Chapitre 6

Bioinformatique

Dans ce chapitre, nous introduisons successivement les principaux objets et leur différents niveaux d'étude en bioinformatique : les macromolécules, leurs interactions ainsi que la formalisation et l'extraction de connaissance, et présentons pour chacun des exemples d'apports de techniques d'intelligence artificielle sur différents problèmes clefs de ce domaine. Cette présentation est nécessairement partielle et partiale, tant la diversité des problèmes abordés et des techniques utilisées est vaste; nous espérons cependant qu'elle illustre la richesse des interactions qui se sont créées et continuent de croître entre ces deux disciplines.

6.1 Introduction

La bioinformatique s'intéresse à toutes les applications de l'informatique et des technologies de l'information à la biologie moléculaire. En tant que domaine de recherche, elle est devenue visible avec le développement rapide des technologies de séquençage de génome au début des années 1990, se focalisant d'abord sur les problèmes d'analyse de séquences d'ADN. De nos jours, la bioinformatique offre un large spectre couvrant la conception de bases de données, d'algorithmes dédiés, de techniques statistiques ou de théories pour résoudre des problèmes formels ou pratiques générés par la gestion et l'analyse de données biologiques diverses, dépassant largement le seul cadre des séquences d'ADN. Elle s'étend maintenant à toutes les approches informatiques visant à améliorer la compréhension des processus biologiques, à toutes les échelles du vivant (ADN, cellule, tissu, organisme, populations).
De fait, la bioinformatique exploite des résultats issus de domaines variés de l'informatique, des statistiques et des mathématiques. La finalité générale, l'interprétation de données brutes, souvent de nature discrète (telles les séquences), et produites dans des quantités importantes par des technologies dites *à haut débit*, a rapidement donné de l'importance aux techniques issues de l'intelligence artificielle. Les méthodes de découverte de motifs, de fouille de données, de traitement du signal et du langage, d'apprentissage automatique, d'optimisation, de satisfaction de contraintes, de raisonnement, et de visualisation de données complexes, sont particulièrement utiles dans ce contexte. L'efficacité des algorithmes est souvent décisive en bioinformatique, compte tenu de la taille des données manipulées (la séquence du génome

Auteurs : FRANÇOIS COSTE, CLAIRE NÉDELLEC, THOMAS SCHIEX et JEAN-PHILIPPE VERT.

humain, par exemple, contient de l'ordre de trois milliards de caractères). Enfin, les données fournies par la biologie moléculaire étant issues de processus expérimentaux, elles sont habituellement entachées d'erreurs. Cela a donné la part belle aux approches permettant de raisonner dans l'incertain, en particulier aux approches probabilistes de l'apprentissage automatique (voir chapitre I.9).

6.2 Les macromolécules

Au cœur de chaque cellule, des bactéries aux organismes les plus complexes, se trouve un ensemble de macromolécules, incluant en particulier les *chromosomes* qui portent le patrimoine génétique de l'espèce et sont transmis de génération en génération. Chaque chromosome est constitué d'une longue double-hélice d'*acide désoxyribonucléique (ADN)* dont chaque brin est un enchaînement de *nucléotides*, pouvant être de quatre types: adénine (A), thymine (T), guanine (G) ou cytosine (C). Chaque nucléotide d'un des deux brins (A, T, G ou C) est apparié par des liaisons hydrogène à un nucléotide complémentaire de l'autre brin (T, A, C ou G respectivement). Ces séquences de nucléotides sont le support dans la cellule de différents processus biochimiques qui permettent, à partir de régions localisées de l'ADN comme les gènes, la synthèse d'autres molécules telles que les *acides ribonucléiques (ARN)*, dont certaines sont ensuite traduites en *protéines* qui sont elles-mêmes des polymères linéaires formés d'une succession d'*acides aminés*, dont il existe 20 types. Chaque acide aminé est représenté par un triplet de nucléotides (appelé codon). L'ADN contient donc l'essentiel de l'information nécessaire à l'exécution du programme de travail de la cellule, que la biologie vise à analyser et comprendre.

La détermination de la séquence de nucléotides qui constitue tout ou partie de l'ADN d'un ensemble de cellules s'appelle le *séquençage*. Les techniques de séquençage, apparues dans les années 1970, étaient initialement assez lentes et coûteuses. Les progrès technologiques rapides permirent cependant de séquencer de nombreux organismes, entre le premier virus séquencé en 1977 et le génome humain séquencé en 2003. Une nouvelle génération de séquenceurs est en train de révolutionner ce domaine: en lisant en parallèle des milliards de courts fragments (typiquement entre 30 et 150 nucléotides), ces méthodes ont drastiquement réduit le coût et la durée du séquençage (Metzker, 2010). L'assemblage de ces fragments soulève des problèmes algorithmiques variés (Miller *et al.*, 2010), alors que l'on annonce simultanément le développement de nouvelles technologies capables de séquencer directement un chromosome entier, ce qui rendrait obsolète ce problème d'assemblage. Ces évolutions technologiques rapides sont fréquentes. Elles peuvent rapidement transformer un problème de bioinformatique crucial en un problème obsolète.

Le séquençage de l'ADN permet, d'une part, de mieux analyser le patrimoine génétique de chaque espèce, et d'autre part d'étudier, au sein d'une espèce, les différences génétiques entre différents individus. Au niveau d'une espèce, le passage de l'ADN séquencé à l'ensemble des molécules (ARN, protéines) que la cellule peut produire reste un problème difficile car les régions de l'ADN impliquées dans ce processus ne sont pas explicites et peuvent être morcelées. Ces problèmes d'analyse de séquence et de prédiction de gènes continuent de faire l'objet de nombreux travaux en bioinformatique. Une fois ce catalogue construit, il faut encore identifier les fonctions biochimiques de chaque molécule, les conditions dans lesquelles leur synthèse est déclenchée et les acteurs avec lesquels elle peut réagir ou interagir. L'annotation fonction-

nelle des génomes cherche à répondre à ces questions. Au sein d'une espèce, la comparaison des génomes de différents individus peut permettre l'identification de mutations ou de variations structurales de l'ADN responsables de différences de phénotype ou de susceptibilités à certaines maladies.

6.2.1 Gènes de protéines

Une des première étapes dans l'analyse des séquences d'ADN consiste à identifier les régions géniques qui expriment des protéines. Le code des protéines est dégénéré puisque plusieurs codons, dits synonymes, représentent un même acide aminé, la fréquence des codons synonymes dans chaque génome présentant une variabilité importante. Une partie des codons jouent un rôle de ponctuation : certains marquent par exemple la fin (obligatoire) d'un gène, alors que d'autres indiquent le démarrage (possible) d'un gène. Il est naturellement crucial de commencer la lecture 3 par 3 des nucléotides d'un gène sur le "bon pied" (dans la bonne phase) pour correctement interpréter les codons.

Prédiction de gènes La prédiction des gènes de protéines peut être d'une difficulté très variable selon le type d'organisme et en particulier la compacité de son génome. Dans les organismes les plus simples (virus, bactéries), une protéine est associée à un intervalle de la séquence d'ADN du génome, ces intervalles pouvant être chevauchants.
Dans les organismes possédant un noyau (eucaryotes), du parasite du paludisme aux vertébrés, en passant par les plantes, un gène est habituellement segmenté en une succession de régions appelées "exons", séparées par des régions non codantes appelées introns. Une ponctuation assez floue encadre ces différentes régions. Un intron est ainsi bordé par des motifs mal caractérisés, appelés signaux d'épissage, dont le contenu est lié à des interactions avec une machinerie moléculaire complexe et mal connue. La part du génome utilisée par ces gènes peut être très faible, ce qui ne facilite pas leur recherche.
Trois grands principes sont utilisés pour détecter les gènes :

1. les propriétés *statistiques* spécifiques liées à l'utilisation du code génétique (périodicité grossière de 3, fréquences des codons...) et à la présence des signaux (épissage, codons spécifiques...). On parle d'approche *ab initio* ou intrinsèque.
2. la similarité (syntaxique) d'une partie du génome avec une séquence expérimentalement caractérisée comme faisant partie d'un gène. On parle d'approche extrinsèque.
3. la conservation de la séquence entre deux régions de deux organismes pas trop proches (ayant suffisamment divergé au cours de l'évolution pour qu'une telle conservation soit inattendue). Si elle est significative, cette conservation doit être le fruit d'une pression sélective liée au fait que la région assure une fonction biologique importante et porte sans doute un gène. On parle d'approche comparative.

L'approche *ab initio* s'est rapidement appuyée sur des modélisations classiques de la segmentation utilisées en reconnaissance des formes et en traitement du signal et du langage. Dans un premier temps, ce sont essentiellement des variantes des chaînes de Markov qui ont été utilisées pour caractériser les gènes des organismes simples (bactéries). Le cas plus complexe des organismes eucaryotes, avec des gènes segmentés, a été abordé en s'appuyant sur des modèles de type chaînes de Markov cachées (**?**, Hidden Markov Models, HMM,)]rabiner1990tutorial

ou des variantes telles les semi-HMM, permettant de représenter explicitement les distributions de longueur des différentes régions, exons, introns, région sans gène, etc... Pour identifier les signaux de ponctuations "flous" dans les gènes, une très grande variétés de méthodes de classification de motifs textuels ont été utilisées : réseaux de neurones, fonctions de discrimination linéaires ou non, séparateurs à vaste marges (SVM) (voir chapitre I.9), modèles probabilistes (chaînes de Markov non homogènes, réseaux bayésiens.... Voir chapitre II.8). Tous ces modèles, qu'ils soient probabilistes ou non, sont généralement appliqués dans un cadre "supervisé" et nécessitent par conséquent la construction préalable de grands jeux d'apprentissage utilisés pour estimer les paramètres des modèles. La construction du jeu d'apprentissage peut être assez lourde, car elle nécessite généralement un travail expérimental délicat. Ces méthodes sont toujours utilisées dans les prédicteurs de gènes *ab initio* (Lukashin et Borodovsky, 1998; Stanke *et al.*, 2006; Burge et Karlin, 1997). Un travail algorithmique non négligeable a été nécessaire pour permettre de tirer parti des propriétés spécifiques du problème de prédiction de gènes pour rendre praticable les approches de type semi-HMM dont les algorithmes associés, de complexité quadratique, sont peu compatibles avec la taille des données traitées.

Les approches extrinsèques et comparatives sont limitées par la connaissance d'un existant avec lequel se comparer. La majorité des travaux ici ne proviennent pas de l'intelligence artificielle mais s'appuient sur des programmes issus du domaine de l'analyse de séquences permettant la comparaison rapide de séquence à grande échelle tels que BLAST (Altschul *et al.*, 1990), qui est l'un des outils les plus utilisés de la bioinformatique. Les régions similaires ainsi identifiées sont rarement complètes ou précises, mais assez fiables.

Depuis les années 2000, l'essentiel des travaux a été consacré à la conception de prédicteurs "consensuels" (ou intégratifs) permettant de prendre en compte simultanément le plus possible d'informations pertinentes : propriétés statistiques, similarités de séquence, conservation de séquence, prédictions issues d'autres sources (expérimentales ou non). Le modèle HMM qui modélise la distribution de probabilité jointe des données observées, c'est-à-dire la séquence d'ADN, et de sa segmentation en exons, introns... qui sont des données cachées, devient mal adapté. En effet, l'intégration d'une variété de données de sources disparates dans un tel cadre est incompatible avec les hypothèses d'indépendance sous-jacentes au formalisme. Les prédicteurs de gènes intégratifs utilisent donc systématiquement des modèles discriminants, modélisant uniquement en quoi les observations rendent plus ou moins vraisemblables différentes segmentations. Ainsi, le formalisme des champs aléatoires conditionnels (**?**, Conditional Random Fields, CRF,)]LaffertyMP01 est maintenant au cœur de nombreux prédicteurs de gènes car il permet de se libérer, dans une certaine mesure, des conditions d'indépendance des HMM en s'intéressant à la probabilité des segmentations sachant les observations. Les CRF sont utilisés dans des logiciels tels que Craig (Bernal *et al.*, 2007) ou CONRAD (DeCaprio *et al.*, 2007) et étaient déjà implicitement utilisés dans certains logiciels plus anciens tels qu'EuGène (Foissac et Schiex, 2005), utilisés en mode intégratif. L'extension de méthodes de classification telles que les SVM à des objets structurés complexes comme les segmentations a également mené à la création d'approches discriminantes non probabilistes telles que les Hidden Markov-SVM (**?**, HM-SVM,)]AltunTH03. Ces méthodes sont utilisées dans le prédicteur de gènes mGene (Schweikert *et al.*, 2009).

Annotation fonctionnelle L'identification de gènes dans un génome ne fournit qu'un catalogue des constituants de la machinerie cellulaire: l'annotation fonctionnelle de ces consti-

tuants vise à élucider leur fonction biologique. L'approche classique procède par homologie: les séquences sont annotées par transfert de l'annotation d'une séquence similaire, généralement trouvée par alignement local de séquences à l'aide d'outils tels que BLAST (Altschul *et al.*, 1990). À partir du moment où des familles de séquences partageant une annotation sont connues, on peut établir une signature caractéristique de chacune des familles et utiliser ces signatures pour annoter les nouvelles séquences. De nombreuses banques de motifs sont disponibles en ligne. Parmi les plus connues, on peut citer Prosite (Hulo *et al.*, 2008), qui propose des signatures calculées de façon semi-automatique pour caractériser des sites fonctionnels, des domaines et des familles de protéines, ainsi que Transfac (Wingender *et al.*, 2000), spécialisée dans les signatures de sites de fixation des facteurs de transcription sur l'ADN, qui sont des sites situés dans les régions promotrices en amont des gènes et impliqués dans la régulation de la quantité de protéines à synthétiser à partir du gène. Interpro (Hunter *et al.*, 2009), banque fédérative de signatures protéiques est un très bon point d'entrée donnant un aperçu de la variété des signatures des protéines. Les signatures sont généralement définies par la recherche de régions conservées au cours de l'évolution dans l'ensemble de la famille, par alignement multiple de séquence ou par découverte de motifs. Elles peuvent caractériser des régions localisées et bien conservées comme par exemple un site actif de protéine. Dans ce cas, des motifs exacts, allant de la séquence dégénérée aux motifs utilisés dans Prosite (à l'expressivité inférieure aux expressions régulières), peuvent être utilisés ainsi que leur version pondérée pour la caractérisation de régions moins conservées. La découverte de motifs (Brazma *et al.*, 1998; Brejova *et al.*, 2003) est un domaine de la bioinformatique ayant tiré profit des techniques d'exploration d'espace de recherche développées en intelligence artificielle et offrant aujourd'hui des outils fonctionnels aux biologistes tels que Weeder (Pavesi *et al.*, 2001), procédant par énumération et sélection de motifs sur-représentés sur l'ADN, Pratt (Jonassen *et al.*, 1995), découvrant des motifs Prosite par exploration en profondeur d'abord et ancrage sur des positions exactement conservées, ou AlignAce (Roth *et al.*, 1998) et MEME (Bailey et Elkan, 1994), identifiant des régions localement conservées respectivement par échantillonage de Gibbs et *Expectation Maximization*. Pour caractériser des régions plus longues, il est nécessaire d'utiliser des modèles probabilistes. Par exemple, une protéine est généralement décomposée en quelques *domaines* (unités structurelles et fonctionnelles autonomes observées chez les protéines) ; les bases de données ProDom (Corpet *et al.*, 1998) et Pfam (Sonnhammer *et al.*, 1997) contiennent plusieurs milliers de modèles de chaînes de Markov cachées de type profile (pHMM) (Durbin *et al.*, 1998) caractéristiques de chaque type de domaine connu et largement utilisés par la communauté pour l'annotation de génomes. Le succès des pHMM tient en leur structure simple de type gauche-droite, modélisant cependant les phénomènes d'insertions, de délétions et de mutations, et surtout dans le schéma d'entraînement des pondérations basé sur une grande utilisation de densités de probabilité définies *a priori*, implémenté dans les deux principales suites logicielles SAM et HMMR (Krogh *et al.*, 1994; Eddy, 1998). Des structures de HMM plus spécialisées intégrant des connaissances expertes sur l'application visée ont été conçues avec succès pour la prédiction de régions transmembranaires dans TMHMM (Sonnhammer *et al.*, 1998) et/ou de signaux peptidiques dans Phobius (Kall *et al.*, 2004), éventuellement couplées à des réseaux de neurones artificiels comme dans SignalP 3.0 (Bendtsen *et al.*, 2004). L'inférence automatique de la structure de HMMs reste un problème difficile et relativement ouvert, même si des avancées, obtenues par le croisement de techniques d'alignement de séquence avec des méthodes d'inférence grammaticale, ont pu être réalisées dans ce domaine

pour la caractérisation de familles de protéines (Kerbellec, 2008).

Prédiction de structures 3D La fonction biologique d'une protéine étant largement déterminée par sa structure tridimensionnelle, la prédiction de cette structure à partir de la séquence d'acides aminés lus sur le génome revêt un grand intérêt. De nombreux outils spécifiques issus de l'intelligence artificielle ont été développés pour tenter de résoudre ce problème considéré comme l'un des grands défis de la biophysique. On peut notamment citer l'utilisation de méthodes d'apprentissage supervisée pour prédire la structure tridimensionnelle locale (Baldi *et al.*, 1999; Ward *et al.*, 2003; Kuang *et al.*, 2004), ou de réseaux de neurones récurrents pour prédire la carte de contacts globale entre les acides aminés (Pollastri et Baldi, 2002). Une autre approche consiste à prédire directement la famille structurale de la protéine, problème qui est souvent formalisé comme un problème d'apprentissage supervisé multiclasse sur séquences: un nombre fini de familles structurales est défini par comparaison des structures connues (Murzin *et al.*, 1995), et le but est de prédire la classe d'une protéine de séquence donnée étant donné les classes d'un ensemble de protéines de séquences et structures connues. Cette formulation a promu le développement de nombreuses méthodes d'apprentissage statistique pour la classification supervisée de séquences de longueurs variables, en particulier les méthodes à noyau utilisant des noyaux pour séquences (Jaakkola *et al.*, 2000; Leslie *et al.*, 2002; Saigo *et al.*, 2004).

Analyse de transcriptome Les problématiques d'annotation de protéines se limitent rarement à l'analyse des séquences d'acides aminés. Une autre source d'information utile pour comprendre le rôle que peut jouer une protéine dans un contexte cellulaire donné est son niveau d'expression, estimé par la mesure de la quantité d'ARN messager codant cette protéine par des *puces à ADN*, ou, depuis peu, via les nouvelles les nouvelles technologies de séquençage (*RNA-seq*). Ces technologies permettent en effet, pour un coût relativement modeste, de mesurer l'expression de l'ensemble des gènes d'un organisme, appelé son *transcriptome*, dans un échantillon donné. En mesurant l'expression des gènes dans des échantillons soumis à diverses conditions expérimentales, provenant de différents patients, ou provenant de souches ayant des différences génétiques connues, on peut caractériser chaque gène par son profil d'activité dans ces différents échantillons. Cette information est évidemment très utile pour l'annotation fonctionnelle des gènes, et de nombreux outils de classification supervisée et non supervisée, visualisation, ou décomposition en composantes additives on été proposés dans ce but. Par exemple, la simple classification hiérarchique est un outil maintenant standard pour détecter des groupes de gènes ayant des profils similaires, et donc susceptibles d'avoir également des fonctions similaires (Eisen *et al.*, 1998). Les algorithmes de classification supervisée ont été utilisés pour prédire la fonction de gènes à partir de leurs profils d'expression (Brown *et al.*, 2000). Un autre exemple est l'utilisation d'analyse en composantes principales ou de factorisation de matrices pour la recherche de variables latentes susceptibles de représenter des modules de gènes participant à une même fonction biologique (Raychaudhuri *et al.*, 2000; Brunet *et al.*, 2004).

Intégration de données Les séquences et les données d'expression des gènes ne sont d'ailleurs pas les seules informations dont on peut disposer pour tenter d'inférer les fonctions des gènes. La protéomique, qui quantifie directement la concentration des protéines dans

un échantillon donné, ou d'autres technologies qui permettent de mesurer à grande échelle les sites de fixation sur l'ADN de nombreuses protéines ou les variations de méthylation de l'ADN, fournissent également de précieuses informations. Une problématique centrale dans ce contexte est donc l'*intégration de données*, visant à combiner ces différents types de données hétérogènes dans des modèles de classification supervisés ou non. On peut citer par exemple les développements récents des méthodes à noyaux (Lanckriet *et al.*, 2004), des méthodes bayésiennes pour l'intégration d'informations fournies par différents types de données (Troyanskaya *et al.*, 2003), ou des méthodes de graphes pour l'intégration de similarités (Nabieva *et al.*, 2005).

6.2.2 Gènes d'ARN

Une partie du fonctionnement de la cellule est rendue possible par d'autres molécules que les protéines. Il s'agit de l'ARN, longtemps considéré comme un support temporaire de l'information entre ADN (génome) et protéines (après traduction). Depuis plusieurs années, le rôle crucial de la molécule d'ARN dans le fonctionnement cellulaire a été largement confirmé, menant à un intérêt croissant pour les gènes dits d'ARN non codant pour des protéines (*non coding* RNAs, ncRNA).

Contrairement à l'ADN, molécule double brin dont la structure est une hélice, bien connue, l'ARN est une famille de molécules simple brin formée d'acides ribonucléiques identifiés par les lettres A, U, G et C (l'Uracile remplace la Thymine). En se liant à elle-même via des liaisons hydrogènes (préférentiellement A-U et G-C – ces dernières étant plus stables – mais pas exclusivement), l'ARN forme des hélices locales définissant une structure tridimensionnelle qui peut lui conférer une fonction biochimique spécifique. Comme elle est déterminante pour leur fonction, et qu'elle permet seule de les comparer de façon informative, la structure d'un ARN et sa détermination à partir de sa séquence seule a fait l'objet d'un intérêt depuis longtemps. Si l'apport de la physique statistique et de la programmation dynamique (Nussinov et Jacobson, 1980; Zuker et Stiegler, 1981) pour déterminer une structure d'énergie optimale reste essentiel, les modèles thermodynamiques restent assez approximatifs, et la vraie structure n'est malheureusement pas souvent la structure optimale. Certains outils de prédiction de structure tels que celui de Gaspin et Westhof (1995) ou MC-SYM (Major, 2003) s'appuient sur des formalismes de l'intelligence artificielle tels que les réseaux de contraintes (Rossi *et al.*, 2006) pour caractériser et identifier les structures possibles d'une séquence d'ARN en intégrant plus d'informations que les données de thermodynamique.

Pour intégrer encore plus d'information, une approche comparative est souvent utilisée. Elle consiste à prédire une structure sur la base de plusieurs séquences d'ARN ayant la même fonction mais issus d'organismes différents. Une structure idéale est alors une structure qui peut être adoptée par ces séquences (formation d'hélices) et qui est d'énergie pas trop éloignée de l'énergie optimale. L'approche fondamentale (Sankoff, 1985) s'appuie sur la programmation dynamique mais est trop lourde en pratique. Des versions heuristiques mais efficaces ont donc été définies (Touzet et Perriquet, 2004; Bernhart *et al.*, 2008).

La question de l'annotation des génomes a conduit à développer des méthodes d'identifications de gènes d'ARN dans les génomes complets. Trois types d'approches existent : (1) en partant d'une description de la structure d'ARN connus et en cherchant les régions qui peuvent se replier selon la même structure, (2) via une approche comparative (exploitant la conservation

des hélices et autres composants essentiels de la structure) ou (3) *ab initio* (en s'appuyant sur les propriétés intrinsèques de l'ARN).
La première approche nécessite de pouvoir formaliser ce qu'est "la structure" d'une famille ARN, ce qui est déjà problématique. En ignorant certaines complexités des structures, celles-ci peuvent en général se décrire via un arbre, et le formalisme des grammaires stochastiques libres de contexte (*Stochastic Context Free Grammar*, ou SCFG), utilisé notamment en intelligence artificielle en analyse du langage, fournit des représentations pouvant être entraînées automatiquement pour caractériser chaque famille de séquences de structures connues. Ces modèles, aussi appelés modèles de covariance (Sakakibara *et al.*, 1994; Eddy et Durbin, 1994), sont ensuite utilisés pour analyser (*parser*) la séquence génomique (par programmation dynamique) et identifier des régions qui sont vraisemblablement des gènes d'ARN de la même famille. Différentes bases de données (RFAM (Griffiths-Jones *et al.*, 2005), ncRNAdb (Szymanski *et al.*, 2007)...) regroupent au niveau international un ensemble de modèles de ce type. Ces modèles probabilistes restent lourds et peu "compréhensibles" par les biologistes. Une autre famille d'approche s'appuie plus sur l'utilisation de langages dédiés et accessibles pour décrire les propriétés que doivent satisfaire les ARN d'une famille donnée (Laferriere *et al.*, 1994; Dsouza *et al.*, 1997; Billoud *et al.*, 1996). Les formalismes de description des connaissances tels que les réseaux de contraintes et les réseaux de contraintes pondérées ont ainsi été utilisés (Thebault *et al.*, 2006; Zytnicki *et al.*, 2008). Les algorithmes associés, de type "recherche arborescente avec filtrage par cohérence locale" (Rossi *et al.*, 2006), permettent ensuite de localiser les régions qui satisfont ces propriétés, par la résolution d'un problème de satisfaction de contraintes . Plus ouverts que les modèles probabilistes, ils présentent l'inconvénient de ne pas disposer de méthodes de construction automatique à partir d'un jeu d'ARN de structure connue. D'autres méthodes suivent une approche hybride (Macke *et al.*, 2001).
L'approche comparative s'appuie assez largement sur les algorithmes de prédiction comparative de structures d'ARN (Touzet et Perriquet, 2004; Bernhart *et al.*, 2008). Ils construisent une structure consensus à partir des régions similaires en séquence dans l'ensemble des génomes considéré. A partir des énergies des structures et de leur conservation dans les différents génomes, l'algorithme décide si ces régions conservées définissent un gène d'ARN. Ce problème de discrimination est résolu, dans le logiciel RNAZ, via l'utilisation d'un discriminant de type SVM (Washietl, 2007).
L'approche *ab initio* pure reste encore très limitée car les propriétés générales des gènes d'ARN sont moins fortes que celles des gènes de protéines. Elle consiste à exploiter la nécessité d'une présence suffisante de nucléotides GC pour stabiliser les hélices, méthode simple qui ne donne des résultats que pour les génomes pauvres en GC, pour lesquels on observe un contraste suffisant.

6.3 Les réseaux biologiques

L'annotation fonctionnelle des gènes et protéines se heurte rapidement au fait que la plupart des fonctions biologiques dans une cellule impliquent de nombreux acteurs qui interagissent entre eux. La description et la modélisation de ces fonctions ne peut donc pas se limiter à la description des acteurs pris individuellement. Les formalismes des graphes, et plus largement des systèmes dynamiques, se sont rapidement imposés pour décrire ces réseaux de protéines et d'autres molécules en interaction, et que l'on appelle collectivement les *réseaux biologiques*.

Le concept de réseau biologique se décline en différentes catégories, relatifs aux fonctions biologiques assurées par les réseaux. Les *voies métaboliques* assurent l'exploitation et la transformation des ressources disponibles en énergie. Elles sont constituées par des réactions chimiques de dégradation et de synthèse travaillant de manière séquentielle, et catalysées par des protéines spéciales appelées *enzymes* (Cornish-Bowden *et al.*, 2005). La transmission de signaux extracellulaires au sein de la cellule passe par des cascades (rapides) de modifications successives de différentes protéines, et sont organisées aux sein de ce que l'on appelle les *voies de signalisation*. La production régulée des protéines s'organise via le contrôle de l'expression de gènes par des protéines spécifiques (les *facteurs de transcription*), induisant éventuellement des phénomènes de rétro-contrôle, et aboutissant à des *réseaux de régulation génétique* (De Jong, 2002).
La construction, la modélisation, l'analyse, la simulation et l'exploitation de ces réseaux sont au coeur de la *biologie des systèmes*, un des domaines les plus actifs de la bioinformatique aujourd'hui. On y distingue deux sous-domaines: les questions relatives à l'identification des interactions et des voies fonctionnelles à partir de jeux de données variés sont rassemblées sous le terme de *biologie intégrative*, ou *inférence de réseau*. La *modélisation dynamique* rassemble par contre les méthodologies permettant de simuler ces systèmes, d'étudier les propriétés de leurs évolutions temporelles, allant jusqu'à la proposition de plan expérimentaux pour contrôler leur comportement.

Niveaux de description Les réseaux biologiques non seulement se distinguent par leur fonction au sein de la cellule, mais varient aussi par leur niveau de description. Dans le cas le plus grossier, seuls les effets à long terme de l'augmentation d'un produit sur le taux de production de ses cibles sont décrits par des graphes d'interaction ou des graphes d'influence, qui peuvent être obtenus par exemple par inférence automatique. Des descriptions plus cinétiques concernent les effets déterministes des réactions moléculaires c'est-à-dire, l'état du système au temps $n + 1$ en fonction de son état au temps n. C'est le cas par exemple, des réseaux de signalisation. Elles sont représentées par des graphes de réaction qui sont interprétés à l'aide de langages et de formalismes booléens (Kauffman, 1971; Chabrier-Rivier *et al.*, 2004). Ces approches déterministes sont parfois prises en défaut par l'existence de différentes échelles de temps, en particulier dans les réseaux de régulation génétique – une transcription est bien plus longue que la dégradation d'une protéine. La dynamique devient alors non déterministe, ses attracteurs à moyen terme qui varient dans l'espace des concentrations sont décrits à l'aide de modèles multivalués ou linéaires par morceaux (Thomas, 1981; De Jong, 2002).

Inférence de réseaux biologiques Les divers réseaux biologiques sont généralement construits à partir d'expériences biologiques longues et coûteuses. Par exemple, l'identification des gènes régulés par un facteur de transcription donné passe par des mutations dirigées au niveau de l'ADN pour empêcher la fixation du facteur de transcription à un endroit qu'il aura déjà fallu déterminer. Afin d'accélérer la construction de ces réseaux et d'étendre leur couverture, sans biais, à l'ensemble des acteurs concerné, des méthodes d'inférence automatique ont été proposées pour la reconstruction *in silico* des réseaux biologiques. Les mesures du niveau d'expression des gènes (puces à ADN, RNA-seq) au niveau de la cellule, dans différentes conditions, à différents instants ou chez différents individus laissent en effet entrevoir la possibilité d'observer des corrélations entre les niveaux d'expression de différents gènes qui

seraient les conséquences de régulations entre ces gènes.
Les modèles graphiques probabilistes (champs de Markov, modèles graphiques gausssiens mais plus particulièrement les réseaux bayésiens) ont été largement mobilisés sur ce type d'approche (Friedman *et al.*, 2000). Ils fournissent en effet un cadre statistique et des méthodes d'inférence relativement efficaces pour détecter des relations d'indépendance conditionnelles entre variables aléatoires en recherchant une structure la plus simple possible expliquant les données. Mais les modèles construits ne sont pas nécessairement causaux (Pearl, 2000). Lorsque l'on dispose de mesures d'expression au cours de séries temporelles, des méthodes d'inférences de systèmes dynamiques ont été proposées, comme par exemple les réseaux booléens (Akutsu *et al.*, 2000) ou les réseaux bayésiens dynamiques (Dojer *et al.*, 2006). Une autre approche appelée génomique génétique (Jansen et Nap, 2001) consiste à exploiter en sus la variabilité génétique (supposée connue) entre individus pour expliquer les différents profils d'expression observés.
D'autres approches cherchent, plus directement, à connecter des gènes ayant des profils d'expression similaires en terme de corrélations ou d'information mutuelle (Margolin *et al.*, 2006; Kharchenko *et al.*, 2004; Jansen *et al.*, 2003). Un autre type d'approche, enfin, consiste à inférer de nouvelles connections dans un réseau partiellement connu par une procédure d'apprentissage automatique, en apprenant en quoi les données disponibles permettent de prédire les arêtes déjà connues (Ben-Hur et Noble, 2005; Bleakley *et al.*, 2007; Mordelet et Vert, 2008).
Enfin, l'acquisition de paramètres fins est réalisée par l'intégration d'observations cinétiques à l'aide d'approches hypothetico-déductives, que ce soit sur des réseaux génétiques ou de signalisation (Corblin *et al.*, 2009; Calzone *et al.*, 2006) ou les voies métaboliques, pour lesquelles on parvient à planifier des tests validant la fonction d'enzymes (King *et al.*, 2009).

Modélisation dynamique Une fois un réseau biologique construit, un certain nombre de méthodologies permettent de modéliser les effets des interactions entre molécules sur le comportement global du système, en s'appuyant sur du raisonnement automatique à base de contraintes. Les graphes d'interaction mettent en évidence des effets à long terme en comparant différents états stationnaires du système. Dans ce but, on intègre une règle causale (D'haeseleer *et al.*, 1999) dans des approches formelles de type résolution de contraintes ou modélisation booléenne, pour tester la cohérence entre données et modèles (Bay *et al.*, 2003; Juvan *et al.*, 2005; Covert *et al.*, 2008; Guziolowski *et al.*, 2009). La formalisation des voies métaboliques, s'effectuant par un raffinement des graphes d'interaction en intégrant une information sur la stoechiométrie des réactions, supportent des contraintes linéaires sur les états stationnaires, qui sont bien plus fortes que les contraintes causales induites par la structure du graphe (Reder, 1988). L'*analyse de flux élémentaires* exploitent ces contraintes linéaires, définissant un programme linéaire, pour contrôler le comportement de ces voies (Schuster *et al.*, 2000; Durot *et al.*, 2009; Larhlimi et Bockmayr, 2009).
Le niveau des graphes de réaction et des systèmes multivalués permet de raisonner plus précisément sur le comportement cinétique du système (Monteiro *et al.*, 2008). Différentes déclinaisons d'algèbres de processus se concentrent sur le comportement moyen de ces système (Regev et Shapiro, 2002; Antoniotti *et al.*, 2003; Danos et Laneve, 2004; John *et al.*, 2009). On utilise enfin des logiques temporelles ou des parcours de diagrammes de décision pour raisonner sur les trajectoires du système, en identifiant ses attracteurs, en vérifiant si les évolutions

du système sont conformes aux expérimentations, ou en prédisant la réponse des systèmes à diverses influences (Calzone *et al.*, 2006; Klamt *et al.*, 2006; Bernot *et al.*, 2004; Monteiro *et al.*, tics; Fauré *et al.*, 2009).

6.4 Formalisation et extraction des connaissances

6.4.1 Ontologies

Le volume de données, leur disparité et le besoin d'interprétation et d'intégration ont ouvert un vaste champ d'application en bioinformatique aux approches d'intelligence artificielle relevant de l'intégration de données structurées et non structurées, à base d'ontologies (voir chapitres I.20 et I.5) et de raisonnement formel (voir chapitre II.4).
De nombreuses ontologies en biologie sont disponibles en ligne (voir le site *BioPortal*) dont la justesse, l'expressivité, la pertinence, et la validité de la sémantique formelle sont très variables. Leur représentation est généralement réduite à une arborescence de concepts. L'ontologie la plus remarquable, Gene Ontology (GO) (Ashburner *et al.*, 2000) est le résultat d'un effort collaboratif initié pour l'annotation manuelle de gènes. Principalement adaptée aux eucaryotes et initialement pour l'annotation de génomes de souris, de mouche et de levure, GO contient plus de 27 000 concepts, organisés en trois hiérarchies, qui décrivent les processus biologiques, fonctions moléculaires et composants cellulaires, plus les relations *part-of* et *regulates*. Plus de 120 génomes ont été manuellement annotés par des concepts de GO. L'intérêt de telles annotations formelles est de permettre des traitements automatiques, dont principalement, la comparaison de gènes et l'annotation de nouveaux génomes par homologie (Conesa *et al.*, 2005). Une autre utilisation fréquente en analyse de transcriptome consiste à interpréter la corrélation de niveaux d'expression de gènes à l'aide des fonctions connues de ces gènes (Marco et Francesco, 2006).
GO est l'objet de deux critiques principales: son manque de pertinence pour certains organismes dont les procaryotes, et sa très grande taille qui rend difficile la recherche des concepts pour l'annotation manuelle ou la comparaison de gènes, malgré les navigateurs comme AmiGO et la gestion des synonymes. De sorte que pour l'annotation de nouveaux génomes, des classifications fonctionnelles réduites à quelques centaines de fonctions, et mieux adaptées peuvent être préférées (voir pour une comparaison (M Rison SCG, 2000)), bien que leur sémantique soit généralement moins formelle et que l'alignement avec GO soit souvent négligé. Une dernière conséquence regrettable de la normalisation à l'aide de GO est que les informations initiales plus riches, exprimées en langue naturelle sont parfois remplacées automatiquement, par des termes très généraux de GO dans les bases généralistes de biologie dans un but louable de normalisation, tel que par exemple dans *Genome Reviews*.
Au-delà de GO où un effort considérable de formalisation est fait, la confusion entre un vocabulaire contrôlé et structuré et une ontologie est fréquente dans les ontologies distribuées et le manque de distinction claire entre le lexique du domaine et les labels qui nomment les concepts est à l'origine de nombreuses formules invalides telles que *detection of osmotic stimulus* $\rightarrow$ *response to osmotic stress*, qui empêchent une exploitation formelle corrrecte. L'absence de relation autre que la relation de généralité oblige également les concepteurs à des contorsions qui nuisent à la sémantique formelle, par exemple, la définition des concepts *response to / humidity / water deprivation / flooding* comme spécialisations de *water response* dans GO

illustre la confusion entre l'eau et l'appréciation de la quantité d'eau par l'organisme dans son environnement. L'utilisation majoritaire de ces ontologies est encore manuelle ; elles servent de référentiel d'annotation ou de classification, plus ou moins normalisé.
A l'opposé, des bases de connaissances à la représentation très expressive (voir section 6.3) telles que la représentation de voies métaboliques, sont utilisées à des fins de modélisation et de simulation, mais la spécificité de leur objet les exclut du cadre ontologique.

6.4.2 Extraction et recherche d'information

Dans le champs de l'intelligence artificielle, la classification de documents, l'extraction d'information (EI) et la recherche d'information (RI) appliquées à la biologie sont des domaines en pleine expansion (Hirschman *et al.*, 2002; Ananiadou et Mcnaught, 2005; Nédellec *et al.*, 2009; Altman *et al.*, 2008; Jensen *et al.*, 2006). Après bientôt dix ans de recherche spécifique à la biologie, des outils opérationnels et intégrés dans des applications bioinformatiques commencent à voir le jour. Conjointement, la maturité des recherches génériques dans ces domaines, la mise à disposition de librairies d'apprentissage automatique (voir chapitre II.10) telle que Weka (Witten et Frank, 2005) et de plates-formes de traitement automatique de la langue (TALN) (voir chapitre III.5) généralistes comme Gate (Bontcheva *et al.*, 2004) ou UIMA, ou spécialisées pour la biologie, par exemple Textpresso (Mueller *et al.*, 2004), BioAlvis (Bossy *et al.*, 2008), les efforts investis dans l'organisation de compétitions telles que BioCreative, GENIA, LLL, TREC genomics sur des tâches relevant de l'EI et de la RI, ont permis le développement, l'adaptation, l'intégration et la comparaison de méthodes relevant des statistiques, de l'apprentissage automatique et du TALN. L'intégration de ces résultats dans des applications bioinformatiques est encore limitée, mais prend de l'ampleur grâce aux efforts de distribution et de publication.
Plus particulièrement, la reconnaissance automatique des entités nommées (REN) consiste à identifier les noms propres tels que les noms de gènes, d'espèces, de lignées cellulaires, de molécules, ou plus généralement les formes figées. Elle est une étape critique de toute application documentaire en ce qu'elle permet d'identifier les objets significatifs dans les documents. La *reconnaissance* elle-même, c'est-à-dire le typage d'un terme du document comme désignant ou non un membre du type est à distinguer de la *normalisation* qui consiste à associer à l'entité reconnue son nom canonique, sans ambiguité. La REN en biologie a fait l'objet de nombreuses compétitions, dont NLPBA (GENIA) (Kim *et al.*, 2004) et BioCreative (Smith *et al.*, 2008; Morgan *et al.*, 2008), et études favorisées par la diffusion de nomenclatures centralisées (GenBank, Hugo, UMLS) et de corpus annotés qui ont permis la mise au point de méthodes d'apprentissage de classification discriminante (surtout IDT, SVM et CRF), voir le chapitre II.10. Combinées à des étapes d'analyse de texte (segmentation, analyse terminologique), elles donnent des résultats de rappel et précision très variables selon les types d'entités et la richesse des dictionnaires. Actuellement, quasiment toutes les recherches en REN concernent la reconnaissance de noms de gène et de protéines avec un net succès comme démontrés par les sites IHOp ou CoCitation. Le système ABNER (Settles, 2005) est apprécié pour ses performances, son interface et son accessibilité pour des non-spécialistes. Les progrès à venir se situent plutôt (1) au niveau de la décomposition des sous-problèmes à traiter : homonymie, noms de mutants, distinction entre entités biologiques (opéron, promoteur, etc.) et rattachement à l'espèce, (2) de l'intégration des méthodes de TALN et d'apprentissage, et (3) de la qualité de l'annotation manuelle (Nédellec *et al.*, 2006), plutôt qu'au niveau des méthodes de classification elles-mêmes.

Contrairement au domaine biomédical, le domaine de la biologie n'a pas suscité beaucoup de travaux d'analyse terminologique, sinon l'exploitation de terminologies plus ou moins structurées par projection directe sur le texte. Les termes proviennent généralement des labels de GO (McCray *et al.*, 2002) ou de MESH et Specialist Lexicon de UMLS, et la projection inclut éventuellement la désambiguisation du sens (Andreopoulos *et al.*, 2008). La question de la variation terminologique se limite en général à la lemmatisation et au comptage des mots en commun entre les termes du texte et de la terminologie.

Les tâches les plus populaires en EI au-delà de la REN concernent l'extraction de relations d'interaction protéine-protéine dans BioCreative ou d'interaction génique dans LLL où de nets progrès en terme de qualité sont observés, mais où le passage à l'échelle reste un verrou malgré des tentatives remarquables comme celle de MEDIE. Par exemple, il s'agit dans *GerE stimulates sigK transcription* d'identifier la protéine *GerE* comme agent de l'interaction dont *sigK* est la cible. L'approche à base de patrons de surface, par exemple un verbe d'interaction encadré par deux noms de protéines, est peu à peu abandonnée au profit de l'analyse syntaxique des dépendances qui permet de rendre compte de cas de dépendances longues distances, d'ellipses ou d'apposition que les patrons échouent à traiter (Rinaldi *et al.*, 2008). Par exemple dans la phrase : *GerE stimulates cotD transcription and cotA transcription [. . .], and, unexpectedly,* i*nhibits [. . .] transcription of the gene (sigK) [. . .]*, l'extraction de l'interaction suppose d'avoir préalablement identifié que *GerE* est le sujet du verbe *inhibits* dont la transcription du gène *sigK* est l'objet. L'extraction de relations à base d'apprentissage automatique (classification discriminante) est rendue difficile par la complexité et la variété des formulations et la relative petite taille des corpus annotés. Le travail pluridisciplinaire de Alain-Pierre *et al.* (2008) compense par exemple cette limitation grâce à la programmation logique inductive (PLI), exploitant une ontologie décrivant les différents modèles biologiques sous-jacents, tels que la fixation (*binding*) de la protéine sur le promoteur, l'appartenance à un régulon ou la mutation du gène cible.

Les avancées dans le domaine de la RI pour l'indexation automatique en texte plein (*à-la-Google*), par opposition à l'indexation contrôlée, sont limitées par la faible adéquation des terminologies existantes. Trois raisons principales en sont à l'origine. Outre la rareté et l'incomplétude des terminologies pour de nombreux domaines à l'exclusion de certains champs du domaine biomédical, les terminologies sont conçues à des fins d'indexation contrôlée c'est-à-dire manuelle, ou de normalisation du vocabulaire, et non pas pour la projection automatique sur du texte. De ce fait, les termes d'indexation sont choisis pour leur compréhensibilité hors contexte quand les rédacteurs vont préférer une formulation moins dense. Enfin, les synonymes sont rarement indiqués, ou de façon incomplète au regard de la richesse des formulations textuelles. L'acquisition plus ou moins automatisée à partir de corpus de terminologies structurées repose sur l'extraction automatique de termes (Nédellec *et al.*, 2009), utilisant ou non des méthodes de variation morpho-syntaxiques, puis l'application de patrons dits *à-la-Hearst* ou la sémantique distributionnelle (classification non supervisée *clustering*) pour l'extraction de relations de synonymie ou d'hyperonymie. Les années à venir devraient voir l'intégration de ces approches automatiques avec des interfaces coopératives dans des applications opérationnelles.

Jusqu'ici, l'hypothèse très forte de bijection parfaite entre les labels des concepts des ontologies et les termes des textes a négligé des phénomènes linguistiques très fréquents tels que la métonymie – *GerE stimule sigK* peut être indifféremment décrit par *GerE stimule l'expression*

de sigK – et inversement a limité les capacités d'inférence et l'expressivité dans les ontologies pour préserver le lien au texte. Par exemple le terme *hyperthermophile bacteria* dans un document devrait être représenté non pas uniquement par le concept *hyperthermophile bacteria* mais à l'aide des concepts d'habitat, de températures normales et extrêmes pour un habitat, et d'intervalles de température, si l'ontologie doit permettre un raisonnement sur les biotopes des bactéries. Cette voie est peu explorée en lien avec l'analyse textuelle et reste un chaînon manquant pour l'exploitation des ontologies dans des applications documentaires. Elle est liée à la question montante de l'implication textuelle (*textual entailment*) en EI.

6.5 Conclusion

Nous avons présenté dans ce chapitre quelques problématiques phares de la bioinformatique qui ont bénéficié de méthodes issues de l'intelligence artificielle, et qui ont également généré de nouveaux besoins de structuration, modélisation et analyse de données et de connaissance. La biologie vit depuis une décennie une véritable révolution caractérisée par l'apparition à un rythme effréné de nouvelles technologies générant d'énormes quantités de données, et par l'importance croissante prise par les outils mathématiques et informatique pour la manipulation et l'exploitation de ces données pour en extraire de la connaissance. On peut prédire que l'intelligence artificielle aura encore une belle place à prendre pour assister le biologiste à travers cette révolution, en n'oubliant jamais - aussi bien du côté informatique que biologique - que les prédictions *in silico* ne restent que des prédictions et qu'elles doivent être confirmées expérimentalement *in vitro* ou, mieux, *in vivo*.

Remerciements : Nous souhaitons remercier pour leur aide à la rédation ce chapitre Philippe Bessières, Robert Bossy, Jacques Nicolas et Anne Siegel.

Références

AKUTSU, T., MIYANO, S. et KUHARA, S. (2000). Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function. *J. Comput. Biol.*, 7(3-4):331–343.

ALAIN-PIERRE, M., ERICK, A. et PHILIPPE, B. (2008). Genic interaction extraction by reasoning on an ontology. *In* SALAKOSKI, T., REBHOLZ-SCHUHMANN, D. et PYYSALO, S., éditeurs : *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008), Turku, Finland*. Turku Centre for Computer Science (TUCS).

ALTMAN, R. B., BERGMAN, C., BLAKE, J., BLASCHKE, C., COHEN, A., GANNON, F., GRIVELL, L., HAHN, U., HERSH, W., HIRSCHMAN, L., JENSEN, L. J., KRALLINGER, M., MONS, B., O'DONOGHUE, S. I., PEITSCH, M. C., REBHOLZ-SCHUHMANN, D., H., S. et A., V. (2008). Text mining for biology - the way forward: opinions from leading scientists. *Genome Biology*, 9:2.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. et LIPMAN, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–10.

ANANIADOU, S. et MCNAUGHT, J. (2005). *Text Mining for Biology And Biomedicine*. Artech House, Inc.

ANDREOPOULOS, B., ALEXOPOULOU, D. et SCHROEDER, M. (2008). Word sense disambiguation in biomedical ontologies with term co-occurrence analysis and document clustering. *IJDMB*, 2(3):193–215.

ANTONIOTTI, M., POLICRITI, A., UGEL, N. et MISHRA, B. (2003). Model building and model checking for biochemical processes. *Cell Biochemistry and Biophysics*, 38:271–286.

ASHBURNER, M., BALL, C., BLAKE, J., BOTSTEIN, D., BUTLER, H., CHERRY, M., DAVIS, A., DOLINSKI, K., DWIGHT, S. et EPPIG, J. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29.

BAILEY, T. L. et ELKAN, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2:28–36.

BALDI, P., BRUNAK, S., FRASCONI, P., SODA, G. et POLLASTRI, G. (1999). Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, 15:937–946.

BAY, S., SHRAGER, J., POHORILLE, A. et LANGLEY, P. (2003). Revising regulatory networks: from expression data to linear causal models. *Journal of Biomedical Informatics*, 35(289-297).

BEN-HUR, A. et NOBLE, W. S. (2005). Kernel methods for predicting protein-protein interactions. *Bioinformatics*, 21(Suppl. 1):i38–i46.

BENDTSEN, J. D., NIELSEN, H., von HEIJNE, G. et BRUNAK, S. (2004). Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, 340:783–795.

BERNAL, A., CRAMMER, K., HATZIGEORGIOU, A. et PEREIRA, F. (2007). Global discriminative learning for higher-accuracy computational gene prediction. *PLoS Comput Biol*, 3(3):e54.

BERNHART, S. H., HOFACKER, I. L., WILL, S., GRUBER, A. R. et STADLER, P. F. (2008). RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, 9:474.

BERNOT, G., COMET, J.-P., RICHARD, A. et GUESPIN-MICHEL., J. (2004). A fruitful application of formal methods to biological regulatory networks: Extending thomas' asynchronous logical approach with temporal logic. *Journal of Theoretical Biology*, 229(3):339–347.

BILLOUD, B., KONTIC, M. et VIARI, A. (1996). Palingol: a declarative programming language to describe nucleic acids' secondary structures and to scan sequence database. *Nucleic Acids Res*, 24(8):1395–403.

BLEAKLEY, K., BIAU, G. et VERT, J.-P. (2007). Supervised reconstruction of biological networks with local models. *Bioinformatics*, 23(13):i57–i65.

BONTCHEVA, K., TABLAN, V., MAYNARD, D. et CUNNINGHAM, H. (2004). Evolving GATE to meet new challenges. *Natural Language Engineering*.

BOSSY, R., KOTOUJANSKY, A., AUBIN, S. et NÉDELLEC, C. (2008). Close integration of ml and nlp tools in bioalvis for semantic search in bacteriology. *In* BURGER, A., PASCHKE, A., ROMANO, P. et SPLENDIANI, A., éditeurs : *Proceedings of the Workshop on Semantic Web Applications and Tools for Life Sciences, Edinburgh, United Kingdom, November 28*. CEUR.

BRAZMA, A., JONASSEN, I., EIDHAMMER, I. et GILBERT, D. (1998). Approaches to the

automatic discovery of patterns in biosequences. *J. Comput. Biol.*, 5:279–305.

BREJOVA, B., VINAR, T. et LI, M. (2003). Pattern Discovery: Methods and Software. *In* KRAWETZ, S. A. et WOMBLE, D. D., éditeurs : *Introduction to Bioinformatics*, chapitre 29, pages 491–522. Humana Press.

BROWN, M. P., GRUNDY, W. N., LIN, D., CRISTIANINI, N., SUGNET, C. W., FUREY, T. S., ARES, M. et HAUSSLER, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA*, 97(1):262–7.

BRUNET, J. P., TAMAYO, P., GOLUB, T. R. et MESIROV, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A*, 101(12):4164–9.

BURGE, C. et KARLIN, S. (1997). Prediction of complete gene structures in human genomic DNA. *J Mol Biol*, 268(1):78–94.

CALZONE, L., FAGES, F. et SOLIMAN, S. (2006). Biocham: An environment for modeling biological systems and formalizing experimental knowledge. *Bioinformatics*, 22:1805–1807.

CHABRIER-RIVIER, N., CHIAVERINI, M., DANOS, V., FAGES, F. et SCHÄCHTER, V. (2004). Modeling and querying biomolecular interaction networks. *Theor. Comput. Sci.*, 325(1):25–44.

CONESA, A., GOTZ, S., GARCIA-GOMEZ, J. M., TEROL, J., TALON, M. et ROBLES, M. (2005). Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676.

CORBLIN, F., TRIPODI, S., FANCHON, E., ROPERS, D. et TRILLING, L. (2009). A declarative constraint-based method for analyzing discrete genetic regulatory networks. *Biosystems*, 98(2):91–104.

CORNISH-BOWDEN, A., JAMIN, M. et SAKS, V. (2005). *Cinétique enzymatique*. EDP Sciences.

CORPET, F., GOUZY, J. et KAHN, D. (1998). The ProDom database of protein domain families. *Nucleic Acids Res.*, 26:323–326.

COVERT, M. W., XIAO, N., CHEN, T. et KARR, J. (2008). Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli. *Bioinformatics*, 18:2044–50.

DANOS, V. et LANEVE, C. (2004). Formal molecular biology. *Theoretical Computer Science*, 325(1):69 – 110.

DE JONG, H. (2002). Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9(1):67–103.

DECAPRIO, D., VINSON, J. P., PEARSON, M. D., MONTGOMERY, P., DOHERTY, M. et GALAGAN, J. E. (2007). Conrad: gene prediction using conditional random fields. *Genome Res*, 17(9):1389–98.

D'HAESELEER, P., WEN, X., FUHRMAN, S. et SOMOGYI, R. (1999). Linear modeling of mrna expression levels during cns development and injury. *In Pacific Symposium on Biocomputing*, pages 41–52.

DOJER, N., GAMBIN, A., MIZERA, A., WILCZYNSKI, B. et TIURYN, J. (2006). Applying dynamic Bayesian networks to perturbed gene expression data. *BMC Bioinformatics*, 7:249.

DSOUZA, M., LARSEN, N. et OVERBEEK, R. (1997). Searching for patterns in genomic data. *Trends Genet*, 13(12):497–8.

DURBIN, R., EDDY, S. R., KROGH, A. et MITCHISON, G. J. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

DUROT, M., BOURGUIGNON, P.-Y. et SCHACHTER, V. (2009). Genome-scalemodels of bacterial metabolism: reconstruction and applications. *FEMS Microbiol Rev*, 33:164–190.

EDDY, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, 14:755–763.

EDDY, S. R. et DURBIN, R. (1994). RNA sequence analysis using covariance models. *Nucleic Acids Res*, 22(11):2079–88.

EISEN, M. B., SPELLMAN, P. T., BROWN, P. O. et BOTSTEIN, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868.

FAURÉ, A., NALDI, A., LOPEZ, F., CHAOUIYA, C., CILIBERTO, A. et THIEFFRY, D. (2009). Modular logical modelling of the budding yeast cell cycle. *Mol. BioSyst.*, 5:1787 – 1796.

FOISSAC, S. et SCHIEX, T. (2005). Integrating alternative splicing detection into gene prediction. *BMC Bioinformatics*, 6:25.

FRIEDMAN, N., LINIAL, M., NACHMAN, I. et PE'ER, D. (2000). Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, 7(3-4):601–620.

GASPIN, C. et WESTHOF, E. (1995). An interactive framework for RNA secondary structure prediction with a dynamical treatment of constraints. *J Mol Biol*, 254(2):163–74.

GRIFFITHS-JONES, S., MOXON, S., MARSHALL, M., KHANNA, A., EDDY, S. R. et BATEMAN, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*, 33(Database issue):D121–4.

GUZIOLOWSKI, C., BOURDE, A., MOREEWS, F. et SIEGEL, A. (2009). Bioquali cytoscape plugin: analysing the global consistency of regulatory networks. *BMC Genomics*, 10.

HIRSCHMAN, L., PARK, J., TSUJII, J., WONG, L. et WU, C. (2002). Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18:1553–1561.

HULO, N., BAIROCH, A., BULLIARD, V., CERUTTI, L., CUCHE, B. A., de CASTRO, E., LACHAIZE, C., LANGENDIJK-GENEVAUX, P. S. et SIGRIST, C. J. (2008). The 20 years of PROSITE. *Nucleic Acids Res.*, 36:D245–249.

HUNTER, S., APWEILER, R., ATTWOOD, T. K., BAIROCH, A., BATEMAN, A., BINNS, D., BORK, P., DAS, U., DAUGHERTY, L., DUQUENNE, L., FINN, R. D., GOUGH, J., HAFT, D., HULO, N., KAHN, D., KELLY, E., LAUGRAUD, A., LETUNIC, I., LONSDALE, D., LOPEZ, R., MADERA, M., MASLEN, J., MCANULLA, C., MCDOWALL, J., MISTRY, J., MITCHELL, A., MULDER, N., NATALE, D., ORENGO, C., QUINN, A. F., SELENGUT, J. D., SIGRIST, C. J., THIMMA, M., THOMAS, P. D., VALENTIN, F., WILSON, D., WU, C. H. et YEATS, C. (2009). InterPro: the integrative protein signature database. *Nucleic Acids Res*, 37(Database issue):D211–5.

JAAKKOLA, T., DIEKHANS, M. et HAUSSLER, D. (2000). A Discriminative Framework for Detecting Remote Protein Homologies. *J. Comput. Biol.*, 7(1,2):95–114.

JANSEN, R., YU, H., GREENBAUM, D., KLUGER, Y., KROGAN, N., CHUNG, S., EMILI, A., SNYDER, M., GREENBLATT, J. et GERSTEIN, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–453.

JANSEN, R. C. et NAP, J. P. (2001). Genetical genomics: the added value from segregation. *Trends Genet*, 17(7):388–91.

JENSEN, L. J. J., SARIC, J. et BORK, P. (2006). Literature mining for the biologist: from

information retrieval to biological discovery. *Nature reviews. Genetics*, 7(2):119–129.

JOHN, M., LHOUSSAINE, C., NIEHREN, J. et UHRMACHER, A. (2009). The attributed pi calculus with priorities. *Transactions on Computational Systems Biology*, XII(594).

JONASSEN, I., COLLINS, J. F. et HIGGINS, D. G. (1995). Finding flexible patterns in unaligned protein sequences. *Protein Sci.*, 4:1587–1595.

JUVAN, P., DEMSAR, J., SHAUNLSKY, G. et ZUPAN, B. (2005). Genepath: from mutations to genetic networks and back. *Nucleic Acids Research*, 33(Web Server issue):W749–W752.

KALL, L., KROGH, A. et SONNHAMMER, E. L. (2004). A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, 338:1027–1036.

KAUFFMAN, S. (1971). Gene regulation networks: a theory for their global structure and behaviors. *Curr Top Dev Biol*, 6(6):145–82.

KERBELLEC, G. (2008). *Apprentissage d'automates modélisant des familles de séquences protéiques*. Thèse de doctorat, Université Rennes 1.

KHARCHENKO, P., VITKUP, D. et CHURCH, G. M. (2004). Filling gaps in a metabolic network using expression information. *Bioinformatics*, 20 Suppl 1:I178–I185.

KIM, J.-D., OHTA, T., TSURUOKA, Y., TATEISI, Y. et N., C. (2004). Introduction to the bio-entity recognition task at jnlpba. *In Proc. of NLPBA/Coling wshp*. International Conference on Computational Linguistics (COLING'04).

KING, R., ROWLAND, J., OLIVER, S., YOUNG, M., AUBREY, W., BYRNE, E., LIAKATA, M., MARKHAM, M., PIR, P., SOLDATOVA, L., SPARKES, A., WHELAN, K. et CLARE, A. (2009). The automation of science. *Science*, 324(5923):85–9.

KLAMT, S., SAEZ-RODRIGUEZ, J., LINDQUIST, J. A., SIMEONI, L. et GILLES, E. D. (2006). A methodology for the structural and functional analysis of signaling and regulatory networks. *BMC Bioinformatics*, 7:56.

KROGH, A., BROWN, M., MIAN, I. S., SJOLANDER, K. et HAUSSLER, D. (1994). Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.*, 235:1501–1531.

KUANG, R., LESLIE, C. S. et YANG, A.-S. (2004). Protein backbone angle prediction with machine learning approaches. *Bioinformatics*, 20(10):1612–1621.

LAFERRIERE, A., GAUTHERET, D. et CEDERGREN, R. (1994). An RNA pattern matching program with enhanced performance and portability. *Comput Appl Biosci*, 10(2):211–2.

LANCKRIET, G. R. G., DE BIE, T., CRISTIANINI, N., JORDAN, M. I. et NOBLE, W. S. (2004). A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635.

LARHLIMI, A. et BOCKMAYR, A. (2009). A new constraint-based description of the steady-state flux cone of metabolic networks. *Discrete Appl. Math.*, 157(10):2257–2266.

LESLIE, C., ESKIN, E. et NOBLE, W. (2002). The spectrum kernel: a string kernel for SVM protein classification. *In* ALTMAN, R. B., DUNKER, A. K., HUNTER, L., LAUERDALE, K. et KLEIN, T. E., éditeurs : *Proceedings of the Pacific Symposium on Biocomputing 2002*, pages 564–575, Singapore. World Scientific.

LUKASHIN, A. V. et BORODOVSKY, M. (1998). GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res*, 26(4):1107–15.

M RISON SCG, Hodgman TC, T. J. (2000). Comparison of functional annotation schemes for

genomes. *Funct Integr Genomics*, 1:56–69.

MACKE, T. J., ECKER, D. J., GUTELL, R. R., GAUTHERET, D., CASE, D. A. et SAMPATH, R. (2001). RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res*, 29(22):4724–35.

MAJOR, F. (2003). Building three-dimensional ribonucleic acid structures. *Computing in Science & Engineering*, pages 44–53.

MARCO, M. et FRANCESCO, P. (2006). Using gene ontology and genomic controlled vocabularies to analyze high-throughput gene lists: Three tool comparison. *Computers in Biology and Medicine*, 7-8(36):193–215.

MARGOLIN, A. A., NEMENMAN, I., BASSO, K., WIGGINS, C., STOLOVITZKY, G., DALLA FAVERA, R. et CALIFANO, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7 Suppl 1:S7.

MCCRAY, A. T., BROWNE, A. C. et BODENREIDER, O. (2002). The lexical properties of the gene ontology. *Proc AMIA Symp*, pages 504–508.

METZKER, M. L. (2010). Sequencing technologies - the next generation. *Nat Rev Genet*, 11(1):31–46.

MILLER, J. R., KOREN, S. et SUTTON, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*.

MONTEIRO, P., DUMAS, E., BESSON, B., MATEESCU, R., PAGE, M., FREITAS, A. et de JONG, H. (BMC Bioinformatics). A service-oriented architecture for integrating the modeling and formal verification of genetic regulatory networks. *2009*, 10:450.

MONTEIRO, P., ROPERS, D., MATEESCU, R., FREITAS, A. et de JONG, H. (2008). Temporal logic patterns for querying dynamic models of cellular interaction networks. *Bioinformatics*, 24(26):i227–33.

MORDELET, F. et VERT, J.-P. (2008). Sirene: Supervised inference of regulatory networks. *Bioinformatics*, 24(16):i76–i82.

MORGAN, A., LU, Z., WANG, X., COHEN, A., FLUCK, J., RUCH, P., DIVOLI, A., FUNDEL, K., LEAMAN, R., HAKENBERG, J., SUN, C., LIU, H. H., TORRES, R., KRAUTHAMMER, M., LAU, W., LIU, H., HSU, C. N., SCHUEMIE, M., COHEN, B. K. et HIRSCHMAN, L. (2008). Overview of biocreative ii gene normalization. *Genome Biology*, 9(Suppl 2).

MUELLER, H.-M., KENNY, E. E. et STERNBERG, P. W. (2004). Textpresso: An ontology-based information retrieval and extraction system for biological literature. *PLoS Biology*, 2(11):309.

MURZIN, A. G., BRENNER, S. E., HUBBARD, T. et CHOTHIA, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247:536–540.

NABIEVA, E., JIM, K., AGARWAL, A., CHAZELLE, B. et SINGH, M. (2005). Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21 Suppl 1:i302–i310.

NÉDELLEC, C., BESSIÈRES, P., BOSSY, R., KOTOUJANSKY, A. et MANINE, A.-P. (2006). Annotation guidelines for machine learning-based named entity recognition in microbiology. *In* HILARIO, M. et NEDELLEC, C., éditeurs : *Proceedings of the Data and text mining*

in integrative biology workshop, associé à ECML/PKDD, pages 40–54, Berlin, Allemagne.

NÉDELLEC, C., NAZARENKO, A. et BOSSY, R. (2009). Ontology and information extraction. *In* STAAB, S. et STUDER, R., éditeurs : *Handbook on Ontologies*, International Handbooks on Information Systems. Springer.

NUSSINOV, R. et JACOBSON, A. B. (1980). Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc Natl Acad Sci U S A*, 77(11):6309–13.

PAVESI, G., MAURI, G. et PESOLE, G. (2001). An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, 17 Suppl 1:S207–214.

PEARL, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge Univ Pr.

POLLASTRI, G. et BALDI, P. (2002). Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics*, 18 Suppl 1:S62–S70.

RAYCHAUDHURI, S., STUART, J. M. et ALTMAN, R. B. (2000). Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac Symp Biocomput*, pages 455–466.

REDER, C. (1988). Metabolism control theory : A structural approach. *J Theor Biol*, 2(135): 175–201.

REGEV, A. et SHAPIRO, E. (2002). Cells as computation. *Nature*, 419(6905):343.

RINALDI, F., SCHNEIDER, G., KALJURAND, K., KLENNER, M., HESS, M., ROMACKER, M., von ALLMEN, J.-M. et VACHON, T. (2008). Ontogene in biocreative ii. *In Proc. of the Second BioCreative Challenge Evaluation Workshop*, pages 193–198, Madrid, Spain.

ROSSI, F., VAN BEEK, P. et WALSH, T. (2006). *Handbook of constraint programming*. Elsevier Science Ltd.

ROTH, F. P., HUGHES, J. D., ESTEP, P. W. et CHURCH, G. M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mrna quantitation. *Nat. Biotechnol.*, 16(10):939–945.

SAIGO, H., VERT, J.-P., UEDA, N. et AKUTSU, T. (2004). Protein homology detection using string alignment kernels. *Bioinformatics*, 20(11):1682–1689.

SAKAKIBARA, Y., BROWN, M., HUGHEY, R., MIAN, I. S., SJOLANDER, K., UNDERWOOD, R. C. et HAUSSLER, D. (1994). Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res*, 22(23):5112–20.

SANKOFF, D. (1985). Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM Journal on Applied Mathematics*, 45(5):810–825.

SCHUSTER, S., FELL, D. et DANDEKAR, T. (2000). A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnology*, 18(3):326–332.

SCHWEIKERT, G., ZIEN, A., ZELLER, G., BEHR, J., DIETERICH, C., ONG, C. S., PHILIPS, P., DE BONA, F., HARTMANN, L., BOHLEN, A., KRUGER, N., SONNENBURG, S. et RATSCH, G. (2009). mGene: accurate SVM-based gene finding with an application to nematode genomes. *Genome Res*, 19(11):2133–43.

SETTLES, B. (2005). Abner: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192.

SMITH, L., TANABE, L., ANDO, R., KUO, C. J., CHUNG, F. I., HSU, C. N., LIN, Y. S., KLINGER, R., FRIEDRICH, C., GANCHEV, K., TORII, M., LIU, H., HADDOW, B., STRUBLE, C., POVINELLI, R., VLACHOS, A., BAUMGARTNER, W., HUNTER, L., CARPENTER, B., TSAI, R., DAI, H. J., LIU, F., CHEN, Y., SUN, C., KATRENKO, S., ADRIAANS, P., BLASCHKE, C., TORRES, R., NEVES, M., NAKOV, P., DIVOLI, A., LOPEZ, M. M., MATA, J. et WILBUR, J. W. (2008). Overview of biocreative ii gene mention recognition. *Genome Biology*, 9(Suppl 2).

SONNHAMMER, E. L., EDDY, S. R. et DURBIN, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, 28(3):405–420.

SONNHAMMER, E. L., von HEIJNE, G. et KROGH, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol*, 6:175–182.

STANKE, M., SCHOFFMANN, O., MORGENSTERN, B. et WAACK, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, 7:62.

SZYMANSKI, M., ERDMANN, V. A. et BARCISZEWSKI, J. (2007). Noncoding RNAs database (ncRNAdb). *Nucleic Acids Res*, 35(Database issue):D162–4.

THEBAULT, P., de GIVRY, S., SCHIEX, T. et GASPIN, C. (2006). Searching RNA motifs and their intermolecular contacts with constraint networks. *Bioinformatics*, 22(17):2074–80.

THOMAS, R. (1981). On the relation between the logical structure of systems and their ability to generate multiple steadt states or sustained oscillations. *Springer Ser. Synergetics*, 9:180–193.

TOUZET, H. et PERRIQUET, O. (2004). CARNAC: folding families of related RNAs. *Nucleic Acids Res*, 32(Web Server issue):W142–5.

TROYANSKAYA, O. G., DOLINSKI, K., OWEN, A. B., ALTMAN, R. B. et BOTSTEIN, D. (2003). A bayesian framework for combining heterogeneous data sources for gene function prediction (in *saccharomyces cerevisiae*). *Proc. Natl. Acad. Sci. USA*, 100(14):8348–8353.

WARD, J. J., MCGUFFIN, L. J., BUXTON, B. F. et JONES, D. T. (2003). Secondary structure prediction with support vector machines. *Bioinformatics*, 19(13):1650–1655.

WASHIETL, S. (2007). Prediction of structural noncoding RNAs with RNAz. *Methods Mol Biol*, 395:503–26.

WINGENDER, E., CHEN, X., HEHL, R., KARAS, H., LIEBICH, I., MATYS, V., MEINHARDT, T., PRUBETA, M., REUTER, I. et SCHACHERER, F. (2000). TRANSFAC: an integrated system for gene expression regulation. *Nucl. Acids Res.*, 28(1):316–319.

WITTEN, I. H. et FRANK, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann.

ZUKER, M. et STIEGLER, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9(1):133–48.

ZYTNICKI, M., GASPIN, C. et SCHIEX, T. (2008). DARN! A weighted constraint solver for RNA motif localization. *Constraints*, 13(1):91–109.