

Scénario pour l'annotation de génomes microbiens

Jean-François Gibrat

27 septembre 2005

1 Introduction

Le but de ce document est de définir un scénario possible pour l'annotation. Je repars du document de Lucie « Stratégie pour l'annotation de génomes microbiens partie IV - version 3 » du 4 avril 2005. Quelques notations pour commencer : les protéines du génome qu'il faut annoter sont notées P_i , i variant de 1 à N le nombre total de protéines dans le génome. Par différentes méthodes (cf. section 2 du document de Lucie) on peut établir des relations entre les protéines du génome à annoter et des protéines stockées dans les bases de données. Nous noterons ces dernières R_i^k . Pour chaque protéine i du génome à annoter on peut trouver M ($k = 1 \dots M$) protéines ayant une relation avec elle (que ce soit une relation d'homologie, une relation issue du contexte des gènes, etc.). M peut varier de zéro à plusieurs centaines selon les protéines i .

2 Arbres de décision pour l'annotation

Les protéines possèdent un certain nombre de propriétés qui leur sont propres (cf. section 3 du document de Lucie). Elles ont une longueur, un point isoélectrique, une masse moléculaire, une localisation cellulaire particulière, elles sont composées de divers domaines soit globulaires soit non globulaire (zones désordonnées, zones de faible complexité, zones de coiled-coils, peptide signal, segments membranaires). On peut aussi inclure comme propriété l'espèce d'où provient la protéine. Il est important de noter que les propriétés citées ci-dessus peuvent être déterminées aussi bien pour les protéines du génome P_i que pour les protéines des bases de données R_i^k .

Cependant, en plus, ces dernières peuvent posséder des attributs (propriétés) liés à leur fonction. Ce sont (cf. section 3.3 du document de Lucie) : la classe fonctionnelle (FunctClass), les mots-clés SwissProt (SP_KW), les termes GO (GO_TERM), le résultat de l'analyse du texte SWISSPROT (SP_TXT), la classification des enzymes (EC number). Ces données ne sont disponibles, bien évidemment, que pour les protéines ayant déjà été annotées. Ces informations sont donc cruciales pour l'annotation de nouvelles protéines.

Le principe de l'annotation consiste à établir une relation entre la protéine annotée et des protéines des bases de données et au vu de la « force » et du type de la relation et des propriétés des protéines des bases de données d'en déduire une annotation. Il est important de noter que toutes les relations ne sont pas équivalentes. Par exemple une relation d'homologie et une relation de voisinage sur le génome ne fournissent pas les mêmes informations. De même la « force » d'une relation peut être variable : une relation d'homologie ayant une E-value de 10^{-75} et une conservation de 95% de la séquence ne fournit pas la même information qu'une relation d'homologie ayant une E-value de 0.2 et une conservation de 15% des résidus. En outre l'annotation va dépendre d'une façon critique des attributs liés à la fonction qui sont disponibles pour les protéines homologues.

La figure 1 présente un « arbre de décision » pour le cas de la relation d'homologie.

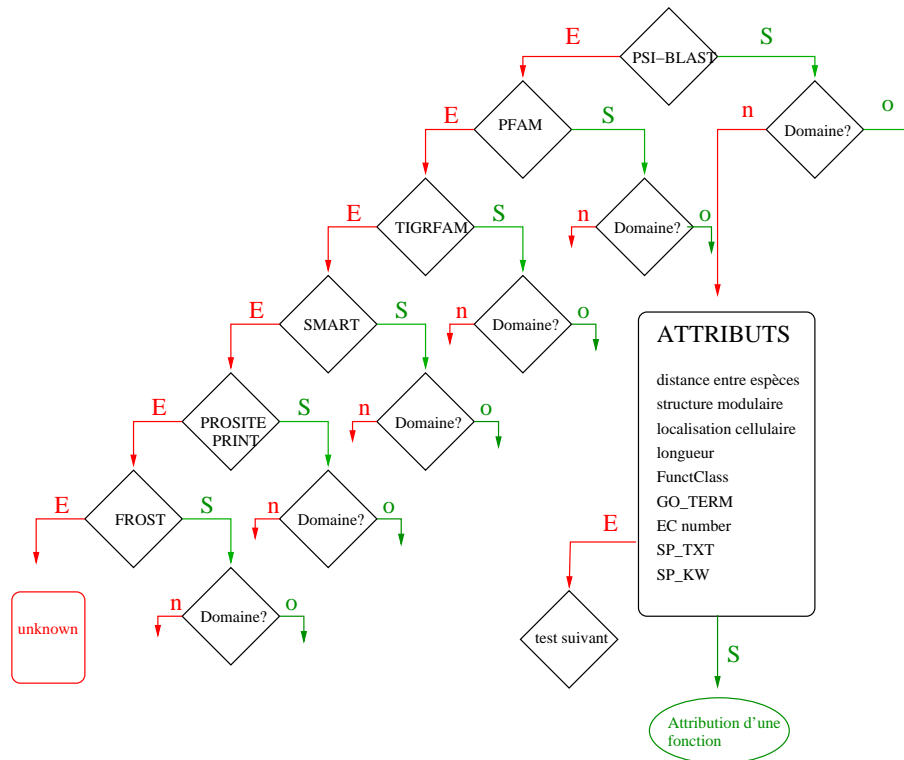


FIG. 1 – Arbre de décision pour la relation d'homologie

Chaque losange représente un test, par exemple le premier consiste à tester avec PSI-BLAST l'existence de protéines homologues. Ici, pour simplifier, il n'y a que 2 possibilités : le succès (branche S) et l'échec (branche E). Le succès peut être défini comme l'existence d'au moins une protéine homologue sortant avec une E-value inférieure à un seuil particulier. On pourrait envisager, comme discuté lors de la dernière réunion, une sortie à 3 branches : i) on obtient des protéines dont on est certain de l'homologie, ii) on obtient des protéines pour lesquelles le résultat est ambiguë, iii) il n'y a pas de protéines homologues.

En cas de succès on passe à un autre test qui vérifie si la ressemblance en terme de séquence est globale ou bien locale (c'est à dire que les 2 protéines n'ont qu'un domaine en commun). Il est important pour l'annotation de bien distinguer ces 2 cas. Quelque soit le cas, protéine complète ou domaine, on utilise ensuite les attributs (propriétés) des protéines homologues. Ici on peut soit créer un nouvel arbre de décision basé sur les attributs soit créer un vecteur (le problème étant qu'il peut manquer des valeurs) qui décrit la protéine homologue et on laisse un algorithme de classification, par exemple un SVM, se débrouiller avec. Une difficulté supplémentaire c'est qu'en général il y a plusieurs proteines homologues. Si ces protéines fournissent toutes le même résultat pas de problème sinon il faut disposer d'un moyen pour décider comment choisir la fonction (un vote pondéré par la ressemblance avec la protéine cible par exemple). Suite à cette procédure soit on peut assigner une fonction dans la hiérarchie et on s'arrête là soit on échoue. Il y a donc deux types d'échec pour un test particulier : direct ou indirect. Pour PSI-BLAST l'échec direct consiste à ne pas trouver de protéine homologue ayant une E-value en dessous d'un seuil et l'échec indirect consiste à trouver des protéines homologues pour lesquelles il n'y a pas d'annotation utilisable.

A la suite d'un échec on passe au test suivant, sur la figure le test PFAM, où on réitère la procédure utilisée pour le premier test. Il faut noter que certains tests pourraient être groupés par exemple les

2 méthodes PFAM et TIGRFAM se ressemblent beaucoup on pourrait envisager d'avoir un test qui réunisse les 2 méthodes (c'est à dire que l'on considère comme un succès si l'on trouve une protéine homologue avec l'une, ou l'autre des méthodes). C'est le cas pour le test PROSITE-PRINT.

Globalement il y a 2 façons de sortir de l'arbre : soit une méthode permet d'attribuer une fonction soit on sort sur la branche la plus à gauche.

Dans un second temps il faudrait aussi considérer les cas où plusieurs méthodes individuelles ont fourni des protéines homologues à partir desquelles il a été impossible d'en conclure une fonction particulière. Il est possible, si l'on considère l'ensemble des informations recueillies, qu'on soit à même de conclure sur la fonction.

La Figure 2 présente les méthodes qui permettent d'obtenir des relations entre protéines à partir du contexte des gènes. Il n'y a pas de cascade de tests comme précédemment car chaque méthode donne des informations différentes. En cas de succès à l'un des tests on obtient une liste de protéines ayant une relation avec une protéine du génome P_i . Chaque protéine P_i^k possède une liste d'attributs similaires à ceux de la figure 1 avec un attribut de plus, à savoir le type de relation la liant à la protéine P_i . Ces différentes listes sont combinées comme dans le cas de la figure 1.

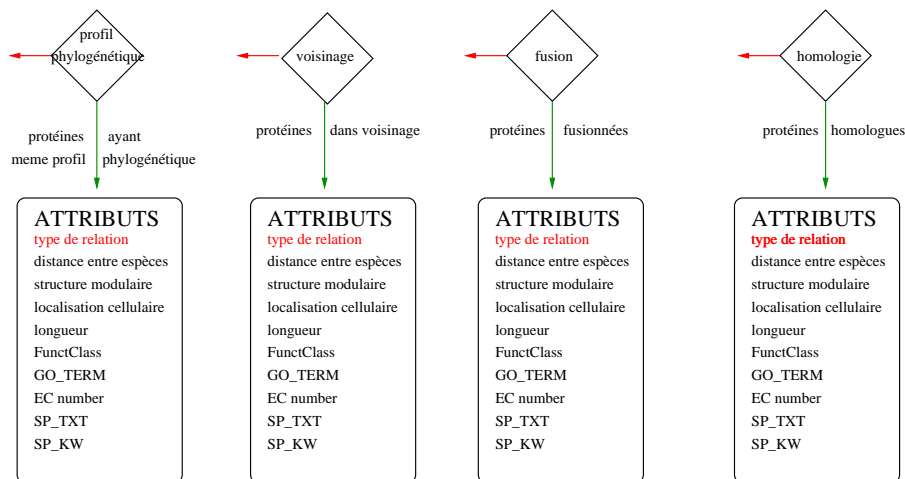


FIG. 2 – Arbre de décision pour les relations fondées sur le contexte des gènes

Dans la figure 1 on pourrait également envisager, au lieu d'avoir une cascade de tests, que chaque test fournisse une liste (pouvant être vide) de protéines homologues et que l'on combine ensuite le tout.