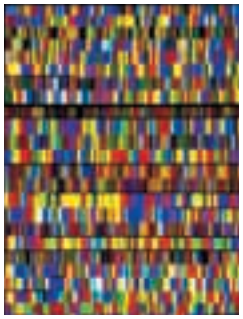
MEDECINE/SCIENCES 2004 ; 20 : 1036-40

Prédire la transcription à partir des séquences génomiques

David Martin, Badih Ghattas, Denis Thieffry

> Les développements technologiques récents ont grandement facilité et accéléré l'obtention des séquences nucléiques à l'échelle génomique. À partir de ces séquences, les bio-informaticiens tentent de délimiter les régions fonctionnelles, y compris bien sûr les gènes, mais aussi les motifs et les conditions qui contrôlent leur expression. Dans un article récemment publié dans *Cell*, M.A. Beer et S. Tavazoie combinent une méthode de classification (*clustering*) de données de transcriptome (puces à ADN), un logiciel de découverte de motifs *cis*-régulateurs, ainsi qu'une méthode d'apprentissage probabiliste pour inférer des règles susceptibles de rendre compte des profils d'expression transcriptionnelle observés. <



Laboratoire de Génétique et de physiologie du développement, LGPD-IBDM, CNRS, Case 907, Université de la Méditerranée, Campus Scientifique de Luminy, 13288 Marseille Cedex 9, France.
thieffry@ibdm.univ-mrs.fr

Dans la foulée du séquençage de génomes complets, la conception et l'exploitation de nouvelles approches expérimentales à haut débit, en combinaison avec de nouvelles méthodes mathématiques et informatiques, ouvrent la voie au décryptage des réseaux de régulation contrôlant les processus cellulaires. Parmi les différents niveaux de régulation impliqués, les mécanismes de contrôle de la transcription sont particulièrement importants et étudiés de manière intensive. Les outils de génomique fonctionnelle les plus répandus - les puces à ADN - visent précisément à caractériser les niveaux d'ARN messagers produits au sein des cellules, des tissus ou des organismes étudiés, dans différentes conditions de culture, situations pathologiques ou contextes génétiques [1] (➜).

(➜) m/s 2004, n° 4, p. 487

Sur la base de jeux de données de transcriptome et des séquences génomiques, il s'agit d'arriver à inférer les mécanismes de régulation transcriptionnelle sous-jacents. Une approche qui a donné de bons résultats chez plusieurs micro-organismes (levure, colibacille) consiste à classer les profils transcriptionnels en classes relativement homogènes (*clustering*), pour ensuite rechercher des motifs (oligonucléotides ou matrices consensus) sur-représentés de manière significative dans les régions promotrices des gènes d'une même classe [2]. Souvent, les motifs ainsi mis en évidence correspondent à des sites de fixation connus pour des facteurs de transcription. D'autres motifs sont inconnus, mais éventuellement susceptibles de fixer des facteurs transcriptionnels par des mécanismes encore méconnus.

Plusieurs hypothèses sous-tendent ce type de travaux bio-informatiques. D'une part, les motifs recherchés doivent se trouver principalement à proximité des sites d'initiation de la transcription ou, lorsque ces derniers sont méconnus (comme dans le cas de la levure), dans une région bien délimitée en amont de la première fenêtre de lecture de chaque unité transcriptionnelle. Ces régions s'étendent typiquement de quelques centaines à un millier de paires de bases pour les micro-organismes les plus étudiés. D'autre part, les motifs étant généralement extraits séparément, le signal permettant à un facteur de transcription de reconnaître la plupart des régions promotrices associées à un profil transcriptionnel doit être suffisamment fort, suggérant un mécanisme de régulation relativement simple,

dominé par un seul ou un petit nombre de facteurs de transcription.

Dans le cas des organismes eucaryotes pluricellulaires, les connaissances encore très partielles des mécanismes de régulation transcriptionnelle contredisent clairement ces deux hypothèses. En effet, chez les animaux et les plantes modèles étudiés, qui sont dotés de régions non codantes beaucoup plus étendues *(Tableau I)*, plusieurs études montrent que les éléments *cis*-régulateurs sont nettement plus dispersés et surtout beaucoup plus complexes, combinant de nombreux sites de fixation (plusieurs dizaines) pour plusieurs facteurs de transcription différents (facilement une demi-douzaine). Cette situation constitue un véritable défi pour les bio-informaticiens et les biostatisticiens qui, en interaction étroite avec les biologistes expérimentaux, s'attèlent au développement de méthodes susceptibles de permettre la délimitation des régions non codantes effectivement impliquées dans la régulation transcriptionnelle, ainsi que l'extraction de règles d'organisation de combinaisons de motifs pour former des modules de régulation transcriptionnelle susceptibles de rendre compte de la fixation coopérative de plusieurs facteurs de transcription.

À cet égard, dans un numéro récent de la revue *Cell*, M.A. Beer et S. Tavazoie ont publié un protocole d'analyse innovant, combinant une méthode de classification (*clustering*) des données de transcriptome, un logiciel de découverte de motifs, ainsi qu'une méthode d'apprentissage probabiliste permettant la définition de règles logiques pour la combinaison de motifs de régulation élémentaires [3]. D'abord appliquée à la levure (*S. cerevisiae*), puis, de manière plus prospective, au nématode (*C. elegans*), cette méthode a conduit à la prédiction de profils d'expression transcriptionnelle directement comparables avec les profils expérimentaux initialement analysés. Enfin, elle a permis d'inférer des hypothèses précises sur des mécanismes de régulation transcriptionnelle, en principe susceptibles d'être vérifiées expérimentalement.

Organisme	Taille du génome	Nombre de gènes	ADN non codant
Colibacille (*E. coli*)	~ 4,6 Mb	~ 4 300	~ 10 %
Levure (*S. cerevisiae*)	~ 13 Mb	~ 6 200	~ 30 %
Nématode (*C. elegans*)	~ 100 Mb	~ 19 100	~ 75 %
Mouche (*D. melanogaster*)	~ 180 MB	~ 13 600	~ 85 %
Vertébrés (*H. sapiens*)	~ 3 200 Mb	~ 24 000	~ 98,5 %

***Tableau I.** Quelques génomes modèles, avec le nombre de gènes prédits, la taille du génome et le pourcentage d'ADN non codant correspondant.*

Une approche génomique systématique pour l'apprentissage de règles régulatrices combinatoires

Les différentes étapes du protocole d'analyse publié par M.A. Beer et S. Tavazoie [3] sont schématiquement décrites dans la *Figure 1*. La première étape consiste à classer les profils transcriptionnels obtenus au cours d'une série de mesures à l'aide de puces à ADN *(Figure 1A)*, en utilisant la méthode du *K-means (Figure 1B)* [4]. Cette méthode nécessite la détermination préalable du nombre de classes recherchées et éventuellement le choix d'un centre initial pour chaque classe. Il en résulte un nombre prédéterminé de classes, dont la cohérence est maximisée en fonction d'un critère de distance par rapport au centre de chaque classe. Les auteurs ont ajouté à cette technique classique un seuil de distance de manière à éliminer les éléments les plus marginaux de chaque classe.

Pour chaque classe ainsi obtenue, l'utilisation d'un logiciel de découverte de motifs permet de mettre en évidence les régions d'ADN non codantes conservées, susceptibles d'expliquer la co-expression de ces gènes [5-7]. Les auteurs ont appliqué l'implémentation d'une variante de la méthode d'échantillonnage aléatoire de Gibbs, le logiciel *AlignACE*, qui permet la mise en évidence de plusieurs sites partiellement conservés dans un jeu de séquences [2]. Le logiciel produit des alignements locaux multiples qui sont modélisés sous la forme de matrices poids-positions, chacune accompagnée d'une estimation probabiliste rigoureuse. Le contenu de cette matrice peut également être représenté sous la forme d'un *logo* (*voir* partie inférieure de la *Figure 1C*), représentant de manière intuitive la diversité (différentes bases) et la conservation (hauteur de chaque lettre, quantifiée en *bits*, avec un maximum de deux *bits* par position) tout au long du motif.

L'originalité de l'approche de M.A. Beer et S. Tavazoie tient surtout dans l'intégration de ces différents signaux, accompagnée d'une caractérisation spatiale simple de ces motifs au sein d'un modèle probabiliste, appelé « réseau bayésien » [8]. Le principe consiste à coder - sous forme d'un graphe acyclique orienté - des liens entre différentes caractéristiques des gènes étudiés : les profils d'expression (« discrétisés ») et l'occurrence de motifs ordonnés dans les régions promotrices des gènes correspondants. Ce graphe est constitué de trois types d'éléments : des nœuds, qui correspondent à certaines caractéristiques des gènes, des arcs unidirectionnels, qui illustrent les liens entre les caractéristiques, et des probabilités conditionnelles qui quantifient ces liens *(Figure 1D)*. Chaque arc (flèche) représente ainsi un lien de causalité entre un profil d'expression donné (cible de l'arc) et une caractéristique des données (source de l'arc), par exemple l'occurrence d'un motif au-dessus d'un seuil, la position par rapport à l'ATG, ou la position relative par rapport à un autre motif. Si la détermination exacte du réseau bayésien le plus optimal pour rendre compte des données est largement hors de portée d'analyse (explosion

combinatoire du nombre de modèles possibles), il est en général possible de développer une approche heuristique qui prend progressivement en compte un ensemble de caractéristiques différentes au sein du réseau.

Pour un groupe de gènes (classe ou *cluster*), la stratégie utilisée considère d'abord un premier motif: (1) en sélectionnant la matrice poids-positions correspondant à la meilleure prédictibilité pour le profil correspondant; (2) en déterminant le seuil à partir duquel une séquence est considérée comme contenant un motif pour cette matrice poids-positions; (3) en définissant les positions préférentielles par rapport à l'ATG; (4) en sélectionnant l'orientation la plus significative. Toujours pour le même groupe de gènes, un second motif est alors sélectionné, en prenant en compte successivement les mêmes critères. Enfin, la position relative et l'ordre de ces motifs sont ultimement pris en compte. En bout de course, des règles d'organisation logique sont ainsi obtenues et associées aux différents profils d'expression.

Afin d'évaluer le gain prédictif de leur méthode, les auteurs ont effectué un grand nombre de tests pour chaque classe, en paramétrant le réseau bayésien sur la base d'un sous-ensemble des gènes (80 %) et en testant le pouvoir prédictif du réseau paramétré résultant sur le reste des gènes (20 %) de la classe. Une analyse systématique du pouvoir prédictif des modèles bayésiens a également été effectuée en comparaison avec le pouvoir prédictif de modèles n'intégrant que la simple présence de motifs, ou encore fondés sur des échantillons aléatoires de gènes (*Figure 2B*). En général, le gain de corrélation entre les combinaisons de motifs *cis*-régulateurs ainsi identifiées et les profils d'expression pour les gènes correspondants est substantiel (on passe d'une corrélation de 0,36 pour les motifs simples à une corrélation de 0,51 pour la modélisation bayésienne complète).

Cette approche a été appliquée à deux jeux de données publics. Dans un premier temps, les auteurs ont testé leur méthode sur un jeu de données de transcriptome de levure, combinant des expériences sur le stress environnemental et le cycle cellulaire, comportant 255 conditions au total. Ils ont ainsi sélectionné 2587 gènes regroupés en 49 classes, clairement enrichies en annotations fonctionnelles spécifiques. L'analyse des régions *cis*-régulatrices a été restreinte aux 800 nucléotides en amont de l'ATG de ces gènes, ce qui a conduit à la découverte d'environ 600 motifs. L'exploration systématique des combinaisons de caractéristiques *cis*-régulatrices mentionnées ci-dessus a ensuite permis la mise en évidence de règles combinatoires précises, en relation avec des sauts qualitatifs dans le pouvoir prédictif de l'expression génique. Ces règles portent non

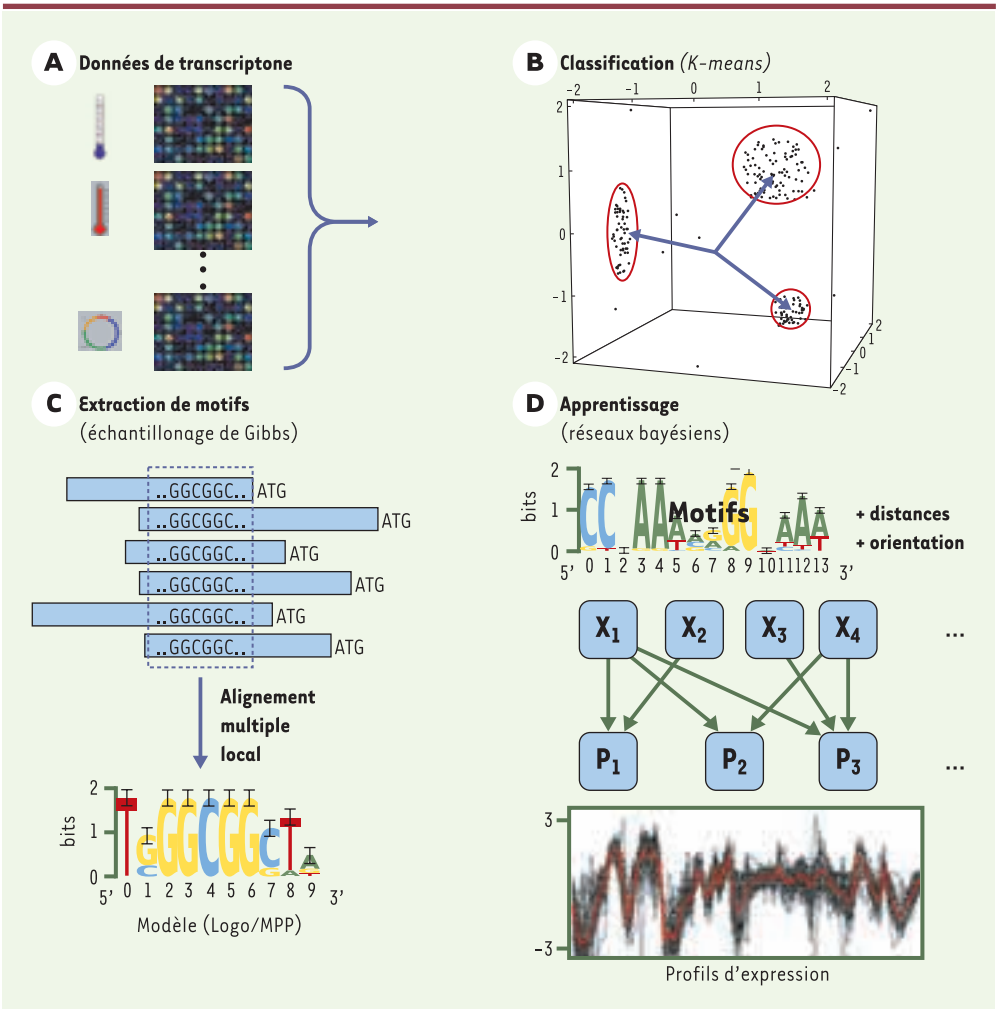


Figure 1. ***Principales étapes de la méthode d'inférence de règles de régulation transcriptionelle à partir des données de transcriptome.***

seulement sur la définition du nombre et de la nature des motifs trouvés, mais également sur leurs ordre, espacement et distance par rapport à l'ATG. Il devrait donc être relativement aisé de tester expérimentalement la pertinence de telles règles, au moins dans le cas de la levure. Globalement, les auteurs estiment le pouvoir prédictif de l'ensemble de ces règles à environ 73% des gènes considérés dans cette étude.

Pour évaluer le potentiel de leur approche dans le cas d'eucaryotes multicellulaires, les auteurs ont exploité un jeu de données Affymetrix combinant 20 points temporels au cours du développement embryonnaire du nématode. Ici, des régions en amont de 2000 nucléotides ont été analysées pour 5 547 gènes significativement exprimés et classés en 30 groupes. Pour plusieurs groupes de gènes co-exprimés, des règles portant sur la présence et les distances relatives de plusieurs motifs ont ainsi été établies. De manière globale, les auteurs estiment avoir pu prédire qualitativement les profils d'expression d'environ la moitié des gènes considérés, malgré la limitation des régions non codantes prises en compte.

De manière générale, les auteurs concluent que les règles logiques mises en évidence impliquent une redondance importante dans les modes de régulation transcriptionnelle (opérateur logique OU). Cependant, de nombreux facteurs fonctionnent de manière synergique (opérateur logique ET). L'absence d'un facteur (opérateur logique NON) est aussi très souvent nécessaire pour permettre l'établissement d'un mode de régulation. Enfin, dans bien des cas, l'orientation d'un site ou l'orientation relative de deux sites, ainsi que les distances relatives entre sites se sont révélées également importantes. Ces règles générales simples justifient ainsi les hypothèses à la base de bon nombre de travaux de modélisation dynamique qualitative des réseaux de régulation génétique [9].

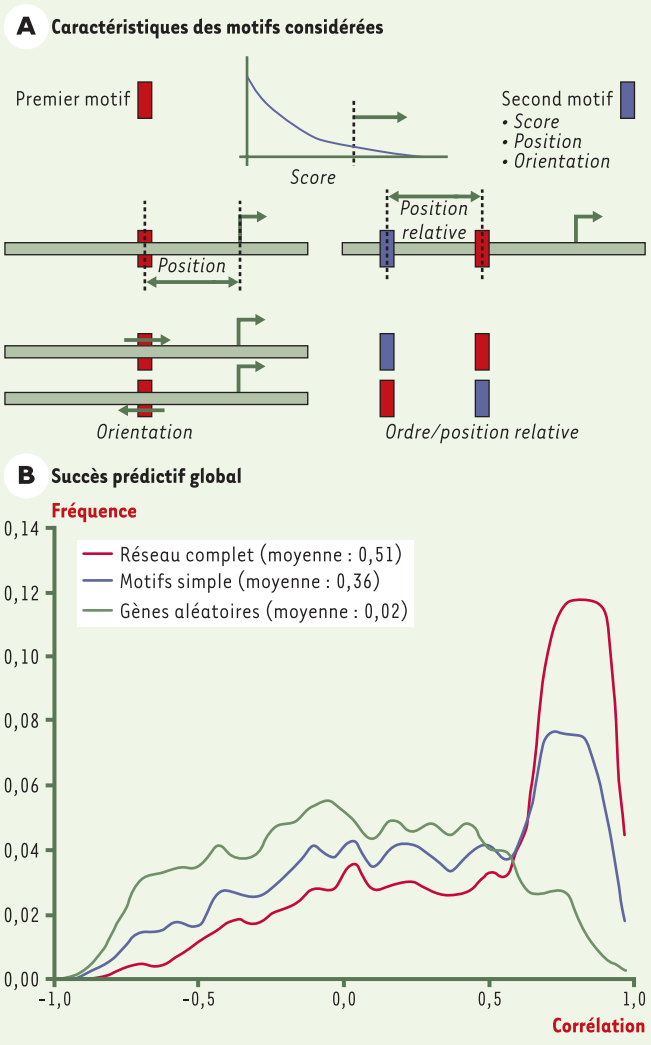


Figure 2. ***Caractéristiques considérées pour les motifs de régulation transcriptionnelle trouvés dans les régions promotrices et impact sur la corrélation avec les profils d'expression étudiés.***

Conclusions et perspectives

L'étude de M.A. Beer et S. Tavazoie combine l'exploitation de données génomiques (séquences et annotations) et post-génomiques (transcriptome) avec des approches bio-informatiques et mathématiques (découverte de motifs, réseaux bayésiens) relativement classiques. Son originalité réside surtout dans les détails du déploiement et de l'articulation des différents éléments utilisés. En particulier, la méthode heuristique d'apprentissage fondée sur les réseaux bayésiens permet d'intégrer de manière cohérente un ensemble sophistiqué de caractéristiques des régions *cis*-régulatrices: combinaisons significatives de motifs, orientations et distances relatives. Du point de vue mathématique, des efforts sont encore nécessaires pour proposer une méthode rigoureuse pour la définition des différents paramètres de la modélisation (par exemple, les seuils de « discrétisation » des données de transcriptome, ou encore le nombre de classes considérées). Du point de vue biologique, il est clair que l'approche utilisée a un certain nombre de limitations, compliquant son application à des organismes aussi complexes que les mammifères. En effet, comme nous l'avons déjà mentionné plus haut, les données disponibles suggèrent que les régions *cis*-régulatrices impliquées sont vraisemblablement beaucoup plus complexes et agissent parfois à de grandes distances. Il s'agirait donc ainsi de repérer un signal relativement faible dans des séquences extraordinairement étendues.

À cet égard, plusieurs approches expérimentales et bio-informatiques devraient permettre à terme de mieux circonvenir les régions non codantes pertinentes.

D'une part, plusieurs méthodes expérimentales, dont le ChIP (liaison covalente des facteurs de transcription, suivie par une immunoprécipitation de la chromatine) et la mise en évidence de régions spécifiquement reconnues par des facteurs chimériques combinant un domaine spécifique de liaison à l'ADN et un domaine méthylase [10], en combinaison avec le développement de puces à ADN génomiques (c'est-à-dire couvrant l'ensemble du génome d'un organisme, y compris les régions non codantes), devraient rapidement nous renseigner sur les parties non codantes impliquées dans des interactions avec des facteurs de transcription ou de remodelage de la chromatine [11].

D'autre part, le séquençage d'un nombre croissant d'organismes à des distances évolutives variées a stimulé le développement de méthodes génomiques comparatives qui permettent la mise en évidence de blocs conservés non codants, appelés « empreintes phylogénétiques », potentiellement impliqués dans des mécanismes de régulation transcriptionnelle [12]. Ces méthodes enchaînent essentiellement trois analyses : (1) l'identification de régions géniques orthologues ; (2) l'alignement global des régions orthologues à l'aide d'algorithmes performants ; (3) l'analyse de la conservation de séquences au sein d'une fenêtre glissante. Couplée à un outil de recherche de motifs *cis*-régulateurs (utilisant par exemple des matrices poids-positions), l'exigence de conservation interspécifique peut ainsi éliminer jusqu'à environ 90 % des prédictions fausses récurrentes avec les meilleurs logiciels. Cependant, une difficulté inhérente à cette approche provient de l'hétérogénéité des mécanismes évolutifs et donc de la conservation de séquences entre organismes (réarrangements, etc.). Plus récemment, l'alignement multiple de plusieurs séquences orthologues relativement proches a permis la mise en évidence de variations de conservation similaires en minimisant les cas de réarrangement. Cette approche dite d'« ombrage phylogénique » est appelée à se généraliser au fil de l'accumulation des génomes complètement séquencés [7].

Dans ce contexte, la modélisation bayésienne utilisée par M.A. Beer et S. Tavazoie devrait permettre la prise en compte progressive de caractéristiques supplémentaires, comme la conservation interspécifique des motifs *cis*-régulateurs ou la répartition des motifs au sein de régions non codantes éloignées les unes des autres (régions 5' et 3', introns, etc.). Cette approche devrait également pouvoir s'articuler avec certains travaux en cours sur la modélisation dynamique des réseaux de régulation génétique, de manière à vérifier la cohérence globale des mécanismes de régulation avec les données cinétiques disponibles pour différentes conditions expérimentales [9]. En tout état de cause, la pertinence des prédictions engendrées, aussi précises soient-elles, devra finalement être validée expérimentalement. Cela nécessitera vraisemblablement la mise au point de nouvelles méthodes expérimentales ou la mise à l'échelle de méthodes existantes pour (in)valider les nombreux mécanismes de régulation transcriptionnelle inférés. ◊

REMERCIEMENTS

Les auteurs remercient S. Tavazoie pour avoir mis à leur disposition plusieurs éléments graphiques à la base de la réalisation de la Figure 1, *ainsi que J. Imbert pour ses suggestions sur une version préliminaire de ce manuscrit.*

SUMMARY

Prediction of transcription and genomic sequences

Technological developments have enhanced DNA sequencing at genomic scale. On the basis of the resulting sequences, computational biologists now attempt to localise the most important functional regions, starting with genes, but also importantly the regulatory motifs and conditions controlling their expression. In a recent paper published in *Cell*, M.A. Beer and S. Tavazoie report the results obtained by combining statistical classifications (clustering) of transcriptome data (DNA chips), software for the discovery of *cis*-regulatory patterns, together with a probabilistic learning method to infer regulatory rules tentatively accounting for the observed transcriptional profiles. ◊

RÉFÉRENCES

1. Jordan B. Jusqu'où iront les puces ? *Med Sci (Paris)* 2000 ; 16 : 950-3.
2. Hughes JD, Estep PW, Tavazoie S, Church GM. Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J Mol Biol* 2000 ; 296 : 1205-14.
3. Beer AM, Tavazoie S. Predicting gene expression from sequence. *Cell* 2004 ; 117 : 185-98.
4. Sherlock G. Analysis of large-scale gene expression data. *Brief Bioinformatics* 2001 ; 2 : 350-62.
5. Bulyk ML. Computational prediction of transcription-factor binding site locations. *Genome Biol* 2003 ; 5 : 201.
6. Van Helden J. Prediction of transcriptional regulation by analysis of the non-coding genome. *Curr Genomics* 2003 ; 4 : 217-24.
7. Wasserman WW, Sandelin A. Applied bionformatics for the identification of regulatory elements. *Nat Rev Genet* 2004 ; 5 : 276-87.
8. Friedman N. Inferring cellular networks using probabilistic graphical models. *Science* 2004 ; 303 : 799-805.
9. Thieffry D, de Jong H. Modélisation, analyse et simulation des réseaux génétiques. *Med Sci (Paris)* 2002 ; 18 : 492-502.
10. Van Steensel B., Henikoff S. Identification of *in vivo* DNA targets of chromatin proteins using tethered Dam methyltransferase. *Nat Biotechnol* 2000 ; 18 : 424-8.
11. Buck MJ, Lieb JD. ChIP-chip : considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* 2004 ; 83 : 349-60.
12. Lenhard B, Sandelin A, Mendoza L, *et al*. Identification of conserved regulatory elements by comparative genome analysis. *J Biol* 2003 ; 2 : 13.

TIRÉS À PART

D. Thieffry