager) accessible avec un navigateur web. Une fiche PAM - AGMIAL est éditée pour chacune des protéines du génome à annoter et rassemble une partie des résultats du lancement de plusieurs logiciels bioinformatiques (voir matériel

supplémentaire 1). Les résultats de chaque logiciel sont présentés dans un tableau de synthèse, enrichi de liens hypertextes afin d'accéder à plus de détails si besoin (*cf.* figure 1). Un accès à la séquence génomique à partir de laquelle la séquence protéique est prédite est possible grâce à un lien vers la fiche CAM – AGMIAL (Contig Analysis Manager) correspondante.

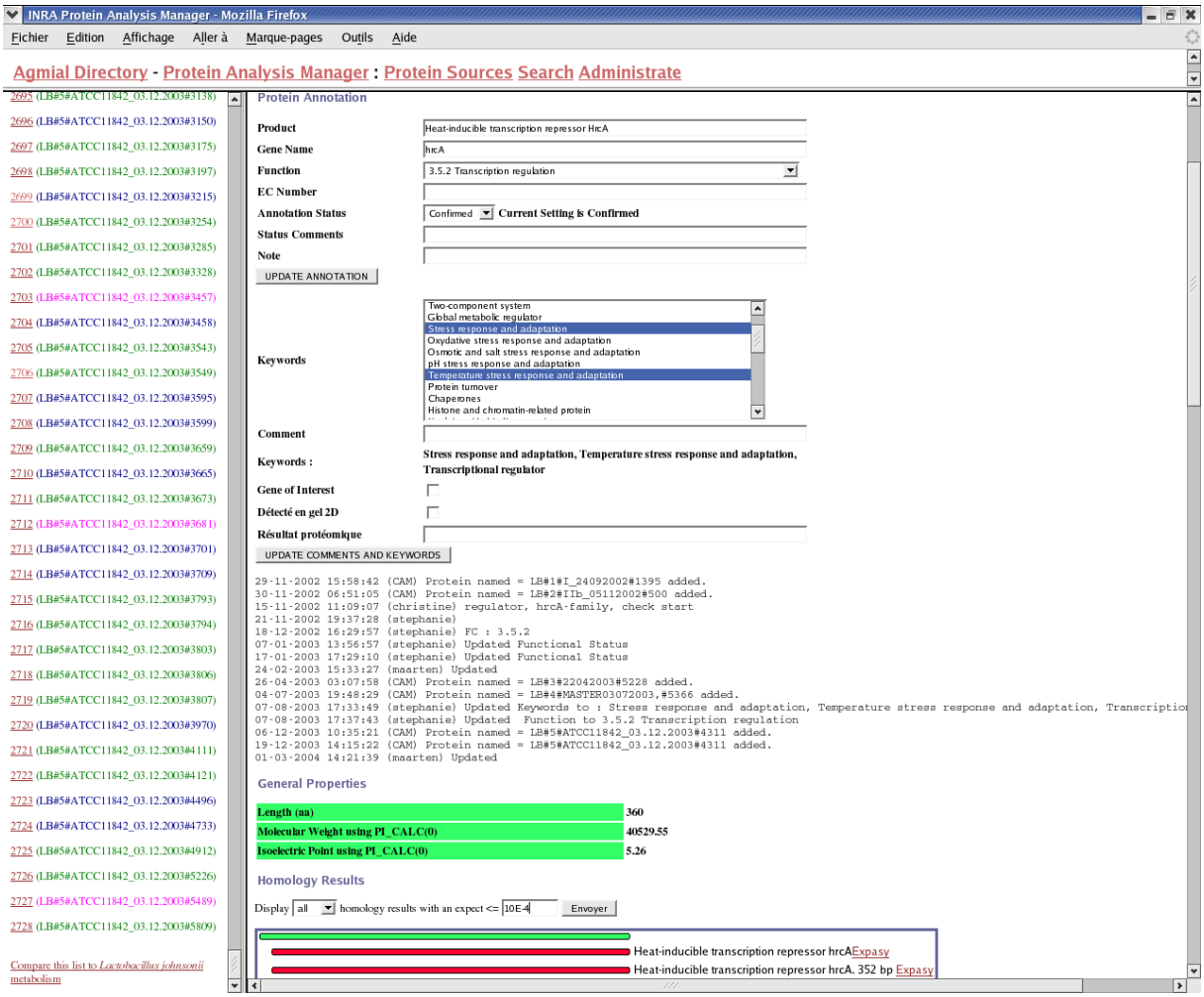

Figure 1 : Copie d'écran du début d'une fiche PAM-AGMIAL

Pour les annotateurs, l'objectif de la tâche d'annotation consiste à remplir le champ « protein product » en langage naturel après l'analyse d'une part, des résultats présentés dans la fiche et d'autre part, de ceux qu'ils ont obtenus en lançant des analyses externes à la plate-forme quand ils ont jugé cela nécessaire.

Nous avons constitué un recueil en texte libre des stratégies mises en œuvre par les annotateurs en suivant pas-à-pas les différentes analyses que les annotateurs ont enchaînées jusqu'à la proposition d'annotation. Un exemple est présenté figure 2. Pour chaque protéine annotée et à chaque étape de l'annotation, nous avons relevé les différentes étapes de navigation effectuées dans la plate-forme AGMIAL. Nous avons aussi répertorié les outils utilisés et les bases de données interrogées à l'extérieur de la plate-forme en recueillant les paramètres choisis lorsque cela était possible. Nous avons porté une attention particulière à la verbalisation des raisons du choix de l'analyse suivante. Certains paramètres de lancement et les valeurs-résultats obtenues ont aussi été notés.

4 - Schématisation des annotations

4.1 - Extraction des actions décrites dans le recueil d'annotation

Nous avons ensuite identifié différentes tâches élémentaires décrites dans le texte des interviews. Dans le cadre de la modélisation de workflow scientifiques, nous avons identifié chaque action de l'annotateur : une analyse d'un résultat présent dans la fiche AGMIAL (ce résultat est issu du lancement d'un logiciel bioinformatique), une transformation de donnée (par exemple, un focus sur une sous-séquence), ou une exécution externe à la plate-forme AGMIAL d'un logiciel bioinformatique ou d'une recherche dans une base de données.

Protéine fbr1970 :
Accès à la fiche PAM par Artemis, fonction « MemD » obtenue par **transfert automatique** depuis l'annotation de F.psychrophilum. Il s'agit d'une protéine dont *l'homologue est clairement identifié*, peu d'erreur sur le transfert de la fonction, ce n'est donc pas mis en doute.
1.Vérification du *start : un aa de différence* entre les deux espèces. Remise en question du start par c**omparaison des deux séquences protéiques** (accès à F.psychrophilum par Agmial) non, *les deux 1er aa sont identiques*, conclusion : le start prédit est confirmé.
2.Résultats **blast des autres homologues** : dans les annotations apparaissent *2 numéros EC* différents.
3.Bascule Uniprot pour retrouver la description de l'orthologue E.coli : 2 numéros EC aussi. Comparaison de la taille des protéines E.coli et F.branchiophilum : 556 et 553 respectivement. Même recherche pour B.subtilis et *mêmes observations* (*deux EC* et taille cohérentes). Conclusion : transfert des 2 numéros EC.
4.Remarque sur les homologues d'autres espèces (défaut des annotations automatiques) : un seul numéro EC (annotation incomplète), et pas forcément toujours le même.
5.On pourrait pousser plus loin l'analyse : MemD doit posséder un domaine transmembranaire selon les features de la fiche Uniprot. Il faudrait donc le chercher sur F.branchiophilum, et s'il est présent, l'identifier aussi sur F.psychrophilum et modifier les deux fiches. Cependant, MemD ne fait pas partie des centres d'intérêts de l'annotateur et ce point n'a donc pas été réalisé.

Figure 2 : Extrait du corpus, protéine fbr1970.
Code de l'analyse du texte : **étapes de navigation** *dans la fiche AGMIAL,* accès externes *aux bases de données et outils, raisons du choix de l'annotation ou de l'analyse suivante,* étapes suggérées *mais non effectuées.*

Ces actions successives ont été ensuite représentées sous forme de schémas qui reflètent le déroulement chronologique de l'annotation d'une protéine.

4.2 - Définition d'une tâche élémentaire

Dans leur classification des tâches bioinformatiques, [Stevens *et al*., 2001] ont établi qu'il existait un lien entre la question biologique que se pose l'utilisateur et la tâche qu'il réalise et qui comprend l'ensemble des actions effectuées pour résoudre cette question. Dans le cas de l'annotation fonctionnelle, nous avons remarqué qu'à chaque étape, l'utilisateur se pose une question biologique et utilise un outil bioinformatique pour y répondre en l'appliquant à une collection de données et ainsi confirme ou infirme une hypothèse biologique. Ici, l'hypothèse biologique consiste en une ou plusieurs propositions d'annotation fonctionnelle. Nous avons donc représenté ce lien par le biais d'une « hypothèse d'annotation ».

A partir du niveau de détail avec lequel a été réalisé le recueil, nous associons à une tâche élémentaire les éléments suivants : une hypothèse d'annotation fonctionnelle, une question biologique et la réalisation d'une tâche qui sera soit de faire appel à un outil bioinformatique en

utilisant une collection de données et des paramètres de lancement, soit d'analyser un résultat proposé par la plate-forme.

L'hypothèse d'annotation fonctionnelle s'enrichit selon le jugement par l'annotateur de l'analyse des résultats obtenus lors de la réalisation de la tâche. Ainsi dans notre représentation, l'enchaînement des tâches n'est pas réalisé sur les objets qui sont manipulés à chaque étape mais par l'hypothèse d'annotation qui évolue à chaque étape. Le type d'enchaînement décrit ici diffère alors de celui des enchaînements des workflows scientifiques. Dans ces derniers, les sorties de chaque étape élémentaire constituent les entrées de la tâche suivante et l'enchaînement est réalisé grâce au fort typage des objets manipulés lors des analyses. Dans notre description, les objets en entrée et sortie de chaque tâche peuvent être différents à chaque étape, les sorties d'une étape ne correspondent pas forcément aux entrées de l'étape suivante. Le flux est celui de l'hypothèse d'annotation et non celui des objets sur lesquels sont réalisées les analyses.

En nous appuyant sur les actions identifiées dans le recueil, nous avons défini chaque tâche élémentaire à l'aide des propriétés suivantes :
- identifiant de la tâche
- description de la tâche
- entrées : type de la collection de données
- sorties : type du résultat
- réalisation : outil logiciel, paramètres de lancement (incluant parfois la base de données de recherche), serveur.

Nous avons identifié 18 tâches élémentaires et leur définition est détaillée dans le tableau constitutif du matériel supplémentaire 2.

Nous avons considéré la confrontation de l'annotation obtenue avec les connaissances de l'annotateur comme une tâche {tca} même s'il n'y a pas de logiciel ou de recherche dans une base de données exploités pour réaliser cette tâche.

4.3 - Règles de schématisation

4.3.1 - Éléments constitutifs

Nous décrivons alors l'annotation fonctionnelle à l'aide d'un ensemble de tâches élémentaires. La succession des tâches est agencée selon l'hypothèse d'annotation qui évolue en fonction de l'analyse des résultats obtenus par l'exécution de chaque tâche élémentaire. Ces analyses sont réalisées par l'annotateur à l'aide de critères de décision et elles lui permettent de composer une nouvelle question biologique à laquelle répondra la tâche suivante.

Le système est donc composé des éléments suivants : tâches élémentaires, outils bioinformatiques et leurs paramètres (incluant la base de données exploitée), hypothèse d'annotation, critères d'analyse, et questions biologiques.

4.3.2 – Représentation des annotations

Chaque tâche élémentaire constitue un noeud du chemin (représenté par un rectangle). Les outils (rectangles) et bases de données (trapèzes) consultés sont reliés au noeud du chemin avec un lien non orienté (trait) et la succession chronologique des différentes étapes est représentée par un lien orienté (flèche) qui enchaîne deux actions successives (voir figures 3 et 4).

L'hypothèse d'annotation est exprimée entre accolades, au niveau du lien entre deux actions. Un point d'interrogation seul correspond à une hypothèse d'annotation « vide », rencontrée en général en début d'annotation sauf dans les cas de ré-annotation pour lesquels une hypothèse d'annotation est émise lors d'une première passe d'annotation ou lorsqu'une passe d'annotation automatique a été réalisée en amont de l'annotation manuelle.
Des critères de décisions évoqués par les annotateurs sont indiqués dans les commentaires précédant l'hypothèse d'annotation. Enfin, après l'hypothèse d'annotation, est aussi renseignée la question d'annotation qui guide le choix de la tâche suivante.
L'hypothèse d'annotation en cours, le(s) critère(s) de décisions et la question biologique sont renseignés les schématisations au niveau du lien entre deux tâches élémentaires.

Pour chaque interview, nous avons schématisé la succession chronologique des actions ou analyses effectuées par un chemin orienté. Origine et fin sont toutes deux représentées par des ovales jaunes. L'origine du chemin identifiée par le numéro d'identification de la protéine à annoter, correspond à l'accès à la fiche PAM-AGMIAL. La fin du chemin précise :

1. la classe d'annotation obtenue (choisie parmi « annotation fonctionnelle », « annotation fonctionnelle putative », « probable protein », « hypothetical protein », « pseudogene », « unknown function »),
2. dans le cas des deux premières classes d'annotation, si la fonction peut être rattachée à une classe fonctionnelle choisie dans la hiérarchie SubtiList [Moszer *et al*., 1995],
3. et les cas où plusieurs annotations (par exemple plusieurs numéros EC[6]) sont conservées par l'annotateur.

La représentation de l'extrait du recueil choisi en la figure 2 est donné en figure 4 à l'aide du codage présenté figure 3.

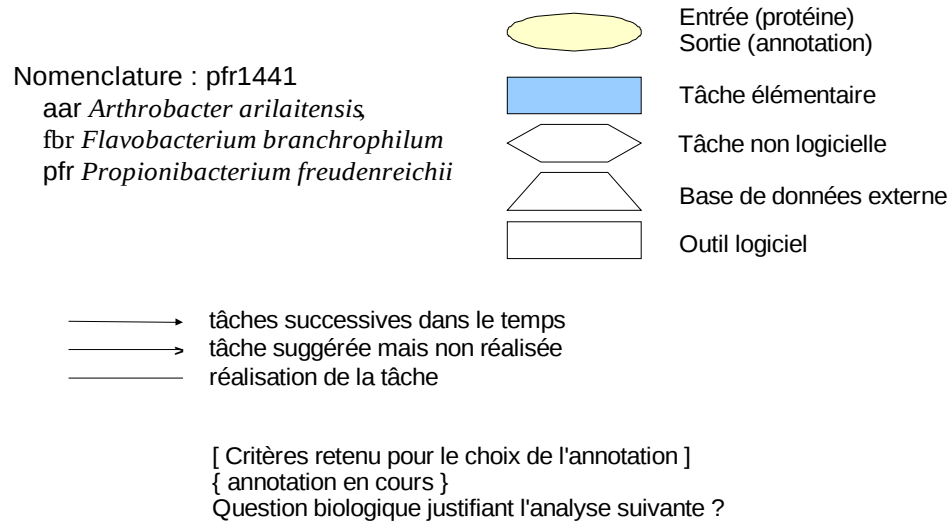

Figure 3 : Codes pour la représentation des stratégies d'annotation.

6 Numéros EC : numéros issus de l'« Enzyme Commission » qui spécifient les réactions chimiques catalysées par les enzymes

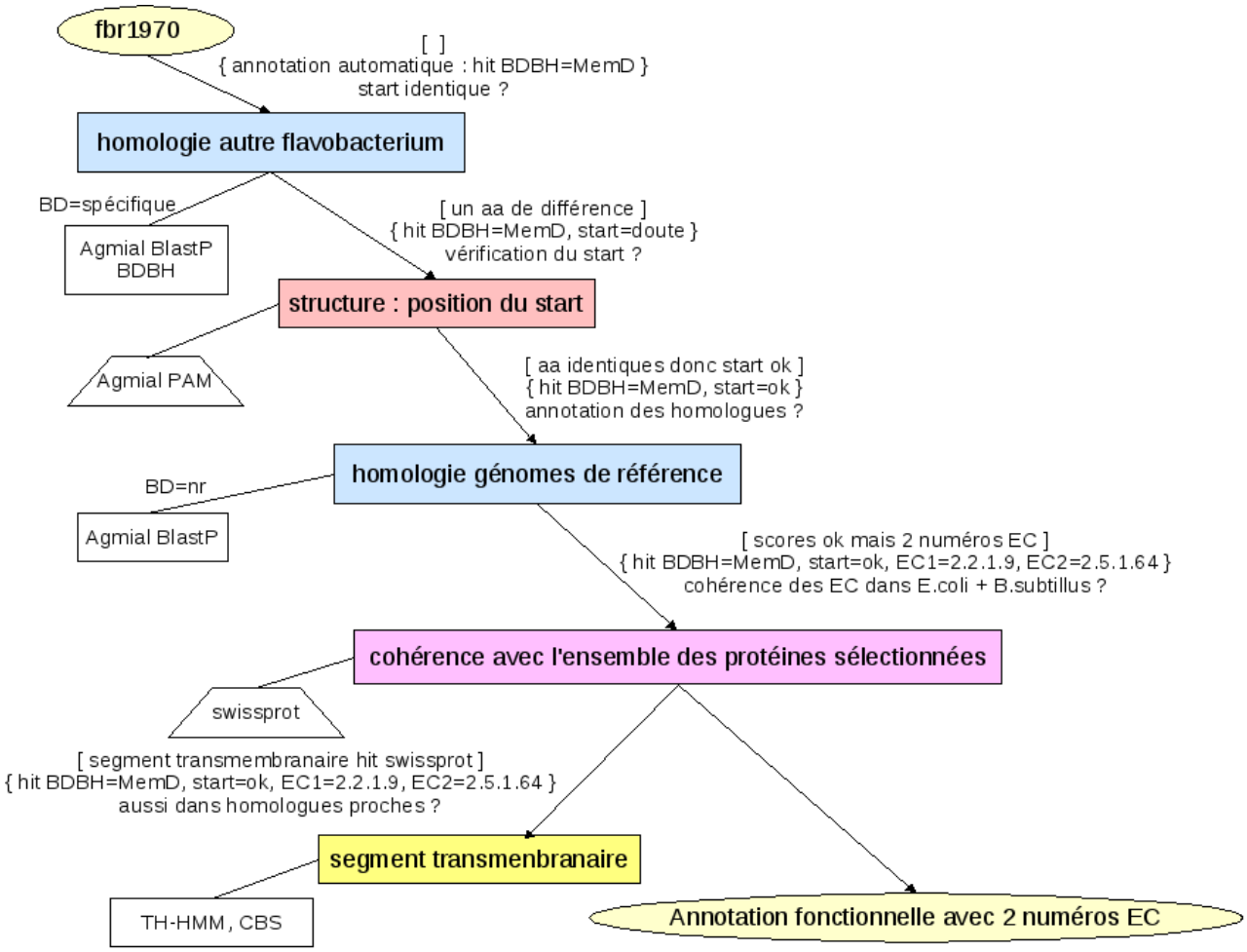

Figure 4 : Schématisation de la stratégie d'annotation pour la protéine fbr1970.
Légende : voir figure 3.

4.3.3 – Tests et critères

L'annotateur analyse les résultats d'une tâche en testant différents critères. Pour un même test évoqué dans plusieurs annotations, il est possible d'associer plusieurs tâches successives différentes et de définir plusieurs sortes de tests selon le nombre de sorties possibles. La figure 5 présente deux exemples : un test général concernant les valeurs des scores obtenus par un logiciel bioinformatique et un test plus précis qui vérifie la cohérence de la longueur des protéines d'un ensemble.

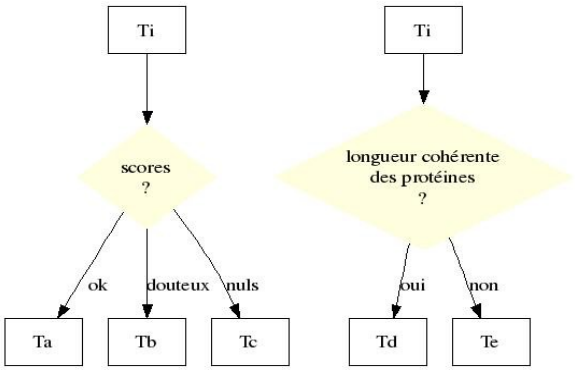

Figure 5 : Deux sortes de tests observées dans le recueil d'annotations selon le nombre de cas possibles en sortie. Ti : $i^{ème}$ tâche, Ta, Tb, Tc, Td, Te : tâches successives à Ti

La succession des tâches élémentaires est directement influencée par les décisions qui sont énoncées à chaque tâche et qui déterminent le choix de la tâche suivante. Ajouté à la mention « scores corrects », il n'y a souvent qu'un seul critère qui est évoqué par l'annotateur afin de justifier le choix de l'étape suivante. Nous avons observé que l'annotateur n'évoque un critère de décision que lorsque le résultat de l'analyse pose un problème de cohérence avec l'hypothèse d'annotation en cours selon le critère évoqué.
Nous supposons que plusieurs critères soient testés par l'annotateur mais que ce dernier ne le verbalise que lorsque la protéine en cours d'annotation présente un problème pour le critère en question. Cela implique que les relevés de critères des annotations du recueil sont incomplets.
Nous proposons alors de regrouper l'ensemble des critères de décision évoqués à partir d'une tâche présente plusieurs fois dans le recueil, afin d'établir l'ensemble des critères testés pour chaque tâche. Le matériel supplémentaire 4 présente une liste de critères établie pour les tâches de recherche de similarité. Ces tâches ont été choisies ici car nous disposions de suffisamment d'exemples pour caractériser plusieurs critères invoqués par l'annotateur.
Organiser ces critères les uns par rapport aux autres à partir d'un recueil plus complet permettrait d'aboutir à une stratégie détaillée et exhaustive de l'annotation fonctionnelle.

4.4 - Cas particulier de la tâche d'extraction d'information, tei.

Une extraction d'information explicite n'a été verbalisée que pour un seul des annotateurs. Ce dernier était un annotateur novice qui, une fois l'annotation choisie, enregistrait les informations jugées finalement pertinentes parmi celles qu'il avait analysées dans les différentes sources de données consultées en complément de la fiche AGMIAL (enregistrement par copier/coller dans un fichier texte en dehors de la plate-forme d'annotation AGMIAL). Cette tâche d'extraction d'informations rendue visible par l'action de copier/coller, n'a pas été relevée auprès des deux autres annotateurs. Nous pouvons penser que ces derniers mémorisaient les informations pertinentes tout au long de l'annotation et n'explicitaient que les quelques mots résumant la fonction à la fin de l'annotation. Dans ce cas, nous supposons que cette tâche tei est sous-jacente et répartie à chaque étape de l'annotation. La figure 6 représente deux stratégies que nous proposons pour la tâche tei.

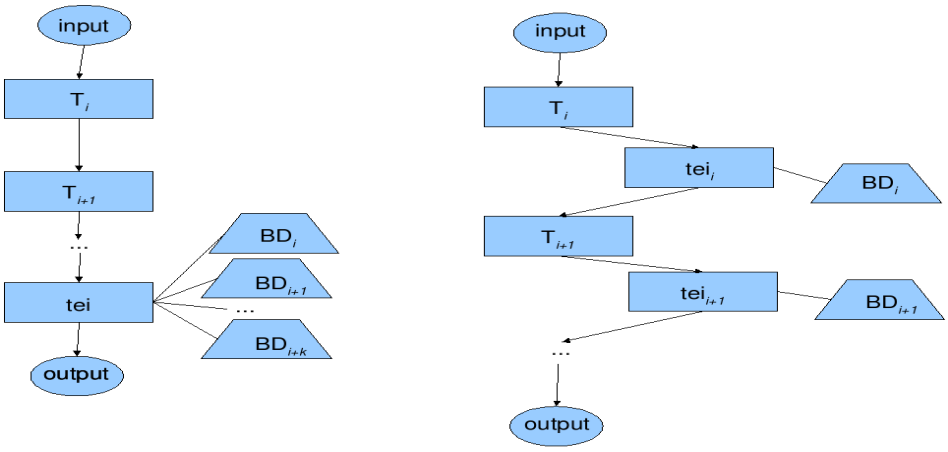

Figure 6 : cas particulier de la tâche d'extraction d'information, tei
Représentation de la situation observée pour un des annotateurs (à gauche) qui réalise une tâche tei à la fin de l'analyse comparée à la situation de mémorisation supposée au fur et à mesure du déroulement de l'analyse pour les deux autres annotateurs (à droite). T : tâche, BD : base de données, tei : tâche tei d'extraction d'information

Dans les schématisations, cette tâche tei n'est représentée que pour l'annotateur qui la verbalise.

5 – Analyse des schématisations

5.1 - Multiplicité de tâches sémantiques pour une même tâche syntaxique

Comme cela a déjà été exprimé dans [Stevens, 2000], la nature à la fois sémantique et syntaxique des tâches bioinformatiques conduit à dissocier pour des raisons sémantiques, deux tâches pourtant syntaxiquement identiques aux paramétrages de lancement près. Par exemple, identifiées dans notre recueil, les tâches thr et thp, ne répondent pas à la même question de l'annotateur : une similarité dans un génome proche non annoté (thp) vérifie que la protéine n'est pas fortuite ainsi que la validité du modèle de gène associé (annotation structurale), alors qu'une similarité dans une base de référence où les protéines sont annotées (thr) permet de proposer en sus une hypothèse d'annotation fonctionnelle. Ces tâches de recherche de similarité thr et thp sont effectuées avec le même outil (Blastp) mais sur des bases de données différentes : par exemple la section des séquences NR (Non Redondantes) du NCBI[7], et une base de données construite à la demande de l'annotateur avec les protéines des génomes proches de celui qu'il annote. La fiche PAM-AGMIAL présente deux tableaux distincts. La recherche de similarité de séquence est donc réalisée par deux étapes de navigation dans la fiche PAM-AGMIAL.

Les exemples d'utilisation des ces deux tâches élémentaires (analyse d'un résultat du logiciel Blast) permettent d'identifier des réponses à des questions différentes selon les caractéristiques de la protéine à annoter. Voici quatre questions et réponses relevées à partir du recueil d'annotations et ordonnées de telle façon que le fait de se poser la question suivante dans cette liste implique que les réponses obtenues aux questions précédentes ont été positives :

i) existence de la protéine ? réponse positive uniquement en cas de détection d'homologues, pas de réponse négative directe mais seulement inférée par l'absence de

7 NCBI : National Center for Biotechnology Information

détection de similarité (on suppose que la richesse de la banque de recherche explorée est suffisante, ce qui est le cas de la division NR du NCBI).

ii) vérification de la définition de la structure du gène ? La définition de la structure du gène est correcte si les longueurs des protéines associées par le blast sont cohérentes. La quantité des alignements du rapport blast commençant à la même position est interprétée comme la probabilité que le début proposé soit le bon

iii) un alignement pertinent permet l'identification de ce qui est conservé et sa localisation dans la séquence (position des motifs protéiques qui indiquent des pistes pour la fonction)

iv) si un alignement pertinent existe avec une protéine déjà annotée, cette annotation constitue l'hypothèse fonctionnelle et conduit la suite de la tâche d'annotation qui aura pour objectif de valider ou non cette hypothèse.

Ce type d'analyse est à mener pour chacune des autres tâches élémentaires identifiées mais il n'y a pas assez d'exemple dans le recueil pour la réaliser, en particulier lorsque les tâches n'apparaissent qu'une seule fois.

5.2 - Enchaînements temporels dans les schématisations

On constate différentes fréquences d'occurrence des tâches. Par exemple, certaines tâches apparaissent dans une seule annotation (cas de cfa, tsz, tgc, thm, thb) et d'autres dans toutes (cas de thp).

Nous nous sommes intéressés aux couples ou triplets de tâches successives que l'on pouvait identifier dans plusieurs annotations afin de rechercher un ordonnancement préféré. Nous développons ci-après deux exemples : le premier dans le cas de la tâche de recherche de similarité avec le couple thp-thr, et le second concerne l'identification de domaines ou de motifs protéiques avec les tâches tmp, tmc, et tmo.

5.2.1 - Tâches de recherche de similarité

Dans le cas de la tâche de recherche de similarité, nous avons relevé thp avant thr dans 14 cas sur 24. Les raisons de préférer un ordre sur l'autre s'expliquent selon l'annotateur (*cf.* tableau 1) ou plus précisément selon l'environnement de la tâche d'annotation.

Ici, la notion d'environnement s'entend comme l'accès aux ressources pour réaliser l'annotation et est principalement définie par la proximité ou non avec un génome évolutivement proche et déjà annoté. Ainsi, la tâche sélectionnée en premier par l'annotateur est celle qui apporte le maximum d'information et qui permet de proposer une première hypothèse pour guider l'annotation. Dans le cas de l'annotation de *F.branchrophilum*, la tâche thp (génome proche) est exploitée en premier puisqu'un génome proche et annoté, *F.psychrophilum,* est disponible. De même pour *P.freudenreichii,* où plusieurs génomes proches sont déjà annotés, la tâche thp est exploitée en premier, voire même réalisée seule. Pour *A.arilaitensis,* où le génome de référence annoté est plus lointain, le recourt à thr (recherche dans les protéines similaires identifiées les protéines appartenant à la base de données UniProtKB/Swiss-Prot) avant thp est majoritaire (12 fois sur 15).

Ces deux tâches sont souvent réalisées en premier, ce qui témoigne de l'importance de la notion de similarité dans la tâche d'annotation. De fait, elle permet de constituer l'hypothèse d'annotation.

Tableau 1: Comparatif de l'ordonnancement des tâches de recherche de similarité, thp et thr. Les chiffres indiqués donnent le nombre d'annotations où chaque ordre apparaît pour l'ensemble du recueil et pour chacun des génomes. Le nombre de fois où les deux tâches ne sont pas successives est précisé entre parenthèses.

	recueil	*A.arilaitensis*	*F.branchrophilum*	*P.freudenreichii*
thp -> thr	14 (3)	2 (1)	6 (2)	6
thr -> thp	10 (2)	10 (2)	-	-
thp seul	5	-	-	5

5.2.2 – Tâches d'identification de motifs protéiques

Le triplet de tâche tmc, tmp et tmo, qui traite des motifs protéiques identifiés dans les séquences protéiques à annoter, est identifié dans deux annotations du recueil (aar1834, aar2437) et cela ne permet pas de conclure quant à un ordonnancement particulier hormis le fait que la tâche tmo (comparaison de l'organisation des motifs entre plusieurs protéines) est par construction, la dernière tâche du triplet. En effet, il est nécessaire d'avoir identifié les motifs (tâches tmc ou tmp), avant de comparer leur présence et organisation (tmo).

Le couple tmc et tmp est observé 7 fois dans le recueil. A chaque occurrence de la tâche tmp est associée une tâche tmc, mais pas l'inverse (il y a aussi 7 tâches tmc sans tmp). Ces deux tâches fonctionnent selon le même principe : l'identification d'un motif se fait grâce à une identification dans la séquence protéique à annoter des séquences des motifs connus et répertoriés d'une base de données. Une autre façon de représenter ces deux tâches pourrait être vue comme une répétition de la tâche « recherche de motifs protéiques », la première fois au sein de la base de données prosite (tmp) et la seconde fois au sein de la base de données CDD (tmc).

5.3 - Hiérarchisation des tâches élémentaires

Nous proposons de regrouper plusieurs tâches élémentaires selon un critère donné. Ce regroupement permet de proposer des solutions pour enrichir les annotations. Par exemple si les tâches rassemblées dans une classe sont complémentaires, il peut être intéressant de proposer à l'utilisateur de les exécuter toutes si l'on peut détecter qu'il en exécute une. Dans le cas où les tâches sont redondantes, il peut alors être possible de proposer à l'utilisateur d'en remplacer une par une autre, et cela pourra être utile lors d'un d'échec de connexion au serveur exécutif. Ainsi, lorsque régulièrement, deux tâches élémentaires sont exécutées successivement ou que l'on trouve l'une avant l'autre autant de fois que l'ordre inverse, on peut considérer qu'il s'agit d'une même tâche globale qui inclue ces tâches élémentaires. Selon leur sémantique, elles peuvent cependant être exécutées sans dépendre l'une de l'autre. Ces tâches pourraient être définies comme un sous-workflow dans les représentations de Taverna.

Nous avons regroupé plusieurs tâches élémentaires sur la base de leur succession temporelle mais aussi en fonction de la sémantique de la tâche (recherche de similarité, synténie, couverture du séquençage) et du type de la donnée manipulée selon qu'il s'agisse d'une séquence protéique complète ou partielle, ou d'une séquence nucléique.

Nous avons déjà détaillé le cas des deux tâches élémentaires, thr et thp, de recherche de similarité

qui utilisent le même outil bioinformatique (blastp) mais qui diffèrent par la base de données explorée pour la recherche et les valeurs des seuils de similarité utilisées pour filtrer le résultat. La fréquente succession temporelle observée de ces deux tâches élémentaires (19 cas sur 24, cf. tableau 1) ainsi que leur syntaxe commune permettent de les regrouper en une tâche plus globale de recherche de similarité, H. On ajoute à cet ensemble la tâche de recherche de similarité dans une base de données expertisées, thb, ainsi que l'identification d'orthologues par la méthode des Bi-Directional Best Hits, thc. En considérant l'aspect sémantique, nous associons les tâches dont l'objectif est d'identifier des signaux ou des motifs dans les séquences protéiques. Il s'agit des tâches tmp, tmc, tmo, et ttm, regroupées sous la tâche M, motifs protéiques. Enfin, un autre regroupement, G, concernant les tâches de traitements liés à la séquence d'ADN à partir de laquelle la séquence protéique est issue (tss, tsg, tgc) est proposé sur la base du type de la donnée manipulée en entrée, séquence génomique ici plutôt que protéique.

Cet ensemble de critère de regroupements de tâches élémentaires en tâches plus globales constitue la liste suivante et est illustré dans la figure 7. Ces regroupements sont aussi représentés dans les schématisations des annotations du corpus au niveau des noeuds par l'utilisation d'un fond de couleurs différentes selon la tâche globale.

- recherche de similarité : H = {thp, thr, thm, thb}
- identification de motifs/signaux intrinsèques (prédiction de signaux transmembranaires, de peptides signal) : M = {tmp, tmc, tmo, ttm}
- identification d'information de synténie : S = {tsa}
- recherche des propriétés liées à la séquence génomique : G = {tss, tsg, tgc}
- extraction d'informations textuelles issues de BD spécialisées : T = {tei, thb}
- test de l'homogénéité d'un ensemble : C = {tct, tmt}
- extraction d'une sous-séquence protéique : E = {tsz}

La question de l'inclusion d'une tâche dans une autre ne se pose pas lorsqu'il s'agit de tâches élémentaires qui correspondent chacune à un lancement d'analyse. Par contre, on peut observer quelques cas d'inclusions de tâches élémentaires dans une tâche de niveau plus global ou encore une inclusion d'une tâche de niveau global dans une autre. Par exemple, lorsque les tâches de recherche de similarités ne sont pas successives (4 cas, *cf.* tableau 1), elles sont interrompues par les tâches suivantes :

- vérification du début de la protéine (fbr1970 et fbr821), où tsz et tss sont incluses dans H
- recherche de motifs (aar251 et aar258), M est incluse dans H

Le premier cas d'inclusion relevé s'explique par le fait qu'un rapport blast fournit en une seule étape différentes informations et donc répond à plusieurs questions biologiques à la fois.
Ces inclusions de tâches élémentaires au sein de tâches plus globales indiquent que ce n'est pas le flux de données qui dirige la succession des tâches ; on a proposé l'existence d'un autre flux directeur, celui de l'hypothèse d'annotation (*cf.* 4.2 – Définition d'une tâche élémentaire).

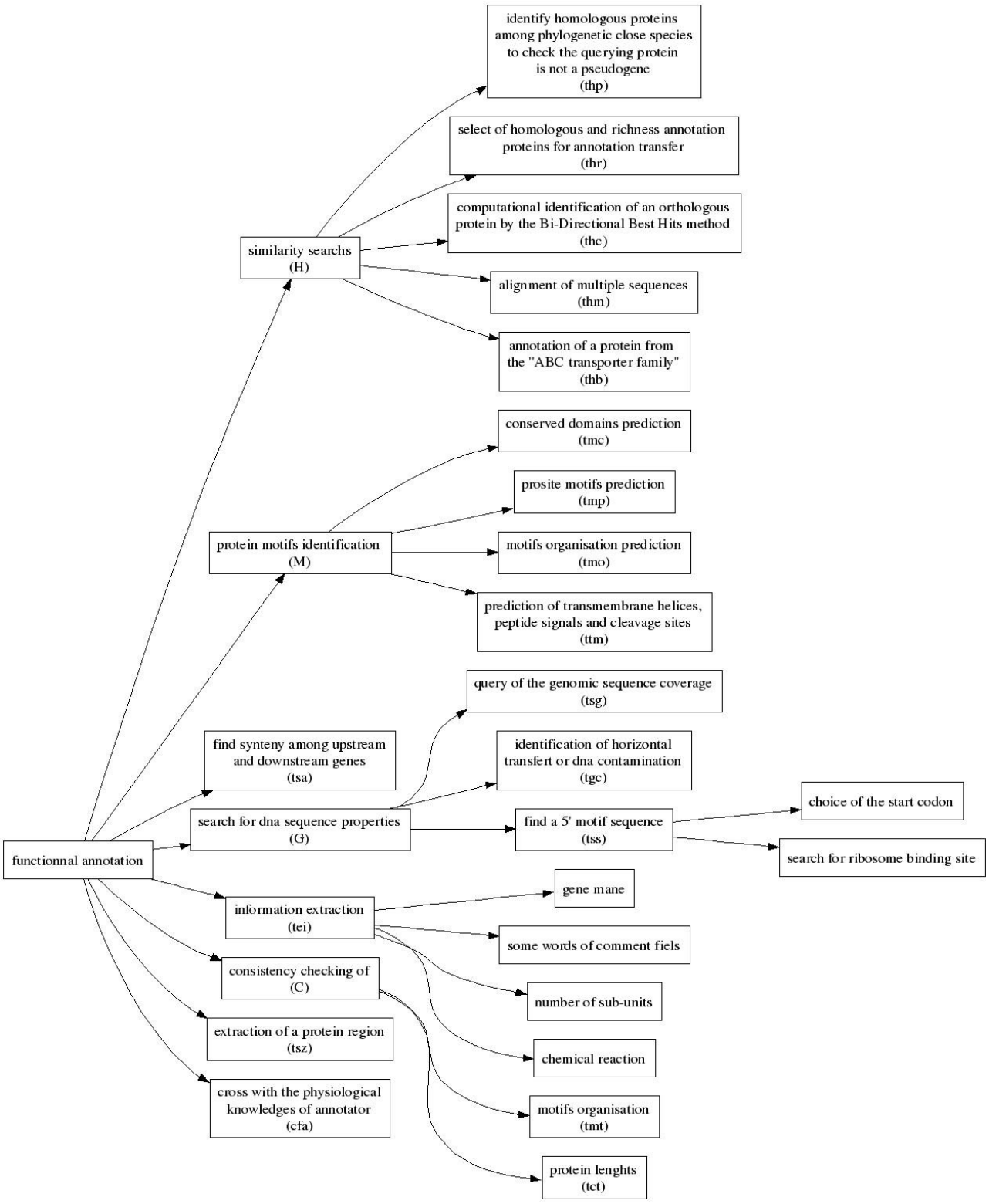

Figure 7 : Classification des 18 tâches élémentaires identifiées dans le recueil.

5.4 - Stratégie globale pour l'ensemble du recueil

On dispose de 29 instances d'exécution d'annotations fonctionnelles où l'enchaînement des tâches est chronologique. Ces dépendances temporelles intègrent à la fois les dépendances d'hypothèses (choix d'une analyse plutôt qu'une autre à cause de l'hypothèse d'annotation en cours de test) et les dépendances de données (une analyse nécessite en entrée les résultats d'une analyse précédente). Une des étapes vers la formalisation d'un workflow générique à partir des instances d'annotations recueillies consiste donc à inférer ces dépendances à partir de l'observation des annotations du recueil. Lorsque deux analyses peuvent être effectuées dans un ordre ou dans l'ordre inverse, on considérera qu'il n'y a pas de dépendance de données entre elles. Ainsi, on observe une dépendance de données avec la tâche tmt « comparaison de l'organisation des motifs » qui nécessite la tâche tmo (« établir l'organisation des motifs ») auparavant et qui, elle, nécessite tmc (« prédiction de motifs avec la base de données CDD ») et/ou tmp (« prédiction de motifs avec la base de données Prosite ») auparavant ; ou encore avec tct (« test de la cohérence de la longueur des protéines d'un ensemble ») qui nécessite H auparavant pour définir cet ensemble. Il n'y a pas assez de stratégies pour identifier d'autres dépendances de flux de données.

En excluant l'évolution de l'hypothèse d'annotation pour se rapprocher de la construction des workflows scientifiques, nous pouvons reconstruire une stratégie globale qui représente l'ensemble du recueil d'annotations (*cf.* figure 8). Dans cette reconstruction et par soucis de synthèse, nous avons utilisé les tâches globales lorsque cela était possible.

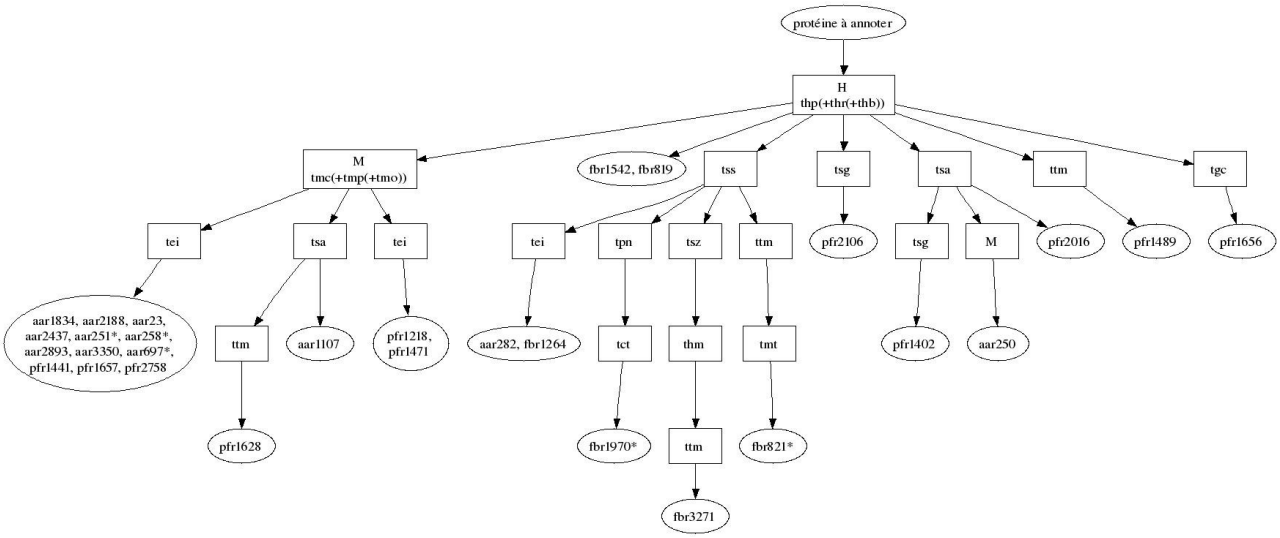

Figure 8 : reconstruction de la stratégie globale de l'ensemble des annotations du corpus. H, M : tâches globales (cf. figure 7), aar pour le génome A.arilaitensis, fbr pour F.branchrophilum et pfr pour P.freudenreichii. Les protéines marquées par un astérix indiquent qu'une inclusion d'une tâche élémentaire est survenue dans la tâche H.

Le chemin protéine->H->M(->tei)->annotation est le chemin majoritaire et concerne 12 annotations sur 29. C'est aussi l'ordre de présentation choisi dans la plate-forme AGMIAL pour présenter les informations pré-calculées : recherche de similarité par blast et de recherche de motifs. Cet ordonnancement, recherche d'un orthologue avant l'identification des motifs protéiques pertinents, correspond à la stratégie susceptible d'apporter le plus d'information le plus rapidement possible.

5.5 - Passer d'une stratégie à un workflow ?

Nous remarquons que la représentation chronologique que nous avons adoptée relie les différentes tâches entre elles mais ne précise aucune information quant à leur dépendance en termes de flux de données. Parce que les annotateurs interviewés annotent à partir de la plate-forme AGMIAL, notre approche est particulière. En effet, l'ensemble des données présentées dans la plate-forme proviennent de logiciels déjà exécutés à la consultation de la fiche PAM-AGMIAL. Dans la fiche et parfois examinés, sont présents des résultats qui n'auraient pas forcément été demandés par l'annotateur. La succession chronologique des analyses reflète alors plutôt un usage cadré par la fiche PAM-AGMIAL qu'une stratégie librement décidée pour aboutir à une annotation. Imposer d'une stratégie particulière qui soit le plus homogène possible était l'un des objectifs de la création de la plate-forme [Bryson *et al.*, 2006].

On suppose alors que chaque annotation du recueil correspond à une exécution d'un workflow générique d'annotation fonctionnelle défini par les concepteurs de la plate-forme AGMIAL et parfois enrichi par les annotateurs lorsqu'ils consultent des ressources externes. Pour aller vers une définition de ce workflow générique, plusieurs étapes ou éléments sont nécessaires :

1. Caractériser les différents éléments en jeu. Nous avons tout d'abord détaillé le type des entrées et sorties de chaque tâche élémentaire (*cf.* Matériel Supplémentaire 2). Les tâches élémentaires d'une annotation doivent ensuite être reprises de façon liée ou parallèle selon l'existence ou non d'un flux de données entre deux étapes successives.
2. Identifier un équivalent WEB service à chaque tâche élémentaire afin de pouvoir sortir du cadre de la plate-forme AGMIAL. En choisissant le formalisme proposé par l'environnement Taverna, une grande partie des étapes identifiées dans le recueil existe déjà et certaines des transformations de données nécessaires alimentent le site « myexperiment » [De Roure *et al.*, 2007]. A une tâche élémentaire telle que nous l'avons définie correspondra selon les cas un ou plusieurs WEB services. Par exemple, la tâche thp de recherche de similarité dans les génomes proches, nécessite de construire la base de données rassemblant les génomes proches avant d'exécuter la recherche.
3. Associer l'ensemble des tests et des critères de décisions qui permettent l'analyse des résultats obtenus pour chaque tâche élémentaire comme cela a été réalisé pour les tâches de recherche de similarité (*cf.* Matériel Supplémentaire 4). Ceci permettra de proposer plusieurs chemins possibles pour choisir la tâche suivante à l'aide de branchements conditionnels. Ceci aboutira à un workflow générique, qui représente l'ensemble des annotations recueillies.
4. La validation du workflow générique obtenu consistera premièrement à retrouver l'ensemble des annotations du recueil, à comparer les annotations obtenues par rapport à celles qui ont été recensées dans le recueil et enfin, à comparer les résultats obtenus avec l'annotation de nouvelles protéines à la fois par la plate-forme d'annotation mais aussi par le workflow.

Ainsi, avec les éléments déjà en place et ceux que nous avons identifiés ci-dessus, il devrait être possible de construire un workflow générique de l'annotation fonctionnelle de génomes bactériens sur le modèle de la stratégie globale présentée figure 8.

6 - Discussions, conclusions et perspectives

6.1 - Bilan

Nous avons défini la tâche l'annotation fonctionnelle comme une tâche d'intégration des nombreuses données et analyses que les annotateurs enchaînent afin de caractériser une séquence. Nous avons proposé une démarche pour construire une stratégie d'annotation. Nous avons recueilli un corpus de 29 annotations. Ces annotations ont été schématisées en identifiant les tâches élémentaires et leur succession temporelle. Nous avons identifié que la succession temporelle de ces différentes tâches est guidée par l'hypothèse d'annotation qui se construit et s'enrichit des analyses effectuées à chaque étape par l'annotateur. A l'aide de différents critères de regroupement, nous avons ensuite hiérarchisé ces tâches élémentaires afin d'identifier des tâches globales et communes à plusieurs annotations.

Cette démarche de construction de stratégie d'annotation pourra être appliquée à d'autres recueils qui ne concerneraient pas seulement des annotations fonctionnelles de protéines bactériennes à l'aide de la plate-forme AGMIAL.

L'hypothèse de travail sous-jacente à l'annotation s'appuyant sur une recherche de similarité entre séquences protéiques, nous avons montré qu'effectivement, nous retrouvons cette tâche dans l'ensemble des annotations du corpus et aussi, qu'elle constitue le point de départ de l'hypothèse fonctionnelle qui guide les analyses pendant l'annotation.

Plusieurs propositions pour l'amélioration de l'annotation fonctionnelles émanent de cette étude et sont exposées dans les paragraphes suivants.

6.2 – Annoter en construisant un ensemble de références

Afin de composer une annotation, l'analyse effectuée par un annotateur se décline en plusieurs étapes que l'on retrouve indépendamment du niveau de détail concerné : protéine, domaine protéique, signal ou positions de début et de fin de traduction (*cf.* tableau 2). En prenant exemple au niveau protéique, voici les trois étapes identifiées :

1. à partir de la protéine à annoter, construction d'un ensemble de protéines qui servira de référence. Cette construction est réalisée par une recherche de similarité dans une banque de protéines. La recherche de similarité est en général réalisée par le logiciel blastp. Les paramètres qui influent cette construction sont d'une part, le seuil utilisé pour borner l'ensemble de référence et d'autre part, le choix de la base de données de recherche qui est souvent une base de données générale. Le recueil d'annotations montre une préférence pour la base UniProt/SwissProt dont la richesse et la fiabilité sont reconnues par les annotateurs. La valeur de score d'alignement à partir de laquelle les protéines sont jugées pour soit faire partie de l'ensemble de référence, soit être rejetées, varie pour chaque protéine et chaque annotateur.
2. vérification de la cohérence des protéines de l'ensemble de référence. Si le critère de similarité des séquences protéiques permet de construire l'ensemble de référence, sa cohérence est évaluée à partir des paramètres intrinsèques des protéines (longueur, domaines, ...) et la cohérence des annotations (fonction, classe fonctionnelle -mots-clefs SwissProt, termes GO, hiérarchie de classes fonctionnelles Funcat ou SubtiList, etc ...-, et dans le cas d'une enzyme, partage de numéros EC). Quand l'une des protéines présente une

différence pour un paramètre, une ré-annotation de cette protéine ou une recherche de cette valeur différente dans les autres protéines de l'ensemble est effectuée pour estimer s'il s'agit d'une information complémentaire pertinente ou d'une divergence. Selon le résultat, la protéine divergente est conservée ou exclue de l'ensemble de référence.

3. vérification de la cohérence des paramètres de la protéine à annoter par comparaison avec ceux des protéines de l'ensemble de référence, les paramètres intrinsèques et estimés doivent être de même ordre de grandeur.

Lorsque les caractéristiques de la protéine sont cohérentes avec celles de l'ensemble de protéines de référence et lorsque les annotations des protéines de l'ensemble sont homogènes entre elles, alors l'annotation consiste simplement en un transfert de l'annotation commune à l'ensemble des protéines sélectionnées.
Lorsque qu'aucun ensemble de protéines ne peut être défini du fait de trop faibles similarités de séquence, seuls les paramètres intrinsèques et estimés constituent l'annotation de la protéine. Et lorsque les paramètres estimés n'apparaissent pas suffisamment établis alors la validité même de la protéine est mise en doute (« hypothetical protein »).

Tableau 2: Critères utilisés lors de l'annotation à l'aide d'un ensemble de référence lors de sa définition et de son utilisation (cohérence des références et appartenance de l'objet en cours d'annotation) pour 4 objets différents : protéines, domaines/motifs protéiques, signaux, ORF.

Critères pour	Protéines	Domaines - motifs	Signaux	ORF
définir l'ensemble de référence (logiciels - BD)	Seuil de score (Blastp sur NCBI-NR ou Swissprot)	Seuil de score de recherche de motifs connus (CDD, prosite)	Prédictions : présence, nombre et score (TM-HMM)	Position partagée de début et fin de l'ORF
décider de la cohérence / appartenance	Partage des annotations	Présence et ordre des motifs	Prédictions partagées	Positions et acides aminés partagés

Cette méthode peut être vue comme la première étape du mécanisme de transfert d'annotations décrit en deux étapes par [Levy *et al*., 2005] : i) établir une liste des protéines connues possédant une similarité suffisante avec la protéine à annoter et ii) sélectionner la ou les séquences connues à partir desquelles transférer l'annotation.

La définition d'un ensemble de référence pour chacune des étapes de l'annotation permet à l'annotateur de disposer d'un outil adapté à chaque protéine pour décider de l'annotation.

6.3 – Seuils et critères de décision

Le facteur seuil régulièrement évoqué lors de l'annotation concerne les seuils de décision qu'utilise l'annotateur pour extraire et ne conserver que la partie jugée pertinente de l'ensemble des informations obtenues lors d'une analyse. Ces seuils interviennent à chaque fois que l'information en jeu n'est pas déterministe : lors de l'estimation de la présence de motifs ou signaux dans la séquence, ou encore lors de la définition de l'ensemble des protéines de référence. Lors des annotations manuelles, ces valeurs de seuils sont définies pour chaque cas, en fonction de l'environnement. On entend ici par le terme « environnement » les informations réunies lors d'une étape, ajoutées aux connaissances de l'annotateur et filtrées par l'hypothèse fonctionnelle en cours

de test par l'annotateur. Lors des programmes d'annotation automatique, des seuils existent aussi pour séparer les informations en deux parties mais ils sont, en général, fixés à l'identique pour l'ensemble des protéines à annoter. L'usage d'une valeur figée dans le cas des annotations automatiques contre celui d'une valeur variable dans le cas de l'annotation manuelle pourrait expliquer une partie des différences observées entre annotations manuelles et automatiques.
Ainsi, une amélioration possible des logiciels de prédiction fonctionnelle pourrait résider dans la définition de valeurs de seuil variables, choisis pour chaque protéine en fonction de l'analyse du résultat d'un logiciel de prédiction (blast pr exemple). Une façon de choisir un seuil variable est de s'appuyer sur un saut des scores ou des *expected values*, parfois observable dans les résultats présentés. Cette solution basée sur un saut de score est utilisée dans [Raes *et al.*, 2003 ; Deveaux *et al.*, 2008], et est pertinente du point de vue biologique.

6.4 – Swissprot : une base de données de confiance

L'utilisation de la base de données Swissprot comme une base de données de référence est confirmée par le recueil puisqu'une des étapes quasiment systématique concerne une recherche de similarité dans cette base de données. La confiance dans la qualité de cette base de données est un fait avéré. Cette remarque, dans le cadre de la définition de workflows d'annotation fonctionnelle, incite à proposer une méthode pour créer automatiquement un workflow ou guider les choix en cours d'exécution d'un workflow générique qui présenterait plusieurs chemins d'annotations possibles selon la protéine à annoter. A partir de l'identification de l'homologue Swissprot le plus proche, on pourrait utiliser la fiche de description de l'homologue pour borner la liste des analyses à mener aux informations rassemblées. Ainsi, pour chaque prédiction correspondant à un résultat jugé pertinent puisque répertorié dans la fiche de l'homologue Swissprot, il faudrait lancer la même prédiction sur la protéine à annoter. L'algorithme que nous proposons est présenté figure 8.

```
pour chaque protéine {
                    rechercher un homologue Swissprot
                    si un homologue existe {
                           construire dynamiquement un workflow :
                               réaliser sur la protéine à annoter les tâches
                                associées à chaque caractéristique décrite
                                pour l'homologue
                            vérifier la cohérence des résultats pour les 2
                             protéines
                            si cohérence, transférer l'annotation Swissprot
                    }
}
```

Figure 8 : Algorithme pour utiliser les descriptions des fiches Swissprot pour construire un workflow d'annotation dynamique et adapté à chaque protéine.

Cette solution présente au moins deux avantages :

i) ne pas proposer à l'annotateur (et bientôt aux systèmes automatiques lorsqu'il leur faudra passer à l'échelle du séqueçage à très haut débit) trop d'information ou de calculs inutiles (par exemple, il ne sert à rien de chercher à identifier une séquence peptide-signal dans une protéine qui n'en a pas parce que non exportée),

ii) si les résultats obtenus sur la protéine à annoter concordent avec les descriptions de l'homologue identifié dans la base de données Swissprot, cela permet de proposer le transfert d'annotation à partir de critères plus riches que le seul degré de similarité habituellement utilisé. En effet, avec cette méthode, le transfert d'annotation prendra alors en compte l'ensemble des informations répertoriées dans la description Swissprot de l'homologue.

Les inconvénients de cette méthode sont de ne pas fonctionner quand la distance entre les deux protéines est trop grande (il n'y a pas d'ortologue) et de ne pas permettre la découverte de nouvelles fonctionnalités. Ce dernier inconvénient est cependant valable aussi dans le cas de l'annotation manuelle réalisée dans le cadre d'une plate-forme si les annotateurs n'utilisent pas des informations complémentaires et récupérées à l'extérieur de la plate-forme.

6.5 - Noter la qualité de l'annotation ?

6.5.1 – Suggestion d'étapes supplémentaires

Plusieurs fois, les annotateurs ont suggéré une étape supplémentaire pour aller plus loin dans l'analyse (fbr1542-tsg, fbr1970-ttm, fbr821-ttm, pfr1402-tsg). Pour ces 4 cas, on peut considérer que les annotations recueillies ne sont donc pas de qualité maximale puisqu'une analyse supplémentaire aurait été nécessaire pour conforter l'hypothèse fonctionnelle choisie. La raison principale évoquée expliquant pourquoi les annotateurs n'ont pas réalisé ces dernières analyses concerne le rapport qualité/temps. En effet, une analyse supplémentaire représente un coût temporel pour un gain en qualité pas forcément assuré.
En particulier et selon l'expérience des annotateurs, la demande de couverture du séquençage (tsg) au génoscope et déjà réalisée dans d'autres cas n'a que rarement apportée une information ayant changé le résultat de l'analyse en cours. De plus la réponse à cette analyse demande du temps puisqu'il faut interagir avec le centre de séquençage, obligeant l'annotateur à mettre de côté l'annotation en cours pour la reprendre après.
De même, vérifier la cohérence concernant la présence d'un segment transmembranaire dans plusieurs orthologues (ttm) nécessite une analyse par orthologue et ne donne pas toujours le résultat attendu, ces logiciels de prédictions n'étant pas encore suffisamment fiables selon les espèces analysées.

6.5.2 – Évolution continue des annotations

Une autre remarque à propos des annotations fonctionnelles concerne leur validité dans le temps. En effet, les annotations sont valables au moment où elles sont posées, mais ensuite, du fait de l'évolution du domaines par l'acquisition de nouvelles connaissances comme de nouvelles données, les annotations posées peuvent devenir incomplètes voir incorrectes. De fait, le séquençage engagé par les différents centres de séquençage engendre une arrivée quasi-continue de nouvelles séquences et la réduction du nombre d'« orphan », protéine sans homologue identifié, illustre cette évolution par acquisition de nouvelles séquences ou de fouille de données efficaces [Lespinet, Labedan, 2006]. Or, ces corrections ne sont pas toujours suivies, ni même répertoriées.

6.5.3 – Qualité des annotations ?

Les deux points précédents montrent d'une part que toutes les annotations manuelles ne se valent pas en terme de qualité par rapport aux informations existantes et d'autre part, que l'évolution des annotations n'est pas idéalement prise en compte.
Dans le cas des annotations manuelles et pour ajouter un élément à leur qualité, si un haut crédit est accordé à Swissprot, il est à comparer avec d'autres projets d'annotation menés par de petites équipes d'annotateur, chacune annotant son génome. Dans ces équipes, l'annotation de l'ensemble des protéines d'un génome peut représenter une quantité trop importante, s'étendre sur une durée trop longue et où la qualité de l'annotation obtenue variera selon l'intérêt des annotateurs pour chaque protéine.

Une fois ce constat posé que la qualité des annotations n'est pas maximale, la question qui a ensuite guidée cette étude est la suivante : comment juger de la qualité d'une annotation ? Comment établir une mesure de qualité et quels critères prendre en compte ?

Effectivement, l'annotation est par essence une hypothèse, et comme toute hypothèse, une indication de son degré de confiance pourrait lui être associée. A partir de notre étude des stratégies d'annotation, nous avons donc cherché à identifier les critères qu'une note pourrait prendre en compte afin de refléter la confiance de l'annotateur vis à vis de son annotation.

6.5.4 – Dissocier qualité de l'annotation et qualité du processus

On peut rencontrer deux cas d'absence d'annotation associée à une protéine : soit aucune annotation n'a été trouvée, en particulier lorsque la séquence protéique ne présente pas assez de similarité avec une autre protéine ou qu'aucune des protéines homologues ne possède d'annotation fonctionnelle, soit, plus simplement, le processus d'annotation n'a pas encore été effectué sur cette séquence qui se trouve alors « nue » dans les bases de données. Dans le premier cas, le processus d'annotation est correct et aucune information n'a été trouvée.
Cette remarque nous permet de dissocier deux concepts reliés à la qualité : d'une part la qualité de l'annotation posée et d'autre part, la qualité du processus d'annotation utilisé. En effet, ce n'est pas parce que l'annotation, l'hypothèse résultante, est mauvaise au sens biologique (car elle ne possède que peu d'information) qu'elle a été posée avec une mauvaise qualité du processus d'annotation. L'annotation est une décision finale qui ne reflète ni les étapes (prédiction, recherche, analyse) ni les seuils de décisions qui ont été utilisés.

Avec les objectifs d'automatisation de l'annotation qui ont initiés cette étude, on a alors la possibilité d'ajouter des informations contextuelles telles que la date, les versions de bases de données, les seuils de décisions, *etc*... à la réalisation de l'annotation et donc la possibilité de juger de la qualité du processus d'annotation suivi.

6.5.5 – Facteurs influençant la qualité de l'annotation

Ainsi, d'après notre étude des stratégie d'annotation, nous pouvons proposer plusieurs facteurs à prendre en compte, à la fois pour juger de qualité de l'annotation (résultat), mais aussi pour estimer la qualité du processus mis en oeuvre.
La qualité de l'annotation est influencée au moins par ces 3 points soulevés dans notre étude :

1. la distance évolutive avec un génome déjà annoté, principalement du fait que les annotations des génomes proches constituent le point de départ d'une nouvelle annotation (revoir le cas de *F.branchrophilum* qui est très proche de *F.psychrophilum*)
2. la confiance dans les annotations antérieures et le haut crédit relevé pour Swissprot. On pourra revenir ici au cas de l'annotation de *P.freudenreichii* pour laquelle autre bactérie séquencée et annotée évolutivement proche existe, mais dont les annotations ne conviennent pas.
3. le partage de la niche écologique entre l'organisme de référence et l'organisme à annoter. On rappelle le cas d'*A.arilaitensis* pour lequel deux autres génomes proches sont annotés mais qui vivent au sein d'une niche écologique différente (fromage - sol).

Du coup, l'usage d'annotations de génomes plus éloignés comme référence implique forcément une moindre qualité des annotations obtenues.

Nous proposons d'augmenter la qualité du processus en l'enrichissant des informations contextuelles, nommées parfois « provenance » [Cohen-Boulakia et al., 2007]. Cela concernerait les

éléments suivants :

- la date de la réalisation de l'annotation
- les versions des logiciels et des bases de données utilisées
- une pondération de « confiance » dans ressources utilisées pourrait être ajustée et pourrait ainsi refléter la confiance et l'expertise des annotateurs
- les seuils de décisions et les scores des résultats

Le nombre d'analyse pour obtenir une annotation, plutôt que d'être associé à la qualité du processus d'annotation indiquerait plutôt une moindre qualité du résultat. En effet, beaucoup d'analyses de prédictions indiquent en général des scores moyens, ne permettant pas de trancher clairement l'hypothèse d'annotation testée. De même, si l'analyse de la première tâche confirme de façon très claire de l'hypothèse d'annotation, lancer d'autres analyses n'est pas nécessaire. Dans ce cas, le critère « nombre d'analyses » n'indique pas forcément un processus d'annotation de mauvaise qualité. Ainsi, sans forcément associer le nombre d'étape à la qualité du processus, ce nombre d'étape donne une indication de la qualité du résultat. De même de la richesse d'une annotation, que l'on pourrait mesurer par le nombre de mot ou la profondeur obtenue dans une hiérarchie de termes, ne reflète pas un processus d'annotation de mauvaise qualité.

Nous voyons ici la difficulté pour qu'une pondération de ces différents éléments, sous forme d'une « note » puisse résumer pour l'annotateur la qualité de son annotation, au sens résultat et processus suivi. En perspectives, cette étude ouvre donc vers la définition d'une telle pondération.

Ajouter une note reflétant la qualité de l'annotation permettrait aux biologistes d'estimer le risque avec lequel ils pourront les utiliser et ainsi leur éviter de refaire par défaut l'annotation comme c'est le cas actuellement. Il existe déjà des annotations qui sont réalisées avec un système de notes. C'est par exemple, le cas des notes du projet GeneFarm [Aubourd *et al.*, 2005] qui reflètent la quantité et la qualité des preuves expérimentales sur lesquelles s'appuient l'annotation proposée. Ces notes sont différentes d'une note de qualité de la tâche d'annotation en elle-même qui a été effectuée. Nous concluons qu'il est pertinent d'associer aux annotations une indication de la qualité de la démarche d'annotation et aussi de l'annotation elle-même.

7 - Références

Altintas I, Berkley C, Jaeger E, Jones M, Ludäscher B, Mock S : Kepler: An Extensible System for Design and Execution of Scientific Workflows. system demonstration, 16th Intl. Conf. on Scientific and Statistical Database Management (SSDBM'04), 21-23 June 2004, Santorini Island, Greece

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ : Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10.

Aubourg S, Brunaud V, Bruyère C, Cock M, Cooke R, Cottet A, Couloux A, Déhais P, Deléage G, Duclert A, Echeverria M, Eschbach A, Falconet D, Filippi G, Gaspin C, Geourjon C, Grienenberger JM, Houlné G, Jamet E, Lechauve F, Leleu O, Leroy P, Mache R, Meyer C, Nedjari H, Negrutiu I, Orsini V, Peyretaillade E, Pommier C, Raes J, Risler JL, Rivière S, Rombauts S, Rouzé P, Schneider M, Schwob P, Small I, Soumayet-Kampetenga G, Stankovski D, Toffano C, Tognolli M, Caboche M, Lecharny A : GeneFarm, structural and functional annotation of Arabidopsis gene and protein families by a network of experts. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D641-6

Audit B, Levy ED, Gilks WR, Goldovsky L, Ouzounis CA : CORRIE: enzyme sequence annotation with

confidence estimates. BMC Bioinformatics. 2007 May 22;8 Suppl 4:S3

Bastien O, Ortet P, Roy S, Maréchal E : A configuration space of homologous proteins conserving mutual information and allowing a phylogeny inference based on pair-wise Z-score probabilities. BMC Bioinformatics. 2005 Mar 10;6:49

Brenner SE: Errors in genome annotation.Trends Genet. 1999 Apr;15(4):132-3

Bryson K, Loux V, Bossy R, Nicolas P, Chaillou S, van de Guchte M, Penaud S, Maguin E, Hoebeke M, Bessières P, Gibrat J-F : AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system. Nucleic Acids Res. 2006; 34(12): 3533–3545

Cohen-Boulakia S, Biton O, Cohen S, Davidson SB : Addressing the Provenance Challenge using ZOOM. In Concurrency and Computation: Practice and Experience, Wiley InterScience, 2007.

De Roure D, Goble C, Stevens R : Designing the myExperiment Virtual Research Environment for the Social Sharing of Workflows. e-Science 2007 - Third IEEE International Conference on e-Science and Grid Computing, 2007. Bangalore, India, 10-13 December 2007. Pages 603-610

Deveaux Y, Toffano-Nioche C, Claisse G, Thareau V, Morin H, Laufs P, Moreau H, Kreis M, Lecharny A : Genes of the most conserved WOX clade in plants affect root and flower development in Arabidopsis. BMC Evol Biol. 2008 Oct 24;8:291.

Devos D, Valencia A: Intrinsic errors in genome annotation.Trends Genet. 2001 Aug;17(8):429-31

Duchaud E, Boussaha M, Loux V, Bernardet JF, Michel C, Kerouault B, Mondot S, Nicolas P, Bossy R, Caron C, Bessières P, Gibrat JF, Claverol S, Dumetz F, Le Hénaff M, Benmansour A : Complete genome sequence of the fish pathogen Flavobacterium psychrophilum. Nat Biotechnol. 2007 Jul;25(7):763-9

Gilks WR, Audit B, De Angelis D, Tsoka S, Ouzounis CA : Modeling the percolation of annotation errors in a database of protein sequences. Bioinformatics. 2002 Dec;18(12):1641-9

Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P, Oinn T : Taverna: a tool for building and running workflows of services. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W729-32

Lespinet O, Labedan B : ORENZA: a web resource for studying ORphan ENZyme activities. BMC Bioinformatics. 2006 Oct 6;7:436.

Lister R, Gregory BD, Ecker JR : Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. Curr Opin Plant Biol. 2009 Jan 19

Ludäscher B, Altintas I, Berkley C, Higgins D, Jaeger-Frank E, Jones M, Lee E, Tao J, Zhao Y : Scientific Workflow Management and the Kepler System. Concurrency and Computation: Practice & Experience, 18(10), pp. 1039-1065, 2006

Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH : CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res. 2002 Jan 1;30(1):281-3.

Mons B, Ashburner M, Chichester C, van Mulligen E, Weeber M, den Dunnen J, van Ommen GJ, Musen M, Cockerill M, Hermjakob H, Mons A, Packer A, Pacheco R, Lewis S, Berkeley A, Melton W, Barris N, Wales J, Meijssen G, Moeller E, Roes PJ, Borner K, Bairoch A : Calling on a million minds for community annotation in WikiProteins. Genome Biol. 2008;9(5):R89

Moszer I, Glaser P, Danchin A : SubtiList: a relational database for the Bacillus subtilis genome. Microbiology. 1995 Feb;141 (Pt 2):261-8

Mulder N, Apweiler R : InterPro and InterProScan: tools for protein sequence classification and comparison. Methods Mol Biol. 2007;396:59-70

Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A, Li P : Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics. 2004 Nov 22;20(17):3045-54

Raes J, Rohde A, Christensen JH, Van de Peer Y, Boerjan W : Genome-wide characterization of the lignification toolbox in Arabidopsis. Plant Physiol. 2003 Nov;133(3):1051-71.

Romano P, Marra D, Milanesi L : Web services and workflow management for biological resources. BMC Bioinformatics. 2005 Dec 1;6 Suppl 4:S24.

Shah SP, He DY, Sawkins JN, Druce JC, Quon G, Lett D, Zheng GX, Xu T, Ouellette BF : Pegasys: software for executing and integrating analyses of biological sequences. BMC Bioinformatics. 2004 Apr 19;5:40

Shendure J, Ji H :Next-generation DNA sequencing. Nat Biotechnol. 2008 Oct;26(10):1135-45

Stevens R, Goble C, Baker P, Brass A: A classification of tasks in bioinformatics. Bioinformatics. 2001; 17 (2): 180-188

Stevens RD, Robinson AJ, Goble CA : myGrid: personalised bioinformatics on the information grid. Bioinformatics. 2003;19 Suppl 1:i302-4

The UniProt Consortium : The Universal Protein Resource (UniProt). Nucleic Acids Res. 2008 January; 36(Database issue): D190–D195

Wieser D, Kretschmann E, Apweiler R: Filtering erroneous protein annotation. Bioinformatics. 2004 Aug 4;20 Suppl 1:i342-7

8 - Matériel supplémentaire

MS 1 : Données accessibles lors de la consultation d'une fiche PAM-AGMIAL

Paramètres intrinsèques à la protéine

Son point isoélectrique, sa longueur en acides aminés, …

Résultats de logiciels bioinformatiques

Les résultats des logiciels bioinformatiques qui ont été lancés avant que l'annotateur accède à la fiche sont présentés sous la forme de tableaux de synthèse et restreints aux dix premiers meilleurs résultats (un lien hypertexte permet aussi de visualiser plus de résultats). Chaque résultat est enrichi de liens hypertextes pour accéder à plus de détails si besoin. Il y a un tableau pour chaque exécution de logiciel : recherche de similarité par alignement de séquences, recherche dans une base de motifs, ou prédiction de signal.

Les trois premiers tableaux concernent les recherches de similarités de la protéine à annoter avec le logiciel d'alignement de séquences Blast (programme blastp). Pour chacune des trois exécutions, la base de données exploitée est différente :

- alignement contre l'ensemble des protéines des protéines du génome à annoter
- alignement contre les protéines de génomes sélectionnés par l'annotateur,
- alignement contre les protéines d'une base de données indépendante, NCBI-division NR

Les tableaux suivants indiquent l'identification de motifs ou de domaines protéiques en résultat d'« interproScan » [Mulder and Apweiler, 2007] ou de « conserved domain search » [Marchler-Bauer *et al*., 2002]. La prédiction de signaux est ajoutée à ce dernier tableau de résultats.

MS 2 : Description des tâches élémentaires identifiées

La description des tâches élémentaires identifiées dans le recueil est répartie dans les deux tableaux suivants pour des questions de lisibilité. Pour chaque tâche, on trouve dans le premier tableau (MS2-1) : l'identifiant de la tâche, sa description et sa classe selon la classification des tâches bioinformatiques proposée par [Stevens et al., 2001, les données qu'elle exploite en entrée (colonne input), et les résultats qu'elle fournie (colonne output). Le deuxième tableau (MS2-2) donne pour chaque tâche élémentaire, les outils et base de données (ainsi que la localisation du serveur entre parenthèses), le critère de décision et la forme de la partition obtenue, et la liste des annotations du corpus concernées.

Table MS2-1: tâches élémentaires

Task	Description *{classification from [Stevens et al., 2001]}*	input	output
thp	Homologous proteins search among phylogenetic close species, checking the querying protein is not a pseudogene	Protein sequence	Alignment, alignment scores, list of homologous

	{Sequence similarity searching, protein vs. protein}		proteins
thr	Homologous and richness annotation proteins search in order to transferring annotation *{Sequence similarity searching, protein vs. protein}*	Protein sequence	Alignment, aligment scores, list of homologous proteins
tmc tmp	Protein motif identification *{Functional motif searching}*	Protein sequence	List of identified motifs
tmo tmt	Check of the conservation of the motifs organisation *{Functional motif searching}*	Protein sequence	Linear graphical representation of identified motifs
tsa	Find synteny among upstream and downstream genes *{nd}*	Extended genomic sequence	Linear graphical representation of annotated genes
tss	Find of a 5' motif sequence : choice of the start codon, or search for ribosome binding site in the genetic context *{Other DNA analysis}*	Genomic sequence	Colored xTG and RBS sequence motifs
tei	Information extraction : 1. gene name 2. words from comment fields 3. number of sub-units 4. chemical reaction *{Litterature searching}*	A data base query result	Words for functionnal description
thp / thr	Computational Identification of an orthologous protein by the Bi-Directional Best Hits method *{Sequence similarity searching, protein vs. protein}*	Protein sequence	One protein identifier
ttm	Prediction of transmembrane helices, peptide signals and cleavage sites *{Functional motif searching}*	Protein sequence	List of prediction
tct	Check of the consistency of the protein length *{Protein analysis}*	begin and end positions of the alignment	Yes or length in aa of the count of the differencies (Nterm or Cterm region)
tsg	Query of the genome coverage of the genomic sequence *{Sequence assembly}*	Position on genomic sequence	Alignment of all the genomic sequences
tsz	Extraction of a protein region *{Protein analysis - Sequence retrieval}*	Protein sequence	A region of the protein sequence
thm	Multiple alignment *{Multiple sequence alignment}*	List of protein sequences	alignment
thb	Annotation of a protein from the “ABC transporter family” *{Litterature searching}*	Protein sequence	Yes or no

tgc	Identification of horizontal transfert or contamination {Other DNA analysis}	Genomic sequence	Horizontal transfert, contamination or not defined
cfa	Cross with the physiological knowledges of annotator	-	-

Table MS2-2: Paramètres des tâches élémentaires

Task	Tools / DB (server)	Parameters	Decision criteria	Annotations
thp	Blast alignment (AGMIAL)	Blastp on specific or NR database	Threshold on expected value partition: one, many or no solution	All annotations
thr	Blast alignment (AGMIAL)	Blastp on specific or NR database	threshold on expected value, provenance from the Swissprot database partition: one, many or no solution	aar1107, aar1834, aar2188, aar23, aar2437, aar250, aar251, aar258, aar282, aar2893, aar3350, aar697, fbr1264, fbr1542, fbr1970, fbr3271, fbr819, fbr821, pfr1402, pfr1489, pfr1656, pfr2016, pfr2638, pfr2758
tmc tmp	CDD and Interpro (AGMIAL)	default	Threshold on score partition: one, many or no solution	aar1107, aar1834, aar2188, aar23, aar2437, aar250, aar251, aar258, aar2893, aar3350, aar697, pfr1441, pfr1657, pfr2638, pfr2758
				aar1107, aar1834, aar2188, aar2437, aar250, aar251, aar3350
tmo tmt	Scanprosite (EBI)	default	From graphic visualisation partition: ordered list of common motifs or not	aar1834, aar23, aar2437, aar258
				fbr3271, fbr821
tsa	GenomeProtMap (NCBI)	default	Synteny on at least two successive genes partition: yes or no	aar1107, aar250, pfr1402, pfr2016, pfr2638
tss	CAM (AGMIAL)	default	From graphic visualisation partition: motif identification or not	aar1107, aar282, fbr1264, fbr1970, fbr3271, fbr821
tei 5.	1. Swissprot (SIB) 2. Interpro,		Following interpretation of textual descriptions	aar1834, aar2188, aar23, aar2437, aar251, aar258, aar282, aar2893, aar3350, aar697, fbr819, pfr1218,

	CDD (AGMIAL) 3. Brenda (Brenda) 4. Enzym (SIB)			pfr1441, pfr1471, pfr1489
thp / thr	BDBH (AGMIAL)	Blastp	Threshold on expected value partition: one orthologue or not	
ttm	MEMSAT (AGMIAL) TM-HMM, SignalP, and LipoP (CBS)	default	Number of predicted motifs, threshold on score, from graphic visualisation	fbr1970, fbr3271, fbr821, pfr1489, pfr2638
tct	Blast alignment (AGMIAL)	Blastp	Same protein length or not	fbr1970
tsg	Query (Genoscope)			fbr1542, pfr1402, pfr2106
tsz	PAM (AGMIAL)			fbr1970, fbr3271
thm	multalin	default	From graphic visualisation	fbr3271
thb	ABCisse (Pasteur)	Blastp	Threshold on expected value partition: yes or no	pfr2016
tgc	GC%			pfr1656
cfa	-	-	-	pfr1402

De l'ensemble des composants de la syntaxe utilisée pour décrire les tâches bioinformatiques (collections, filters, transformers, transformer-filters, conditional and parralel forks) de [Stevens *et al.*, 2001] seuls les filters, transformers, transformer-filters correspondent à une classification des tâches. Nous avons donc réparti les 16 tâches dans ces trois classes :

- filters (3 entrées : une collection restreinte, une source de données à filter, une projection pour filter ; une sortie : la collection issue de la projection de chaque donnée sur la base de données) : thp, thr, tmc, tmp, ttm, tei, thb,
- transformers (entrée : une collection, sortie : la collection transformée pas le processus) : tss, tct, tsg, tsz, thm, tgc
- transformer-filters (filtre et transformation non ordonnés) : tmo, tmt, tsa

MS 3 : Listes de critères de décision associés aux tâches

extraction à partir des fichiers .dot :

```
grep "{" *.dot | gawk 'BEGIN{FS=":"}{print $2":"$3":"$4":"$5}' | sort > tests.txt
```

Tâches de recherche de similarité de séquences (thr et thp)

La liste des critères est présentée suivie des ensembles de valeurs entre accolades (en italiques sont indiquées les valeurs des critères non observées mais nécessaires pour compléter l'ensemble).

- score {ok, faible}
- rapport (longueur alignement)/(longueur protéine) {*ok*, court}
- positions des alignements {*centré*, plusieurs régions différentes de la protéine}
- annotations des protéines hits :
    - cohérence des annotations {*oui*, fonctions biologiques ou biochimiques (EC numbers) différentes, annotation incohérente avec la voie métabolique supposée}
    - absence d'annotation {annotation « putative » , pas de fonction identifiée}
- compatibilité des longueurs des protéines {*ok*, start variables, région Nterm tronquée, protéine trop courte (manquent 12 aa)}
- type du start {*atg*, gtt peu fréquent}
- domaines identifiés {*non, un seul,* plusieurs domaines identifiés ou seulement sur une partie de la séquence}
- conservation de la région Nterm {*oui*, peu conservée ce qui est signe d'un segment particulier}
- relation taxonomique du génome duquel provient le hit {*même taxon*, taxon éloigné}
- position du hit swissprot {*appartient aux premiers hits*, pas en tête de Blast}

Motifs/domaines (tmc et tmp)

- score {*ok*, faible}
- annotation d'un domaine {*aucune, une*, deux annotations possibles}
- comparaison avec l'hypothèse d'annotation {en accord, en désaccord}