SIGffRid : Programme de recherche des sites de fixation des facteurs de transcription par approche comparative

Fabrice Touzain, Sophie Schbath, Isabelle Debled-Rennesson, Bertrand Aigle, Pierre Leblond, Gregory Kucherov

SIGffRid : Programme de recherche des sites de fixation des facteurs de transcription par approche comparative

F. Touzain [a] , S. Schbath [b] , I. Debled-Rennesson [a] , B. Aigle [c] , P. Leblond [c] , et G. Kucherov [a]
[a]Laboratoire Lorrain de Recherche en Informatique et ses Applications,
Vandoeuvre-Lès-Nancy, France,
[b]Unité Mathématique Informatique et Génome INRA,
Jouy-en-Josas, France,
[c]Laboratoire de Génétique et de Microbiologie de l'Université Henri Poincaré,
de Nancy, France.
email : touzain@loria.fr

23 août 2005

Résumé

Notre objectif est la recherche des sites de fixation des sous-unités σ de l'ARN polymérase dans des génomes bactériens, sites généralement composés de deux « boîtes » dites -35 et -10 en référence au point d'initiation de la transcription. En utilisant la génomique comparative, nous souhaitons nous assurer de la conservation de couples de boîtes statistiquement intéressantes (programme R'MES [1]) liée à la présence d'un motif régulateur fonctionnel. Nous autorisons un espacement variable entre boîtes -35 et -10 conservées dans les séquences intergéniques amonts de gènes supposés orthologues[1]. Cette démarche, généralisable à tout couple de bactéries proches, permet de prédire les sites de fixation des facteurs de transcription (SFFT) qui leur sont communs. Un descriptif du fonctionnement du programme est présenté. Plusieurs résultats prometteurs ressortent de son utilisation. Trois SFFT connus sont retrouvés ou confirmés, avec un grand nombre de nouveaux gènes co-régulés candidats pour chacun. Deux groupes de motifs ressemblent à divers SFFT référencés, suggérant certaines hypothèses biologiques sur les résultats connus. Au moins deux nouveaux SFFT sont proposés, à la fois chez Streptomyces coelicolor et Streptomyces avermitilis.

[1]Sont dits « orthologues » deux gènes de bactéries différentes issus d'un gène d'une bactérie ancestrale commune.

1 Introduction

De nombreux programmes ont été conçus dans le but de découvrir des SFFT. Certains sont comparés dans un article récent [2] qui montre la diversité et le nombre des solutions avancées pour la résolution de ce problème fondamental en bioinformatique, qu'un paragraphe ne saurait résumer. La plupart d'entre eux ne peuvent pas utiliser d'espacement variable entre deux mots (MEME [3]) ou font appel à des méthodes pour lesquelles le bruit de fond peut

interférer avec le motif biologique (Bioprospector [4]). Il existe un programme permettant de fixer précisément les contraintes structurelles des motifs recherchés : Smile [5]. Néanmoins, à vocation plus généraliste, il n'oriente pas ses comparaisons en fonction de données phylogéniques et impose de fixer un quorum pour la représentation d'un motif dans les séquences d'intérêt.
Dans le cadre de la recherche de SFFT dans des génomes bactériens, nous avons ciblé nos comparaisons et permis un espacement variable entre boîtes -35 et -10 potentielles, définies en nous appuyant sur des statistiques rigoureuses. Telles sont quelques-unes des améliorations apportées par le programme SIGffRid (SIGma Factor (binding site) Finder using R'mes to select Input Data). Décrit ci-après, il s'appuie sur l'analyse simultanée de couples de séquences extraits de deux génomes de bactéries phylogénétiquement proches, et l'utilisation du programme R'MES [1].

2 Données initiales

Le programme principal nécessite plusieurs types de données :

- la séquence totale du génome qui nous intéresse (chromosome et plasmide(s)),
- toutes les séquences amonts de gènes, fusionnées si elles se chevauchent et sont de même orientation (nous considérons les deux brins d'ADN distinctement), pour chaque bactérie (Fig 1),
- les mots statistiquement sur-, ou sur- et sous-représentés chez la bactérie d'intérêt (sorties de R'MES modifiées),
- les probabilités de transition d'un modèle de Markov d'ordre 3 ajusté sur l'ensemble du génome pour chaque bactérie,
- des fichiers comportant chacun deux séquences intergéniques amonts de gènes orthologues (une par bactérie).

Des scripts ont été réalisés pour obtenir ces données à partir de fichiers embl, genbank et des sorties standard de R'MES. Plusieurs critères biologiques sont pris en considération (données pouvant évoluer en fonction des connaissances que nous avons des SFFT) :

- longueurs minimale et maximale de l'espacement entre boîtes -35 et -10 tous facteurs sigma confondus (10 et 25 respectivement par défaut),
- variabilité de cet espacement pour un facteur sigma et une bactérie donnés (1 par défaut),
- variabilité de cet espacement pour un facteur sigma donné entre deux bactéries phylogénétiquement proches (1 par défaut).

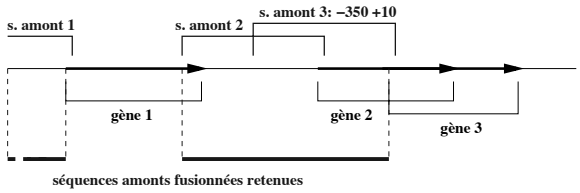

Fig. 1 – Définition des séquences amonts utilisées

3 Approche

Les techniques décrites supposent plusieurs hypothèses biologiques :

- les SFFT sont localisés en amont des gènes,
- des gènes orthologues de bactéries phylogénétiquement proches ont de bonnes chances d'avoir conservé le même type de régulation transcriptionnelle,

- la conservation de la structure de l'ADN reconnue par une sous-unité σ de l'ARN polymérase se traduit par une conservation de portions de séquences nucléotidiques (souvent nommées boîtes -35 et -10 en référence à leur position approximative par rapport au site d'initiation de la transcription).

La première étape consiste à déterminer les boîtes à rechercher. Nous sélectionnons les mots donnés par R'MES [1] (http ://www-mig.jouy.inra.fr/ssb/rmes) comme statistiquement sur-représentés sur l'ensemble du génome de la bactérie d'intérêt. L'étude des boîtes -35 et -10 connues chez Streptomyces coelicolor a en effet montré que les mots correspondants, ou des sous-mots de taille raisonnable les composant, présentaient un score d'exceptionnalité positif et très élevé dans la sortie de R'MES.
Intuitivement, si nous considérons un motif de SFFT, nous devrions avoir un faible nombre d'occurrences pour chaque boîte sur la totalité d'un génome comparativement aux autres mots. En effet, si une protéine se fixe sur les bases correspondant à ce motif, il est fort possible qu'elle puisse interagir avec chacune de ses composantes. Celles-ci devraient donc avoir une influence négative sur la disponibilité de cette protéine, donc sur le processus de transcription qui lui est associé. Néanmoins, les boîtes des SFFT devraient être significativement sur-représentées par rapport à leur nombre attendu si nous tenons compte des probabilités des sous-mots les composant. En revanche, la sous-représentation de ces derniers explique le plus souvent le faible nombre d'occurrences observées d'une boîte sur la totalité du génome. Nous nous appuyons sur les relations d'orthologies issues de la base de données MBGD [6], et regroupons les paires de séquences amonts de gènes orthologues via les familles définies dans celle-ci, pour diminuer le nombre de séquences traîtées simultanément. Ces familles ne sont que de grandes catégories de gènes permettant de scinder les paires de séquences amonts d'orthologues en sous-groupes, sinon cohérents du point de vue de la régulation, au moins logiques par rapport aux fonctions supposées des gènes. Elles permettent de limiter la mémoire nécessaire au programme qui traîtera successivement chaque groupe de gènes. Via des scripts perl additionnels, nous récupérons les séquences intergéniques amonts correspondant à des gènes orthologues probables groupés par paires (une séquence intergénique par bactérie pour chaque relation d'orthologie, des positions -350 -au mieux- à +10 par rapport au site d'initiation de la traduction et d'une taille minimale de 30 nucléotides). Nous y recherchons alors les couples de mots intéressants conservés avec un espacement compatible avec la fixation d'un facteur de transcription (Fig 2). Pour chaque paire de séquences amonts d'orthologues, nous obtenons donc une liste de candidats en tant que SFFT.

Ces résultats intermédiaires sont alors groupés (et dupliqués si besoin) par similarités de doublet de trinucléotides (un pour chaque boîte) et d'espacement (avec une variation de 1 autorisée) (Fig 3).

A partir de cette étape, nous allons traîter les séquences de chaque bactérie séparément, ceci afin de pouvoir mettre à jour des motifs proches mais ayant évolué différemment pour chaque bactérie.

Un tri des séquences concernées est réalisé de concert avec l'extension du motif qui leur est commun et l'évaluation du motif consen-

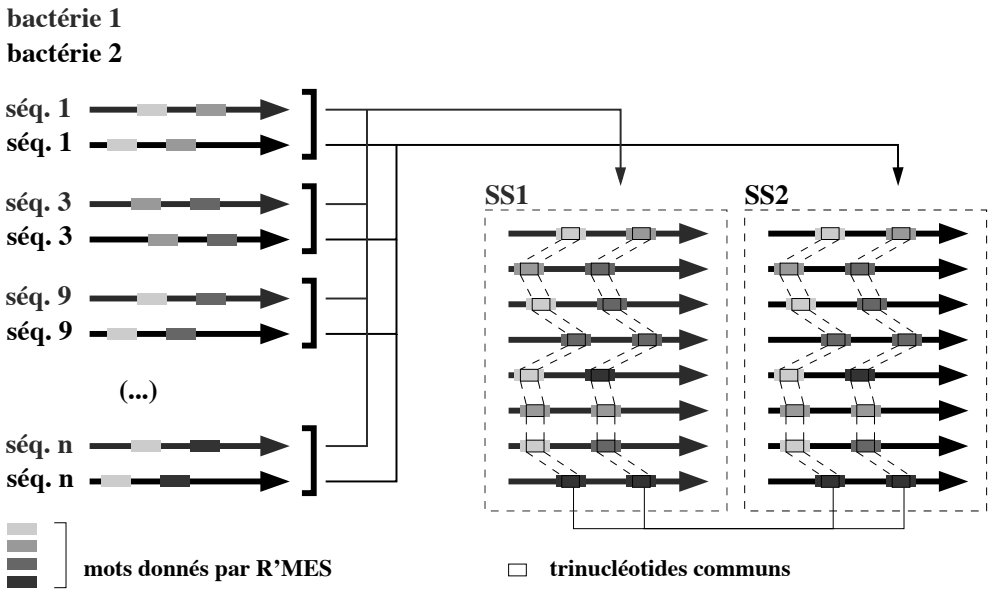

Fig. 3 – Conservation des mots intéressants dans un ensemble de paires de séquences intergéniques amonts d'orthologues

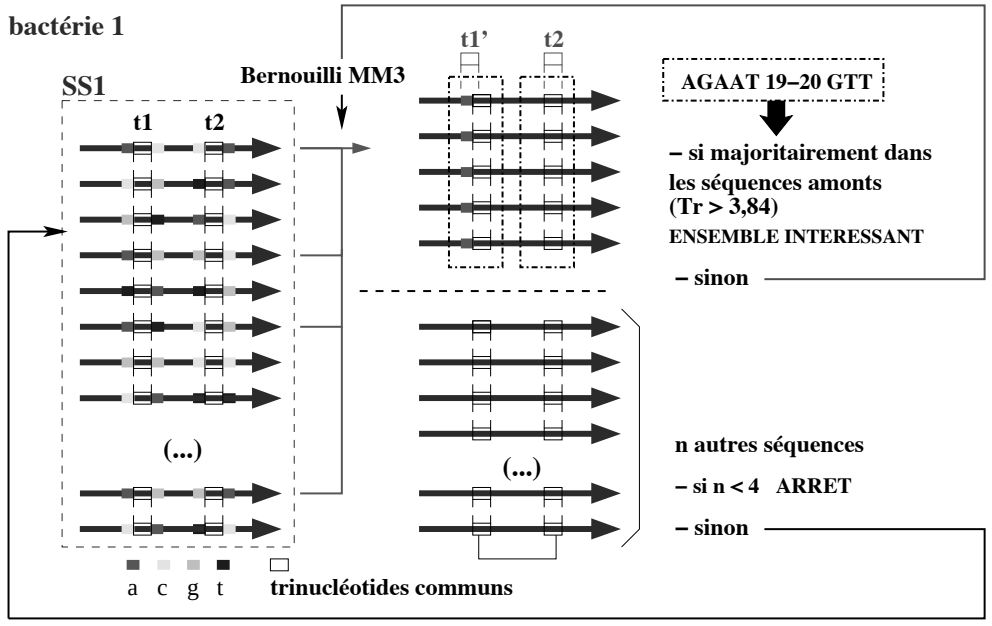

Fig. 4 – Extension des trinucléotides communs, tri des séquences

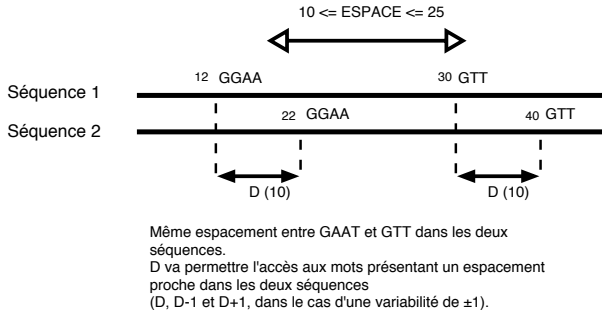

Fig. 2 – Conservation des mots intéressants dans les séquences intergéniques amonts d'une paire d'orthologues

sus résultant. Cette extension de motif s'appuie sur un modèle probabiliste (Fig 4). Elle se poursuit récursivement tant qu'aucun motif intéressant n'a été mis en exergue et que le nombre de séquences est suffisant. Nous décrivons précisément dans le paragraphe suivant les techniques utilisées.

3.1 Algorithme

3.1.1 Définition des mots recherchés

Vu les tailles de mots (¡8) et de génomes (¿8 Mb) considérées, nous avons utilisé R'MES dans sa version approximation gaussienne du comptage particulièrement adaptée aux mots fréquents. Nous avons analysé conjointement les mots et leur complémentaire inversé (option -fam). Les scores d'exceptionnalité calculés par R'MES correspondent aux transformations probit des p-values approchées, une p-value étant la probabilité qu'un mot soit observé autant de fois dans une séquence aléatoire de même composition que le génome analysé (modèle markovien d'ordre maximal). Ainsi, ces scores sont comparables à des variables de loi $\mathcal{N}(0,1)$. Pour décider de l'ensemble des mots de fréquence exceptionnelle, nous avons défini, pour chaque longueur h de mots, un seuil t inspiré du critère de Bonferroni :

$$P(\mathcal{N}(0,1) \geq t) = \frac{\alpha_{r'mes}}{4^h}, o\grave{u} \quad \alpha_{r'mes} = 5.10^{-3}$$

Cela nous a donc donné un ensemble W de mots exceptionnels de longueur $3 \leq h \leq 7$ sur l'alphabet $\mathcal{A} = \{a, c, g, t\}$.
Ces mots sont recherchés dans chaque paire de séquences intergéniques amonts d'orthologues.

3.1.2 Propriétés des motifs retenus comme pouvant être des SFFT

Soient d_{min} et d_{max} les espacements minimaux et maximaux autorisés entre les boîtes -35 et -10 (données biologiques).
Soient D la variation biologique d'espacement acceptée entre les SFFT des deux séquences amonts, et $sp1$ et $sp2$ deux espacements $\in [d_{min}..d_{max}]$.
Soit un triplet $C_i = \{w_i^1, w_i^2, \{s_{1i}, s_{2i}\}\}$ correspondant aux mots w_i^1 et $w_i^2 \in W$ dans les séquences amonts d'orthologues s_{1i} et s_{2i}. C_i est considéré comme intéressant si w_i^1 et w_i^2 sont présents dans s_{1i} et s_{2i} avec des espacements $sp1$ et $sp2$ respectivement tels que $sp2 = sp1 \pm D$.
Si $p_{s_{1i}}(w_i^1)$, $p_{s_{1i}}(w_i^2)$ sont les positions de w_i^1 et w_i^2 respectivement dans s_{1i} et $p_{s_{2i}}(w_i^1)$, $p_{s_{2i}}(w_i^2)$ les positions de w_i^1 et w_i^2 respectivement dans s_{2i}, nous avons la relation suivante :

$$p_{s_{1i}}(w_i^1) - p_{s_{2i}}(w_i^1) = p_{s_{1i}}(w_i^2) - p_{s_{2i}}(w_i^2) \pm D \quad (1)$$

Cela nous permet de grouper les mots par espacements proches. Nous ne gardons pour chaque paire de séquences d'orthologues que les couples C_i qui vérifient la relation (1), candidats potentiels en tant que SFFT.

3.1.3 Extension de motifs et tri des séquences

Après cela, nous regroupons les résultats intermédiaires (ensemble des C_i) en fonction des trinucléotides qui composent w_i^1 et w_i^2.
Soient $t1$ et $t2$ deux trinucléotides et d_{t1-t2} l'espace qui les sépare.
Soit e un entier.
Pour chaque triplet $(t1, t2, d_{t1-t2})$ possible en considérant l'ensemble des C_i obtenus pour toutes les paires de séquences amonts d'orthologues, nous allons créer un ensemble $\mathcal{C}$ de tous les C_i qui vérifient :

$$(t1 \subset w_i^1) \wedge (t2 \subset w_i^2) \wedge (d_{t1-t2} \in [e..e+D])$$

De chaque $\mathcal{C}$, nous récupérons deux ensembles de séquences $SS_1 = \bigcup s_{1i} \in \mathcal{C}$ et $SS_2 = \bigcup s_{2i} \in \mathcal{C}$, un pour chaque bactérie (Fig 3).
Soit min_{SS} le nombre de séquences distinctes minimal intervenant dans l'obtention d'un motif de SFFT candidat. Nous ne gardons chaque ensemble SS_1 ou SS_2 que s'il présente au moins min_{SS} séquences distinctes.
$t1$ est le trinucléotide de gauche qui sera inclus dans la boîte -35 d'un éventuel SFFT potentiel et
$t2$ est le trinucléotide de droite qui sera inclus dans la boîte -10 du même SFFT potentiel.
Pour chacun des ensembles SS_1 et SS_2, les séquences sont triées par comptage et évaluation statistique des lettres jouxtant $t1$ et $t2$. Notre critère statistique s'appuie sur les probabilités de transition d'un modèle de Markov d'ordre 3 ajusté pour chaque bactérie sur l'ensemble du génome.
Soient les positions :

- 1 : immédiatement à gauche de $t1$,
- 2 : immédiatement à droite de $t1$,
- 3 : immédiatement à gauche de $t2$,
- 4 : immédiatement à droite de $t2$.

Soient la position $g \in \{1, 3\}$ et la position $d \in \{2, 4\}$.
Soit $\ell \in \mathcal{A}$, le nucléotide dont nous considérons la probabilité d'obtention à une position donnée.
Soit n le nombre de séquences concernées.
Soit t le trinucléotide à étendre.
Soit $j \in [1..2]$ fixé, l'indice permettant de préciser l'ensemble de séquences traîté.
Pour l'extension d'une lettre de droite, posons :

$$Y_i^d(\ell) = \begin{cases} 1 & \text{si la ième séquence de } SS_j \text{ possède} \\ & \text{le nucléotide } \ell \text{ en position } d, \\ 0 & \text{sinon.} \end{cases}$$

Le nombre $N^d(\ell)$ de séquences possédant le nucléotide ℓ en position d, $N^d(\ell) = \sum_{i=1}^{n} Y_i^d(\ell)$, suit alors une loi binomiale $\mathcal{B}(n, N(t\ell)/N(t))$, où $N(\cdot)$ désigne le comptage et $t\ell$ le tétranucléotide formé de t suivi de ℓ. Nous pouvons ainsi calculer la significativité $p^d(\ell)$ du nombre de séquences avec un ℓ en position d :

$$p^d(\ell) = 1 - \sum_{y=0}^{x-1} C_n^y (\frac{N(t\ell)}{N(t)})^y (1 - \frac{N(t\ell)}{N(t)})^{n-y}$$

Pour l'extension d'une lettre de gauche, posons :

$$Y_i^g(\ell) = \begin{cases} 1 & \text{si la ième séquence de } SS_j \text{ possède} \\ & \text{le nucléotide } \ell \text{ en position } g, \\ 0 & \text{sinon.} \end{cases}$$

Le nombre $N^g(\ell)$ de séquences possédant le nucléotide ℓ en position g, $N^g(\ell) = \sum_{i=1}^{n} Y_i^g(\ell)$, suit alors une loi binomiale $\mathcal{B}(n, N(t\ell)/N(t))$, où $N(\cdot)$ désigne le comptage et ℓt le tétranucléotide formé de t précédé de ℓ. Nous pouvons ainsi calculer la significativité $p^g(\ell)$ du nombre de séquences avec un ℓ en position g :

$$p^g(\ell) = 1 - \sum_{y=0}^{x-1} C_n^y (\frac{N(\ell t)}{N(t)})^y (1 - \frac{N(\ell t)}{N(t)})^{n-y}$$

Nous choisissons le nucléotide k et la position $i \in \{1,2,3,4\}$ les plus significatifs (minimisation des probabilités $(p^d(\ell), p^g(\ell)))$ avec $N^{d|g}(\ell) \geq 4$. Les séquences possédant la lettre k à la position i sont regroupées pour les étapes suivantes (Fig 4). Un motif correspondant à cet ensemble de séquences est généré et évalué (cf. § 3.1.4).
- S'il est considéré comme intéressant, le processus d'extension se poursuit sans test sur R et T_R (cf. § 3.1.4), pour trier les séquences et faciliter leur comparaison visuelle (jusqu'à ce que l'extension concerne moins de 4 séquences), et nous marquons cet ensemble de séquences pour son affichage ultérieur dans les résultats (en enregistrant l'intervalle des indices de séquences intéressantes, la matrice d'évaluation et le motif correspondant (cf. § 3.1.4)),
- Si le nombre de séquences concernées devient trop faible ($< min_{SS}$), le processus s'arrête,
- Si le motif n'est pas intéressant, nous poursuivons l'extension, en remplaçant :
  - $t1$ par $t1' = k.t1[1].t1[2]$, si $i = 1$,
  - $t1$ par $t1' = t1[2].t1[3].k$, si $i = 2$,
  - $t2$ par $t2' = k.t2[1].t2[2]$, si $i = 3$,
  - $t2$ par $t2' = t2[2].t2[3].k$, si $i = 4$,

  où . est l'opérateur de concaténation.
  (déplacement d'une lettre dans toutes les séquences concernées)

Les autres séquences sont traîtées distinctement suivant la même démarche.

3.1.4 Génération d'un motif consensus et son évaluation

A chaque étape de regroupement, un motif générique est déduit correspondant à deux mots avec un espacement variable. Il est construit en ajoutant au couple de tinucléotides les lettres présentes dans 70% des séquences concernées (par extension de ces derniers en s'appuyant sur une double matrice position-spécifique) (Fig 5). Ce motif est recherché dans l'ensemble des séquences amonts fusionnées pour chaque brin, et dans le génome entier dans les deux sens. De ces comptages est déduit un rapport R :

$$R = \frac{|motif_{\text{séq amonts}}|}{|motif_{\text{séq totale 2 sens}}|}$$

Ce rapport mesure la spécificité du motif pour les séquences amonts. Il est généralement admis que les SFFT sont localisés en amont des gènes. Pour tester la significativité du rapport R, nous effectuons un test du rapport de vraisemblance [7] dont la statistique de test T_R suit une loi $\chi^2\{1\}$ et est donnée par :

$$T_R = 2\left[N_1 log\left(\frac{\frac{N_1}{L_1}}{\frac{N}{L}}\right) + N_2 log\left(\frac{\frac{N_2}{L_2}}{\frac{N}{L}}\right)\right]$$

avec
$L_1 = l_{amont} - (l_{motif} \times nbseq)$,
$L_2 = 2(l_{genome} - l_{motif})$,
$L = L_1 + L_2$, et $N = N_1 + N_2$,
où
l_{amont} est la somme des longueurs des $nbseq$ séquences amonts de gènes,
l_{genome} la longueur totale du génome complet et l_{motif} la longueur maximale pouvant être prise par l'expression régulière correspondant au motif dénombré.
N_1 est le nombre d'occurrences du motif dans les séquences amonts, et
N_2 le nombre d'occurrences dans le génome total et son complémentaire inversé.
T_R conditionne la poursuite ou l'arrêt de l'extension du motif consensus par tri des séquences. Une sélection des résultats les plus intéressants est faite via les rapports R et T_R. La relation :

$$(R \geq R_{min}) \wedge (T_R \geq T_{R\_min})$$

doit être vérifiée,
avec R_{min} le seuil minimal de spécificité (pour le moment empirique, fixé à 0.35, mais qui devrait à terme être déduit et adapté pour chaque bactérie), et T_{R_min} le quantile à 5% ($\alpha_{T_{R\_min}}$) de la loi du χ^2.

3.1.5 Visualisation des résultats

La figure 5 montre l'aspect d'un motif résultat et l'ensemble des séquences ayant permis de le générer.

Ils sont complétés par une recherche automatique du motif générique dans l'ensemble des séquences amonts de la bactérie concernée, fournissant ainsi les identifiants des gènes, et les positions des occurrences par rapport au début de la traduction.

Les seuls travaux pour validation de ces résultats sont donc d'ordre biologique : vérification de la cohérence des fonctions des gènes liés par un même motif de régulation et expériences. Aucun post-traitement manuel des résultats ne sera nécessaire avant interprétation dans la version finalisée du programme.

Toutes les étapes, de la récupération des orthologues et l'extraction des séquences, en passant par l'utilisation de R'MES jusqu'à l'affichage des résultats sont automatisées en PERL.

4 Discussion

4.1 Point de vue informatique

Nombre d'idées intéressantes pour la recherche des sites SFFT avaient été utilisées isolément : emploi de motifs composites [5, 4], de relations d'orthologies pour cibler les comparaisons [8, 9], de statistiques pour post-traîtement des résultats [5]. Elles n'avaient néanmoins jamais été combinées. Le programme présenté ici n'est pas exhaustif, puisqu'il nécessite encore le regroupement des gènes des bactéries concernées en grandes fonctions (16 dans le cas présent) du fait de la grande taille des génomes utilisés. Une évolution prochaine devrait permettre le traîtement à partir de toutes les relations d'orthologie disponibles entre deux bactéries, quelles que soient les tailles de leur génomes. Ceci est rendu possible par une sélection statistique rigoureuse des mots recherchés.
D'autres caractéristiques propres à ce programme tiennent mieux compte de la nature des SFFT. Ainsi, des variations d'un même SFFT peuvent exister dans deux bactéries phylogénétiquement proches [10]. Nous les distinguons par l'alignement des séquences de chaque bactérie séparément. Nous obtenons des variantes éventuellement différentes d'un même SFFT dans deux bactéries proches, les différences pouvant concerner aussi bien les boîtes que la longueur de l'espacement qui les sépare.
Une limitation actuelle du programme est le choix unique qu'il fait pour l'extension des boîtes d'un motif donné. Il est possible qu'il ne détecte pas certains motifs simplement parce que ceux-ci recoupent d'autres motifs dont les caractéristiques statistiques sont plus significatives.

4.2 Point de vue biologique

Nous avons utilisé des bactéries phylogénétiquement proches de la famille des Actinomycètes, Streptomyces coelicolor et Streptomyces avermitilis. Outre leur intérêt économique (les Actinomycètes sont responsables de la production de plus de 70% des antibiotiques connus), ces Streptomyces

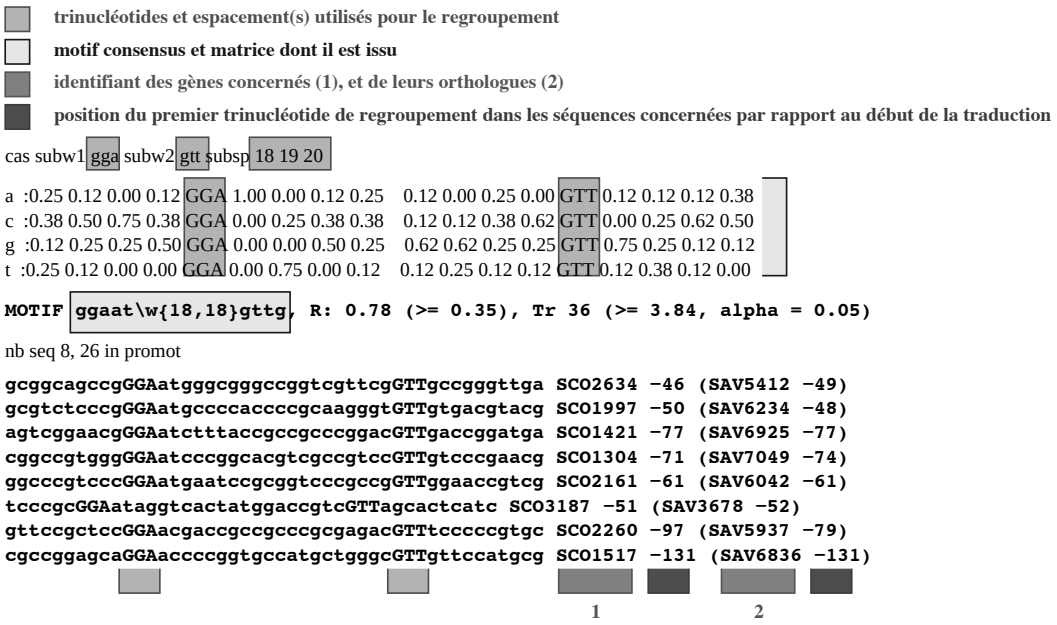

Fig. 5 – Exemple de résultat pour un motif

présentent des génomes de grande taille et un nombre de facteurs sigma prédits sans commune mesure (plus de 60), ce qui en fait des modèles de choix dans notre étude. Notre programme a donc été éprouvé sur ces deux bactéries.

Les fonctions des gènes concernés par nos prédictions peuvent être employées comme première validation des résultats. Seuls des tests biologiques seront à même de les confirmer.

Néanmoins, des SFFT connus ont été retrouvés (sigR [10], ggaat-n(18)-gtt, 30 SFFT répertoriés ; BldN [11], cgtaac-n(16)-cgttga, un seul SFFT répertorié ; HrdB [12, 13, 14, 15, 16]) et de nombreux gènes potentiellement corégulés sont proposés. Certains motifs identifiés via d'autres méthodes de prédiction, ont pu être retrouvés (tgtcagtgc-n(7)tagg, tgac-n(19)-tgac [8]). De nouveaux candidats sont en outre avancés.

Seuls sont donnés les motifs connus jusqu'à présent, le programme en fournit parfois des variations plus ciblées ou plus généralistes. Il est à noter qu'approximativement 30 des 65 facteurs sigma supposés de Streptomyces coelicolor peuvent être groupés en trois grandes familles de facteurs sigma aux domaines de fixation supposés aux boîtes -35 et -10 extrêmement proches (régions 2.4 et 4.2). Bien qu'on ne puisse déduire directement une conservation des boîtes -35 et -10 à partir de la conservation des domaines protéiques 2.4 et 4.2 des facteurs sigma, il est fort possible que les boîtes reconnues soient proches pour une famille de facteurs sigma donnée.

De nombreuses validations seront nécessaires, mais au vue des recoupements déjà observés avec des SFFT connus, l'approche paraît très prometteuse.

5 Conclusion

Les résultats obtenus sont intéressants à plus d'un titre. D'une part parce qu'ils recoupent parfaitement la plupart des SFFT connus et précisément définis, d'autre part parce que certains des autres motifs avancés sont très proches de ceux supposés pour la régulation de certains gènes (cas de HrdB [12, 13] par exemple). En outre, d'autres motifs spécifiques des régions amonts ont été mis à jour et pourraient aussi être impliqués dans des mécanismes de régulation.

Evidemment, ce programme ne dispense pas d'une vérification expérimentale des résultats, mais il apporte la réponse la plus spécifique pour la prédiction de SFFT.

Il nécessite néanmoins un minimum de six orthologues co-régulés pour la déduction d'un motif consensus de SFFT.

6 Remerciements

Nous souhaitions remercier tout particulièrement M. Sylvain Blondeau pour sa contribution dans les investigations biologiques et l'automatisation des traîtements de fichiers dans le cadre de son stage de maîtrise, et M. Laurent Noé pour les éclaircissements qu'il a pu apporter dans l'élaboration de certaines parties d'un sous-programme. Il est à noter que ces travaux n'auraient pu voir le jour sans les concours de l'ACI IMPBio (Informatique, Mathématiques et Physique pour la Biologie) et de la région Lorraine auxquelles nous exprimons toute notre reconnaissance.

Références

[1] S. Schbath. An efficient statistic software to detect over- and under-represented words in dna sequences. J. Comp. Biol., 4 :189-192, 1997. http ://www-mig.jouy.inra.fr/ssb/rmes

[2] M. Tompa, N. Li, T.L. Bailey, G.M. Church, B. De Moor, E. Eskin, A.V. Favorov, M.C. Frith, Y. Fu, W.J. Kent, V.J. Makeev, A.A. Mironov, W.S. Noble, G. Pavesi, G. Pesole, M. Régnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu. Assessing computational tools for the discovery of transcription factor binding sites. Nature Biotech., 23(1) :137-144, 2005.

[3] T.L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. of the Sec. Int. Conf. on Intel. Sys. for Molec. Biol., pages 28-36, 1994.

[4] X. Liu, D.L. Brutlag, and J.S. Liu. Bioprospector : discovering conserved dna motifs in upstream regulatory regions of co-expressed genes. Pac. Symp. Biocomput., pages 127-138, 2001.

[5] L. Marsan and Sagot M.F. Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. J. Comput. Biol., 7(3-4) :345-362, 2000.

[6] I. Uchiyama. MBGD : microbial genome database for comparative analysis. Nucleic Acids Research, 31(1) :58-62, 2003.

[7] S. Robin and S. Schbath (2005), Un motif est-il significativement plus concentré dans une séquence que dans une autre ?, En préparation.

[8] D.J. Studholme, S.D. Bentley, and J. Kormanec. Bioinformatic identification of novel regulatory dna sequence motifs in Streptomyces coelicolor. BMC Microbiology, 4(14), 2004.

[9] Touzain, F. , Debled-Rennesson, I. , Aigle, B. , Leblond, P. et Kucherov, G.. Poster at the European Conference for Computer Biology. Identification of Transcription Factor Binding Sites in Streptomyces coelicolor A3(2) by Phylogenetic Comparison, 2003.

[10] M.S.B. Paget, V. Molle, G. Cohen, Y. Aharonowitz, and M.J. Buttner. Defining the disulphide stress response in Streptomyces coelicolor A3(2) :

identification of the regulon. Molecular Microbiology, 42(4) :1007-1020, 2001.

[11] M.J. Bibb, V. Molle, and M.J. Buttner. , an extracytoplasmic function rna polymerase sigma factor required for aerial mycelium formation in Streptomyces coelicolor A3(2). Journal of Bacteriology, 182(16) :4606-4616, 2000.

[12] A. Saito, M. Ishizaka, P.B. Francisco Jr, T. Fijii, and K. Miyashita. Transcriptional co-regulation of five chitinase genes scattered on the Streptomyces coelicolor A3(2) chromosome. Microbiology, 146 :2937-2946, 2000.

[13] J.S. Hahn, S.Y. Oh, and J.H. Roe. Regulation of the fura and catc operon, encoding a ferric uptake regulator homologue and catalase-peroxidase, respectively, in Streptomyces coelicolor A3(2). Journal of Bacteriology, 182(13) :3767-3774, 2000.

[14] J.-G. Kang, M.-Y. Hahn, A. Ishihama, and J.-H. Roe. Identification of sigma factors for growth phase-related promoter selectivity of rna polymerases from Streptomyces coelicolor A3(2). Nucleic Acids Research, 25(13) :2566-2573, 1997.

[15] K.L. Brown, S. Wood, and M.J. Buttner. Isolation and characterization of the major vegetative rna polymerase of Streptomyces coelicolor A3(2) ; renaturation of a sigma subunit using groel. Mol. Microbiol., 6 :1133-1139, 1992.

[16] I. Delic, P. Robbins, and J. Westpheling. Direct repeat sequences are implicated in the regulation of two streptomyces chitinase promoters that are subjet to carbon catabolite control. Proc. Natl. Acad. Sci. USA, 89 :1885-1889, 1992.