Recherche d'éléments régulateurs communs, application aux gènes cibles des facteurs de transcription Rel/NF-κB

Matthieu Defrance[1], Hélène Touzet[1], Karo Gosselin[2], and Corinne Abbadie[2]

[1] LIFL UMR 8022 - Université Lille 1, 59 655 Villeneuve d'Ascq cedex, France
matthieu.defrance, helene.touzet@lifl.fr
[2] UMR 8117 - Institut de Biologie de Lille, Institut Pasteur de Lille, Université Lille 1, 1 rue Calmette, 59 021 Lille Cedex
karo.gosselin, corinne.abbadie@ibl.fr

Résumé. Le but de ce travail est l'analyse des régions régulatrices pour des ensembles de gènes partageant des mécanismes de régulation transcriptionnelle. Nous présentons une stratégie originale qui s'appuie sur un algorithme de recherche de sur-représentations locales des sites de fixation de facteurs de transcription. Cette approche est mise en œuvre dans un logiciel dénommé TFM-Scan, dont nous démontrons la fonctionnalité sur un jeu de données complet de gènes cibles des facteurs de transcription Rel/NF-κB chez l'homme.

1 Introduction

Ce travail s'inscrit dans la problématique générale de l'étude du contrôle de l'expression des gènes codant des protéines chez les eukaryotes supérieurs, et en particulier chez l'homme. Un des modes de ce contrôle implique la fixation de facteurs de transcription dits spécifiques au niveau des promoteurs, qui vont ainsi favoriser, ou au contraire défavoriser, le positionnement de l'ARN polymérase II sur l'ADN et l'initiation de la transcription. La fixation du facteur de transcription sur l'ADN implique un domaine de liaison à l'ADN capable de reconnaître une séquence d'ADN spécifique, longue généralement de 6 à 10 nucléotides. L'action sur l'ARN polymérase implique un domaine transactivateur capable d'interagir avec l'ARN polymérase directement, ou indirectement via des coactivateurs ou des corepresseurs. Il est admis que la nature et le nombre de facteurs de transcription fixés sur un promoteur vont déterminer la spécificité et l'intensité de transcription [13]. Une question de base se pose alors : les gènes régulés par un même facteur de transcription présentent-ils des similitudes au niveau de leur région promotrice en terme de nombre de sites de fixation pour le facteur de transcription et de position de ces sites dans le promoteur ? Cette question trouve notamment des applications pour l'analyse de puces à ADN.

La recherche *in silico* de sites de fixation de facteurs de transcription permet d'apporter des éléments de réponse à ce problème. La première approche consiste simplement à localiser les motifs approchés pouvant être reconnus par un facteur de transcription donné [20]. Les sites prédits comprennent toutefois une grande proportion de faux positifs. Cela s'explique par le faible contenu informationnel des séquences de fixation d'une part, et par la longueur des régions potentiellement régulatrices à considérer d'autre part. Dans le cas qui nous intéresse, celui de jeux de gènes régulés par un même facteur de transcription, il est possible de mener l'analyse plus finement avec la recherche de motifs sur-représentés. L'hypothèse est que les motifs présents dans les régions régulatrices et qui ont une fréquence d'occurrence exceptionnelle sont impliqués dans la régulation de

gènes. Cette approche s'est révélée féconde avec la découverte de motifs exacts ([4] pour la levure), ou de motifs approchés par maximum de vraisemblance [2] ou par échantillonnage de Gibbs ([6] pour la levure, [19] pour plusieurs modèles eucaryotes). Il peut également s'agir de la sur-représentation de motifs connus modélisés sous la forme de matrices, comme dans [1].

Dans cet article, nous présentons une stratégie de prédiction de sites de fixation de facteurs de transcription qui tire parti des positions conservées entre sites. Notre point de départ est un ensemble de sites de fixation candidats. Nous appliquons un algorithme de filtrage qui permet d'éliminer une grande partie des faux positifs. Pour cela, nous nous appuyons sur une définition locale de la sur-représentation basée sur un modèle positionel, qui se prête à l'analyse de longues séquences dont la composition est hétérogène. Les régions significatives extraites ne demandent pas d'information préalable sur la position ou sur la taille de la fenêtre d'intérêt. Cette approche est mise en œuvre dans un logiciel appelé TFM-Scan. L'algorithme de TFM-Scan est suffisamment efficace pour traiter de grands jeux de données constitués de longues séquences.

Nous illustrons notre démarche par l'étude des gènes régulés par les facteurs de transcription de la famille Rel/NF-κB. Ces facteurs sont connus pour être impliqués dans le développement de différents types de tumeurs chez l'homme, de par leur participation au contrôle des mécanismes de prolifération, apoptose et sénescence [9]. Ce sont également des régulateurs majeurs des réponses immunitaires et inflammatoires [10].

2 Analyse des régions régulatrices

La méthode que nous présentons pour l'analyse des régions régulatrices utilise une étape préliminaire avec la localisation des sites de fixation potentiels de facteurs de transcription. Nous appelons ces sites prédits des *hits*. Pour la facilité de l'exposé, nous supposons que les sites sont décrits par des matrices de fréquences positionnelles (PFM), car ce sont les modèles les plus courants et il existe de nombreux programmes permettant de localiser des PFM [5, 15]. Les hits pourraient également être obtenus par des motifs exacts, des expressions régulières, des modèles de Markov cachés, des éléments de structure secondaire conservés. C'est une donnée extrinsèque. La seule condition est que l'espérance d'occurence soit faible, inférieure à 0,05, et la méthode trouve son intérêt quand les prédictions sont peu spécifiques. Pour des PFM issues de la base de données Transfac et des séquences promotrices de plusieurs milliers de bases, le taux de faux positifs est supérieur à 90%.

Nous nous plaçons donc dans le contexte où nous disposons d'un ensemble de hits obtenus de manière systématique et associés à chaque séquence promotrice. Les hits sont repérés par leur position sur ces séquences, et éventuellement le brin, direct ou indirect. Pour prévenir les artéfacts dus aux PFM auto-chevauchantes ou palindromiques, nous ne conservons que le premier hit d'un train de hits chevauchants pour une PFM donnée. Les séquences sont alignées sur le site d'initiation de la transcription. La question posée est la suivante : existe-t-il des fenêtres où les hits d'une PFM sont sur-représentés par rapport à la densité attendue ? Pour y répondre, nous proposons une méthode organisée en triptyque : modélisation des séquences promotrices avec le choix d'un modèle de fond, algorithme d'extraction des fenêtres denses et calcul de la significativité d'une fenêtre.

Tout au long de la présentation, nous utilisons deux jeux de données de référence : 20 000 gènes humains d'une part, et 13 000 gènes de la souris d'autre part. Il s'agit plus précisément des séquences en amont, localisées de -10 000 à +1000, 0 correspondant au

site d'initiation de la transcription, extraites de [21][3]. Pour la prédiction initiale des hits, nous prenons l'ensemble des PFM de vertébrés disponibles dans Transfac 6.0 [23], soit 243.

2.1 Modèle de fond

Pour pouvoir estimer quelles sont les régions significativement denses, il faut tout d'abord se donner les moyens de décrire la distribution des hit le long de la séquence promotrice, avec le choix d'un modèle de fond. Il y a principalement deux options : choisir un modèle théorique, obtenu par apprentissage à partir des données, ou construire un modèle empirique.

Notre choix a été guidé par un impératif : tenir compte de la grande hétérogénéité de la composition des séquences. La figure 1 montre la variation du pourcentage en *GC* le long des séquences promotrices, particulièrement dans la région proximale, pour l'homme et la souris.

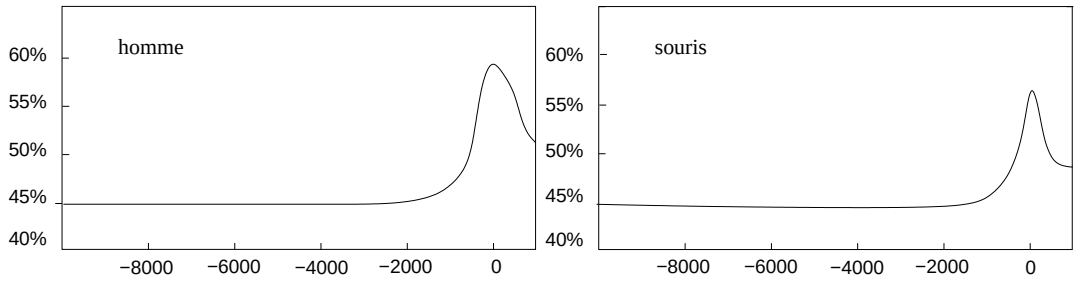

Fig. 1. Variation du pourcentage en GC. Ce graphique montre l'évolution du pourcentage en GC (en ordonnée) en fonction de la position sur la séquence (en abscisse). Les valeurs sont lissées par fenêtres de 500bp.

Ce seul pourcentage en GC ne suffit pas à expliquer la fréquence observée des différents hits le long des séquences. Nous avons tout d'abord cherché à modéliser les séquences promotrices par des modèles de Markov, en essayant différents ordres. C'est cette approche qui est retenue pour la découverte de motifs sur-représentés par échantillonage de Gibbs, par exemple [18]. Il serait ensuite facile de calculer la probabilité d'occurence d'un hit pour une PFM donnée. Mais la taille moyenne d'une PFM est d'environ 13 positions. Pour avoir une modélisation satisfaisante, il faudrait donc définir un modèle de Markov d'ordre élévé. L'hétérogénéité de la composition des séquences demande ensuite d'utiliser une combinaison de modèles de Markov, ce qui multiplie le nombre de paramètres, et rend critique la taille de l'échantillon d'apprentissage.

Face à ce constat, nous avons décidé de construire une distribution pour chacune des PFM, et de raisonner sur cette distribution. Le modèle de fond est calculé à partir d'un ensemble de séquences extraites des régions promotrices de gènes de l'organisme considéré (Figure 2). La robustesse de l'échantillon a été testée par rééchantillonnage, avec des techniques de Jackknife, systématiquement pour toutes les matrices.

[3] Pour l'homme, il s'agit de la séquence de référence, et pour la souris du *working draft*.

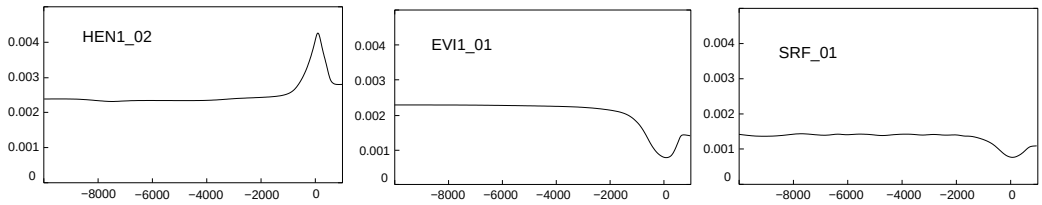

Fig. 2. Densité des hits prédits. Ces graphiques montrent trois exemples de répartition de hits pour des PFM de composition et de contenu informationnel différents pour des gènes humains. L'axe des abscisses indique la position sur la séquence relativement au site d'initiation de la transcription et l'axe des ordonnées la fréquence d'observation d'un hit à cette position.

2.2 Recherche efficace des fenêtres d'intérêt

À partir du modèle de fond, nous recherchons les fenêtres où les occurrences d'une PFM sont significativement sur-représentées. Nous ne faisons aucune hypothèse sur la localisation, ni sur la longueur des fenêtres. L'examen exhaustif de toutes les fenêtres mènerait à un algorithme quadratique par rapport à la longueur des séquences, ce qui est handicapant si on souhaite développer un outil interactif pouvant traiter de longues séquences. Nous utilisons une stratégie heuristique qui permet de détecter des fenêtres candidates en temps linéaire, à la volée. Notre point de départ est la comparaison des deux modèles : le modèle de fond, basé sur l'échantillon de référence, et le modèle cible, décrivant les fenêtres significatives que nous recherchons. Pour chaque position i, ces modèles peuvent être simplement décrits par les paramètres p_i et q_i : p_i est la probabilité d'observer un hit à la position i dans l'échantillon de référence et q_i est la probabilité d'observer un site dans une "bonne" région de l'échantillon d'étude.

À chaque position i, on associe un score $s(i)$ qui indique l'appartenance au premier ou au second modèle. $s(i)$ est positif dès lors que l'observation est plus probable dans le modèle cible que dans le modèle de fond :

$$s(i) = \log P(q_i, n, k) - \log P(p_i, n, k) \tag{1}$$

où $P(p_i, n, k)$ est la probabilité d'observer k hits à la position i pour les n séquences dans le modèle de fond, et $P(q_i, n, k)$ la probabilité du même événement dans le modèle cible décrit par q_i. Pour estimer $P(q_i, n, k)$ et $P(p_i, n, k)$, le calcul est simplifié en faisant l'hypothèse que les positions des hits sont indépendantes. Cela est raisonnable, notamment parce que nous ne considérons que les hits non chevauchants et que la distance moyenne entre deux hits est toujours beaucoup plus importante que la longueur d'un hit. La probabilité d'observer k sites à la position i pour les n séquences dans le modèle de fond décrit par p_i se calcule avec une loi binomiale :

$$\begin{aligned} P(p_i, n, k) &= \binom{n}{k} \; p_i^k (1-p_i)^{(n-k)} \\ P(q_i, n, k) &= \binom{n}{k} \; (1-q_i)^{(n-k)} \end{aligned} \tag{2}$$

En remplaçant 2 dans 1, on obtient simplement

$$s(i) = k \log \frac{q_i}{p_i} + (n-k) \log \frac{1-q_i}{1-p_i} \tag{3}$$

Il reste ensuite à déterminer p_i et q_i. Les valeurs de p_i sont naturellement calculées à partir du modèle de fond. Pour q_i, comme nous recherchons les fenêtres avec une forte densité relative, nous appliquons un ratio fixe : $r = \frac{q_i}{p_i}$. En pratique, r est fixé à 1,8 par défaut. Enfin, on extrait les fenêtres candidates en recherchant les positions où le score reste croissant (Figure 3). Pour cela, on utilise un score cumulatif, ayant 0 comme valeur plancher : $S_i = \max(0, S_{i-1} + s_i)$. Ce système est additif par construction. Il est calculé de manière incrémentale le long des séquences. Les fenêtres ainsi trouvées sont ensuite classées en fonction de leur significativité.

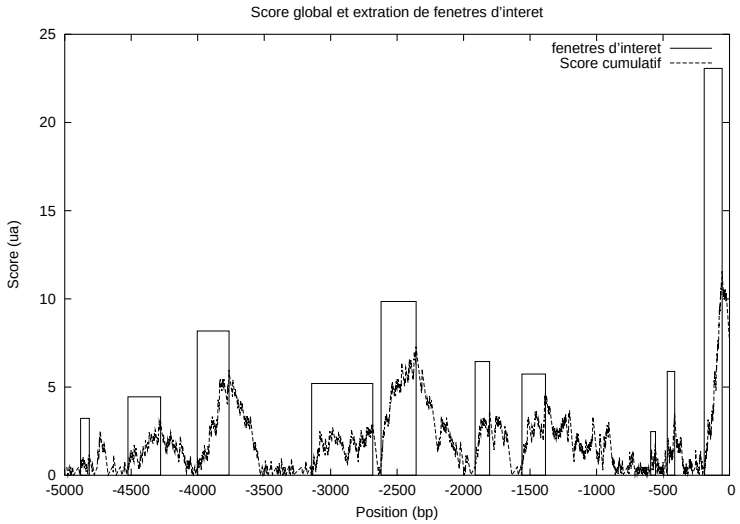

Fig. 3. Exemple de profil du score et extraction de fenêtres significatives.

2.3 Estimation de la significativité d'une fenêtre

La significativité d'une région dense en hits est déterminée en calculant sa P-valeur. La P-valeur d'une fenêtre dépend de plusieurs paramètres: les positions i et j de début et fin de la fenêtre, la matrice M, le nombre de hits k, le nombre de séquences n. On note $P(i, j, M, k, n)$ la probabilité d'observer au moins k hits pour la matrice M dans la fenêtre $[i, j]$ pour un jeu de données de n séquences. Pour estimer $P(i, j, M, k, n)$, il faut étudier la loi de probabilité du nombre d'occurrences de hits de la matrice M. Cette fois, nous raisonnons sur une fenêtre donnée, et non sur l'ensemble du jeu de séquences. Nous faisons l'hypothèse que la distribution des hits sur cette région est uniforme. La loi de comptage pour les motifs exacts sans chevauchement est bien étudiée [16]. Sa loi limite est une loi de Poisson, dès lors que l'espérance d'occurrence du motif est petite. Le seuil appliqué aux scores des PFM assure que l'espérance d'occurrence est suffisamment faible.

Pour voir dans quelle mesure cette hypothèse pouvait être étendue aux PFM, nous avons systématiquement comparé la distribution observée avec une loi de Poisson par un calcul de χ_2 pour chacune des PFM, dans les promoteurs humains (Table 1). Cela nous conduit à définir trois classes de confiance, suivant le risque associé au test de χ_2 : vert, orange et rouge. Il s'avère que pour la grande majorité des matrices, l'approximation est suffisante pour le calcul de la P-valeur. Pour les PFM problématiques, celles de la

classe rouge, la distribution présente une variance plus élevée que celle utilisée dans l'approximation de Poisson. Dans ce cas, la significativité calculée à travers la P-valeur est surestimée. À défaut d'une solution plus satisfaisante, c'est que nous avons retenu. Cette approximation devrait pouvoir être améliorée par une analyse plus individualisée, tenant de compte la variabilité de la matrice, de la taille du jeu de données et de la longueur de la fenêtre.

Ajustement Observation/Poisson par test χ^2			
α *(erreur type I)*	$[100\%, 5\%]$	$]5\%, 0.5\%]$	$]0.5\%, 0.0]$
proportion de PFM	68%	11%	20%
classe de confiance	vert	orange	rouge

Table 1. Ajustement de la distribution de la loi de comptage des hits par une loi de Poisson pour les 243 matrices de vertébrés de TRANSFAC 6.0 dans la fenêtre [-200,-1] (test du χ^2). Nous avons mesuré la distribution du nombre de hits par groupe de 20 séquences dans la fenêtre $[i, j]$ et formulé l'hypothèse (H0) suivante: La distribution des hits dans ce cas est une distribution de Poisson. Nous avons classé les matrices en 3 catégories suivant le niveau de risque pour l'hypothèse H0, en fixant deux niveaux de seuil pour α à 5.0% et 0.5% (risque de rejeter H0 à tord). Pour 168 matrices (68%), on accepte l'hypothèse H0 avec un risque d'erreur (type I) de 5.0% (stricte) Pour 27 matrices (12%), on accepte H0 avec un risque de 0.5% (souple) mais rejette avec un risque de 5.0% Pour les 48 matrices restantes (20%), on rejette H0 avec un risque aussi bas que 0.5%. Les résultats dans d'autres fenêtres, des fenêtres distales par exemples, sont analogues.

3 TFM-Scan

Nous avons mis en œuvre cette stratégie d'analyse de régions régulatrices dans un logiciel dénommé TFM-Scan (pour *Transcription Factor Module* - Scan). TFM-Scan propose actuellement deux organismes : l'homme et la souris, ce qui permet d'appliquer des recoupements par empreinte phylogénétique [22]. Ces modèles sont construits pour les régions $[-10000, +1000]$, à partir de l'intégralité des gènes annotés de l'UCSC [21] et l'intégralité des matrices de vertébrés de Transfac 6.0. La localisation des hits est faite avec le logiciel Patser [5], en cherchant sur les deux brins, mais en supprimant les chevauchements pour les hits d'une même PFM. Pour avoir un critère uniforme, les hits sont séléctionnés en fonction de leur P-valeur [17]. L'ensemble des hits est pré-calculé, sommé et stocké dans une table à deux entrées : position, facteur. Cela permet une implémentation efficace de l'algorithme de recherche de fenêtres ainsi que du calcul de la P-valeur. Pour un jeu de séquences donné et une sélection de PFM (éventuellement, l'ensemble des PFM répertoriées dans Transfac), TFM-Scan calcule les fenêtres les plus significatives : positions de la fenêtre, facteur de transcription impliqué et gènes possédant un hit dans la fenêtre. Afin d'aider l'analyse, les informations complémentaires suivantes sont également fournies :

- la valeur de la P-valeur et de la E-valeur (la E-valeur prend en compte la taille des séquences, ainsi que le nombre de matrices);
- le nombre moyen de hits par séquence dans la fenêtre, assorti de la distribution du nombre de hits par gène, comparé aux mêmes paramètres dans le modèle de fond;

– le pourcentage en GC de la fenêtre, comparé au pourcentage en GC de la matrice impliquée et au pourcentage en GC de la même région dans le modèle de référence. Cette indication permet de corriger les résultats de TFM-Scan dans le cas de biais de composition locaux par rapport au modèle de fond;
– la qualité de l'approximation par une loi de Poisson de la distribution de la PFM associée au facteur de transcription trouvé.

4 Les facteurs Rel/NF-κB et leurs gènes cibles

Pour valider notre approche, nous avons appliqué TFM-Scan à l'analyse d'un jeu de gènes régulés par les facteurs de transcription de la famille Rel/NF-κB. Nous décrivons en détail ce jeu de données, avant de présenter les résultats de TFM-Scan.

Chez les vertébrés, cinq protéines de la famille Rel/NF-κB sont connues : c-Rel, RelA (ou p65), RelB, NF-κB1 (ou p50) et NF-κB2 (ou p52). Toutes ces protéines ont en commun dans leur moitié N-terminale un domaine conservé d'environ 300 acides aminés, le Rel Homology Domain (RHD) qui représente le domaine de fixation à l'ADN. c-Rel, RelA et RelB possèdent dans leur domaine C-terminal un ou des domaines transactivateurs. NF-κB1 et NF-κB2 ne possèdent pas de domaines transactivateurs. Les facteurs Rel/NF-κB sont actifs sous forme de dimères. L'interface de dimérisation se situe au niveau du RDH. Pratiquement toutes les combinaisons d'homodimères ou d'hétérodimères entre les différents membres sont possibles. Seul RelB n'est capable que de s'hétérodimériser avec NF-κB1 ou NF-κB2. Ces deux derniers facteurs peuvent agir comme des répresseurs passifs sous forme d'homodimères, puisqu'ils sont dépourvus de domaine transactivateur, ou sous forme de co-facteur sous forme d'hétérodimère avec un autre membre de la famille [3]. Les facteurs Rel/NF-κB se fixent sur l'ADN sous forme dimérisée en reconnaissant spécifiquement des sites κB, dont la séquence consensus, un quasi-palindrome, est `5'-GGGRNYYYCC-3'` où R est une purine, Y une pyrimidine et N un nucléotide quelconque. Cette affinité est particulièrement élevée (10-10 à 10-13 M pour l'hétérodimère p50/RelA), ce qui distingue cette famille de nombreux autres facteurs de transcription. Bien sûr, l'affinité pour l'ADN des différents dimères varie en fonction de la séquence exacte du motif κB qui varie d'un promoteur à l'autre. Par exemple, les homodimères p50 ont une meilleure affinité pour les sites parfaitement palindromiques qui permettent la liaison de chaque protéine à un demi-site, tandis que les hétérodimères p50/RelA préfèrent des sites non totalement symétriques [3]. Dans la plupart des types cellulaires, les dimères Rel/NF-κB sont présents constitutivement sous une forme inactive retenue dans le cytoplasme par une protéine inhibitrice de la famille IκB qui masque leur signal de localisation nucléaire et empêche leur liaison à l'ADN. La stimulation de la cellule déclenche une cascade de réactions conduisant à une phosphorylation suivie d'une ubiquitination et dégradation d'IκB par le protéasome. Les dimères Rel/NF-κB ainsi libérés peuvent migrer dans le noyau, se fixer à l'ADN et moduler la transcription de leurs gènes cibles. Ce mécanisme permet donc une mobilisation rapide de ces facteurs en réponse à un signal [11].

Nous avons construit une liste de gènes cibles des facteurs de transcription de la famille Rel/NF-κB. Cette compilation a été établie par recherche bibliographique d'articles présentant des résultats expérimentaux qui démontrent l'implication d'un ou plusieurs sites κB dans la régulation de l'expression d'un gène donné. Nous avons utilisé pour cela trois sources d'information : une liste de gènes cibles des facteurs Rel/NF-κB publiée

dans l'article de revue [12], une seconde liste disponible sur internet sur le site *Rel/NF-kB transcription factors* [7], et des recherches complémentaires grâce à la base de données bibliographiques PUBMED [14]. Nous avons concentré nos efforts sur les gènes humains. Nos critères de sélection pour considérer qu'un gène est un gène cible vérifié des facteurs Rel/NF-κB ont été de deux types.

La première possibilité est l'existence d'une étude de promoteur qui permet de démontrer l'importance de la présence et de l'intégrité d'un ou plusieurs sites κB dans la régulation de l'expression du gène considéré. Généralement ces études emploient des tests de transactivation dans lesquels la partie de promoteur contenant le ou les sites κB potentiellement fonctionnels sont clonés devant un promoteur minimal et un gène rapporteur; la mesure de l'expression du gène rapporteur reflète l'importance du ou des sites pour l'induction ou la répression de la transcription; cette expression doit être modifiée lorsque le ou les sites sont délètés et/ou mutés. Ces expériences sont souvent complétées par des expériences de retard sur gel qui permettent de montrer que, *in vitro* au moins, les facteurs Rel/NF-κB sont bien capables de se lier sur la séquence exacte du ou des sites κB potentiellement fonctionnels.

Le second cas est la réalisation d'expériences de précipitation de la chromatine qui permettent de déterminer si *in vivo*, les facteurs Rel/NF-κB sont bien liés à un ou des sites κB repérés à une position précise de la région promotrice du gène étudié. Ces résultats ne peuvent être pris en compte que s'ils sont accompagnés d'expériences à l'échelle cellulaire telles que l'induction de la transcription du gène par un stimulus connu pour induire une activité Rel/NF-κB et/ou l'inhibition de la transcription du gène lorsqu'on inhibe l'activité Rel/NF-κB par des inhibiteurs pharmacologiques ou la surexpression d'IκB.

La compilation obtenue compte 100 gènes humains. Chaque gène est référencé par son identifiant RefSeq, et est accompagné des informations complémentaires suivantes : la fiche PUBMED de l'article mettant en évidence le contrôle par un facteur Rel/NF-κB, l'identifiant OMIM, la position et la séquence exacte des sites vérifiés expérimentalement, la séquence en amont [-10 000, +1000] et le gène orthologue chez la souris. Cette liste est complétée par un ensemble de 34 gènes cibles potentiels. Il s'agit de gènes pour lesquels les résultats expérimentaux suggèrent très fortement une régulation de leur expression par les facteurs de transcription de la famille Rel/NF-κB, mais pour lesquels une étude de promoteur complète n'a pas été réalisée, ou de gènes pour lesquels l'implication directe de NF-κB a été mise en évidence pour un orthologue.

5 Validation de TFM-scan par l'analyse des promoteurs des gènes cibles Rel/NF-κB

Nous avons appliqué TFM-Scan à l'ensemble des gènes cibles humains vérifiés de Rel/NF-κB. Transfac compte six PFM pour les facteurs Rel/NF-κB, qui reflètent les spécificités des facteurs de la famille : `CREL_01`, `NFKAPPA50_01`, `NFKAPPAB65_01`, `NFKB_Q6`, `NFKAPPAB_01` et `NFKB_C`. La recherche a été faite de manière exhaustive pour l'intégralité des PFM de Transfac, pour des séquences de -10 000 à +1000. Nous avons sciemment inclus des régions distales dans l'analyse, à titre de contrôle. Les résultats sont indiqués pour les dix fenêtres les plus significatives, tous facteurs confondus, en Table 2.

Parmi les six meilleures fenêtres, cinq concernent des matrices Rel/NF-κB, avec des P-valeur allant de 7.09×10^{-12} à 8.70×10^{-11}. Nous avons testé la robustesse de ce résultat sur des ensembles de même taille de gènes sélectionnés aléatoirement, sans retrouver de fenêtre dont la P-valeur était inférieure à 10^{-8}. Les fenêtres ainsi détectées ont une

matrice (PFM)	fenêtre	nbre de hits	nbre de séquences	P-valeur	E-valeur
TATA_01	[-74: -9]	42	42	3.57e-14	9.55e-05
NFKB_C	[-507: -16]	204	87	1.71e-13	4.57e-04
NFKAPPAB_01	[-624: -17]	237	92	3.32e-13	8.87e-04
NFKAPPAB65_01	[-520: -13]	190	83	6.51e-13	1.74e-03
NFKB_Q6	[-696: -20]	235	91	1.65e-11	4.41e-02
CREL_01	[-511: -17]	173	75	1.58e-10	4.22e-01
TATA_C	[-60: +42]	39	35	3.61e-10	9.65e-01
RREB1_01	[-4382:-3855]	240	88	7.05e-10	1.88e+00
NFAT_Q6	[-231: -16]	98	62	6.13e-08	1.64e+02
CDXA_02	[-5849:-5523]	100	49	1.25e-07	3.34e+02

Table 2. Résultats de TFM-Scan pour les gènes cibles de Rel/NF-κB chez l'Homme.

longueur variable, mais ont des positions similaires et qui se recoupent dans le promoteur. En globalisant les données, on peut limiter cette fenêtre aux nucléotides -16 à -520. Il s'agit donc d'une fenêtre proximale. Les hits détectés par les différentes PFM ne sont que partiellement redondants. 91% des gènes présentent des sites κB dans cette fenêtre, avec une densité moyenne de 0.344 sites pour 100 pb, avec un maximum pouvant dépasser 6 sites par fenêtre. La position proximale des fenêtres prédites est conforme aux informations disponibles concernant les sites connus, puisque sur les 149 sites identifiés expérimentalement que comprend au total la liste des gènes cibles vérifiés, 110 sites se retrouvent dans la fenêtre [-520, -17], soit 74%. La localisation des sites résiduels est très variable : certains sont situés plus en amont du site d'initiation de la transcription, jusqu'à -4000, et d'autres après le site d'initiation de la transcription, avant ou après le codon start, dans un exon ou même dans la région flanquante 3'. Par contre, sur ces 110 sites κB de la fenêtre [-520, -17], seuls 60 sont effectivement prédits par le logiciel. Deux raisons peuvent être avancées pour expliquer cela. Premièrement, la localisation des sites d'initiation de la transcription retenue par les expérimentateurs peut ne pas concorder avec l'annotation proposée par l'UCSC. Deuxièmement, certaines séquences des sites κB déterminées expérimentalement sont considérablement dégénérées par rapport aux consensus reconnus par les PFM, et échappent à l'étape initiale de localisation des hits potentiels.

TFM-Scan détecte également, en première et en huitième positions, une fenêtre avec une sur-représentation importante en boîtes TATA. Les boîtes TATA sont des sites présents en amont des sites d'initiation de la transcription en position -25 à -35. Ils permettent la fixation de facteurs généraux de transcription, dont la TBP et TFIID, qui facilitent le positionnement de l'ARN polymérase II sur le site d'initiation de la transcription. Tous les gènes ne contiennent pas de boîte TATA. Dans ce cas, le positionnement de l'ARN polymérase requiert la présence de séquences riches en GC ou des boîtes CAAT, mais le site d'initiation de la transcription est alors moins précis. Il est admis que les gènes à boîte TATA sont des gènes facilement et rapidement activables, tandis que les gènes dépourvus de boîte TATA seraient transcrits à des taux plus bas et plus constants [13]. La fenêtre riche en boîtes TATA la plus significative est bien localisée par TFM-Scan. Elle est délimitée de -59 à -9 et contient toujours un seul site. 38% des gènes cibles Rel/NF-κB présentent une boîte TATA non dégénérée. C'est une particularité significative, qui est attendue pour des cibles de facteurs rapidement activables.

Les PFM suivantes détectées par TFM-Scan sont moins significatives. Elles présentent des E-valeurs supérieures à 1. Par exemple, apparaiten septième position une fenêtre enrichie en hits pour le facteur RREB, avec une E-valeur affichée égale à 1,88. La fenêtre est très distale, de -4832 à -3855, et présente 1 à 3 hits par gène. Aucune donnée dans la littérature ne permet d'attendre la présence de sites pour ce facteur dans les gènes régulés par Rel/NF-κB. De plus, la matrice RREB-1 est effectivement une des PFM pour laquelle le calcul de la P-valeur est sur-estimé. Il est fort probable que ce résultat soit un faux positif, qui s'explique comme un artefact de TFM-Scan.

Nous avons également analysé le jeu de données des gènes cibles de NF-κB avec les principales approches de recherche de motifs sur-représentés. La forme en diade du site de fixation fait que la détection des sites κB se prête mal à une approche par motif exact ou par Gibbs sampling, qui effectivement ne donnent pas de résultat. Nous avons également appliqué MotifScanner de la suite Toucan [1], qui fonctionne comme TFM-Scan en recherchant des sites de fixation produits à l'aide de PFM. Les résultats sont reproduits en Table 3. Quand on considère l'intégralité du jeu de données, ou même des régions plus courtes, plusieurs facteurs sont prédits avec une significativité élevée. Mais aucune prédiction ne correspond au résultat attendu. Cela pourrait s'expliquer par la taille de la région et l'utilisation d'un modèle uniforme. Il faut réduire la fenêtre à la région [-500,-1] pour voir apparaitre les facteurs de la famille NF-κB. Les boîtes TATA, quant à elles, ne sont pas détectées. Les autres facteurs prédits sont associés à des PFM riches en GC. On voit là la sensibilité d'une approche locale, par rapport à une approche globale.

MotifScanner : fenêtre [-5000, + 1000]		
matrice (PFM)	nbre de hits	P-valeur
M00081-V$EVI1_04	151	0.0
M00096-V$PBX1_01	124	0.0
M00116-V$CEBPA_01	75	0.0
M00129-V$HFH1_01	95	0.0
M00131-V$HNF3B_01	157	0.0
M00145-V$BRN2_01	56	0.0
M00148-V$SRY_01	55	0.0
M00155-V$ARP1_01	77	0.0
M00159-V$CEBP_01	88	0.0

MotifScanner : fenêtre [-500, -1]		
matrice (PFM)	nbre de hits	P-valeur
M00196-V$SP1_Q6	80	0.0
M00052-V$NFKAPPAB65_01	50	5.6-13
M00054-V$NFKAPPAB_01	47	6.7-13
M00208-V$NFKB_C	29	2.1-11
M00053-V$CREL_01	42	3.2-11
M00194-V$NFKB_Q6	37	5.3-11
M00051-V$NFKAPPAB50_01	19	5.7-6
M00255-V$GC_01	52	5.0-4
M00063-V$IRF2_01	9	0.02

Table 3. Résultats de MotifScanner pour les gènes cibles de NF-κB chez l'Homme. Le modèle de fond est basé sur un modèle de Markov d'ordre 3 construit à partir d'un échantillon d'apprentissage fourni par EPD. Ce sont les paramètres recommandés.

Disponibilité et perspectives

TFM-Scan dispose d'un site web, à l'URL `http://bioinfo.lifl.fr/TFM-Scan`, et la compilation des gènes cibles des facteurs de transcription de la famille Rel/NF-κB est accessible à `http://bioinfo.lifl.fr/NF-KB`.

Dans l'état actuel, la recherche de sites potentiels se fait matrice par matrice, de manière indépendante. Cela pose deux limites. D'une part, nous ne pouvons pas traiter globalement des familles de matrices. Cela supposerait de prendre en compte l'information

mutuelle entre PFM. D'autre part, nous ne gérons pas les associations entre facteurs, à moyenne ou à longue portée. Cette approche permettrait d'augmenter la sensibilité, tout en réduisant les faux positifs.

matrice (PFM)	fenêtre	nbre de hits	nbre de séquences	P-valeur	E-valeur
NFKAPPAB_01	[-726: -46]	169	70	2.02e-12	5.40e-03
CREL_01	[-757: -56]	170	62	1.04e-11	2.77e-02
NFKAPPAB65_01	[-892: -9]	190	66	7.23e-11	1.93e-01
NFKB_Q6	[-735: -57]	164	64	1.88e-10	5.03e-01
AP2_Q6	[-9431:-7755]	281	64	1.68e-09	4.50e+00
NFKB_C	[-436: -19]	118	56	2.58e-09	6.89e+00
TATA_01	[-59: +230]	45	42	1.86e-08	4.97e+01

Table 4. Résultats de TFM-Scan pour les gènes orthologues chez la Souris.

Enfin, le nombre croissant de génomes eucaryotes disponibles permet également de tirer partie des informations conservées par l'évolution, par empreinte phylogénétique. Sur l'exemple des gènes cibles humains de NF-κB, nous avons recherché les gènes orthologues chez la Souris (grâce à Homologène, NCBI) et appliqué TFM-Scan avec un modèle de fond construit à partir du génome de la Souris: on retrouve bien les facteurs de la famille NF-κB (table 4). Il serait souhaitable de pouvoir combiner les modèles d'espèces différentes, pour mettre en lumière les mécanismes communs.

Remerciements

Ces travaux ont été financés par les CNRS, l'Université Lille 1, l'ACI Jeunes Chercheurs 2003, l'ARC, la Ligue contre le Cancer (comité du Nord), l'Institut Pasteur de Lille, le conseil Régional Nord-Pas de Calais et le Fond Européen de Développement Régional. Karo Gosselin est financée par l'Institut Pasteur de Lille et la Région Nord-Pas de Calais.

References

1. Aerts, S., Thijs, G., Coessens, B., Staes, M., Moreau Y., and De Moor,. B. TOUCAN: deciphering the cis-regulatory logic of coregulated genes. Nucleic Acids Research, 31(6), 1753-1764, 2003
2. Bailey, T.L., and Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, 1994.
3. Chen, F. E., and Ghosh, G. Regulation of DNA binding by Rel/NF-kappaB transcription factors: structural views. Oncogene 18, 6845-6852, 1999.
4. van Helden, J., Andre, B., and Collado-Vides, J.: Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotides frequencies. Journal of Molecular Biology 281(5), 827-842, 1998
5. Hertz, GZ., Hartzell III, GW., Stormo, GD. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. Computer Applications in the Biosciences 6(2): 81-92, 1990)
6. Hughes, JD, Estep, PW, Tavazoie S, Church GM. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae, Journal of Molecular Biology, 296(5):1205-14, 2000.
7. Gilmore, TD. `http://people.bu.edu/gilmore/nf-kb/index.html`
8. Karin, M. and Lin A. (2002), NFkB at the crossroads of life and death Nature Immunology 3 (3), 221-227

9. Lin, A., and Karin, M. NF-kappaB in cancer: a marked target. Semin Cancer Biol 13, 107-114, 2003
10. Liou, H. C. Regulation of the immune system by NF-kappaB and IkappaB. J Biochem Mol Biol 35, 537-546, 2002
11. Mercurio, F., and Manning, A. M. Multiple signals converging on NF-kappaB. Curr Opin Cell Biol 11, 226-232, 1999
12. Pahl, H. L. Activators and target genes of Rel/NF-kB transcription factors. Oncogene 18, 6853-6866, 1999
13. Patikoglou, G., and Burley, S. K. Eukaryotic transcription factor-DNA complexes. Annu Rev Biophys Biomol Struct 26, 289-325, 1997
14. Pubmed, `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed` National Center for Biotechnology Information
15. Quandt, K., Frech, K., Karas, H., Wingender, E., Werner, T. MatInd and MatInspector - New fast and versatile tools for detection of consensus matches in nucleotide sequence data Nucleic Acids Research 23, 4878-4884, 1995
16. Robin, S., Rodolphe, F., Schbath, S. ADN, mots et modèles, édition Belin, collection Échelle, 2003
17. Staden R, Methods for calculating the probabilities of finding patterns in sequences. CABIOS 5 89-96, 1989
18. Thijs G., Lescot M., Marchal K., Rombauts S., De Moor B., Rouzé P., Moreau Y. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics, 17 - 12, 1113-1122, 20O1
19. Thijs G, Marchal K, Lescot M, Rombauts S, DeMoor B, Rouze P, Moreau Y. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. J Comput Biology, 9(2):447-64, 2002
20. Tronche F., Ringeisen F., Blumenfeld M., Yaniv M., Pontoglio M. Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. Journal of Molecular biology, 266, 231-245, 1997
21. UCSC Genome Bioinformatics (University of California, Santa Cruz) , `http://genome.ucsc.edu/`
22. Wasserman WW, Palumbo M, thompson W, Fickett JW, Lawrence CE: Human-mouse genome comparisons to locate regulatory sites. Nature Genetics, 26:225-228, 2000
23. Wingender, E., Chen, X., Hehl, R., Karas, I., Liebich I., Matys V., Meinhardt T., Pruss M., Reuter I., and Schacherer F. TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Research, 28, 316-31, 2000