Bio-algorithmique des ARN : petite promenade aux interfaces

Yann Ponty

Bio-algorithmique des ARN : petite promenade aux interfaces

Yann Ponty*

28 décembre 2014

Cet article de dissémination scientifique est paru dans :

Yann Ponty. Bio-algorithmique des ARN : petite promenade aux interfaces. In Eric Sopena, editor, *1024 - Bulletin de la société informatique de France*, volume 4, pages 23–53. SIF - Institut Henri Poincaré, 11 rue Pierre et Marie Curie, 75231 Paris Cedex 05, Octobre 2014

1 Introduction

La bioinformatique est un champ disciplinaire consacré à l'analyse des données biologiques par des méthodes informatiques. S'agissant d'une discipline transverse, elle concerne de nombreux aspects de la recherche en informatique, allant des bases de données au calcul haute performance, en passant par l'algorithmique combinatoire. La bioinformatique aide principalement la biologie à exploiter, et parfois à établir, des relations causales entre des phénomènes apparaissant en biologie. De façon plus large, elle vient s'inscrire dans une démarche de modélisation, traditionnellement centrale à la biologie. Elle y permet alors non seulement de compléter l'éventail des méthodes expérimentales, autorisant une validation/invalidation des modèles proposés, mais offre aussi la possibilité d'inférer des modèles compatibles avec les données. Dans ce cas en particulier, elle requiert une confrontation à des grands volumes de données, parfois si énormes que des complexités temps/mémoire surlinéaires sont totalement exclues. Elle soulève donc de nombreuses questions spécifiques, inspirant des développements informatiques *amont*, motivant par exemple des études poussées de complexité algorithmique, ou de problèmes combinatoires dans des fonctions d'évaluation complexes.

La pratique quotidienne de la bioinformatique n'est cependant pas sans difficultés. L'évolution rapide des techniques, des volumes de données disponibles et des problématiques biologiques rend parfois difficile la pérennisation des approches algorithmiques. Il n'est

*Yann Ponty est Chargé de Recherche CNRS au laboratoire LIX de l'École polytechnique à Palaiseau, Yann.Ponty@lix.polytechnique.fr

d'ailleurs pas rare qu'un traitement *ad hoc* hyper-spécialisé soit préféré à un algorithme élégant à fort potentiel de généralisation, et offrant de meilleures garanties de performances, mais arrivé quelques mois trop tard, ou publié dans des revues plus théoriques que ses compétiteurs... Par ailleurs, l'appétence modérée des biologistes pour les approches trop formelles rend parfois difficile la dissémination des méthodes basées sur une algorithmique non-triviale. Une étude en 2012 de Fawcett et Higginson [11] observe par exemple une corrélation négative entre la densité d'équations et la visibilité d'une étude [1], initiant un débat passionné outre-atlantique [16, 12] sur la nécessité d'introduire un meilleur bagage formel dans les formations en biologie. Cette corrélation n'implique certes pas une causalité, mais elle illustre *a minima* une difficulté récurrente de communication aux interfaces. La dissémination de résultats ayant une dimension informatique, dans des journaux du domaine d'application, s'apparente alors parfois à un exercice de schizophrénie douce, mais très certainement passionnant et enrichissant !

1.1 Principe de parcimonie et optimisation combinatoire

En dépit de ces difficultés, de nombreux problèmes algorithmiques bien définis, à l'origine d'outils à l'impact démontré sur la biologie, jalonnent la jeune histoire de la bioinformatique. Les algorithmes, et plus particulièrement les algorithmes d'optimisation combinatoire, y jouent un rôle prépondérant. Des problèmes d'optimisation combinatoire complexes apparaissent en effet très naturellement dans tous les contextes où plusieurs explications existent pour une observation empirique. Le principe de parcimonie, aussi parfois appelé *rasoir d'Ockham*, amène à privilégier, toutes choses étant égales par ailleurs, l'explication la plus simple (ou la plus probable) pour un phénomène observé. L'ensemble des explications admissibles pour l'observé peut être alors assimilé à un espace de recherche, sur lequel on cherchera à optimiser le nombre d'événements atomiques. Ce principe se généralise au principe du maximum de vraisemblance en associant à chaque explication une probabilité, ou vraisemblance, qu'on cherchera à maximiser. Les deux notions coïncident évidemment quand tous les événements sont indépendants et équiprobables, mais ce paradigme permet aussi de modéliser des phénomènes de dépendances, potentiellement complexes, entre les événements.

Un exemple emblématique est celui de l'alignement de séquences, que l'on présente traditionnellement comme la mise en correspondance des positions de deux séquences, de façon à mettre en évidence leurs similarités et différences. Un point de vue parcimonieux sur ce problème consiste à considérer les deux séquences comme homologues, c'est-à-dire des versions contemporaines d'une même séquence ancestrale. On suppose alors que les deux séquences ont évolué indépendamment, chacune subissant une suite d'opérations atomiques (insertion/délétion/mutation d'un caractère). Imaginons maintenant que la séquence ancestrale soit indisponible, et que nous ne disposions que des deux séquences produites par l'évolution. Toutes choses étant égales par ailleurs, un principe de parcimonie nous amènera à favoriser une suite d'opérations faisant apparaître un nombre minimal de modifications. Ce nombre correspond alors exactement à la distance de Le-

1. Environ 28% de citations en moins, en moyenne, pour chaque équation supplémentaire par page !

venshtein, efficacement calculable par programmation dynamique dès les années 60 [22]. En modifiant très légèrement les équations de programmation dynamique utilisées pour son calcul, de façon à maximiser la vraisemblance, on obtient l'algorithme de Smith-Waterman [34], qui permet de détecter des similarités locales. Ce dernier fut jadis l'un des algorithmes les plus utilisés de la bioinformatique (cité plus de 7 000 fois), au point de faire l'objet, à son heure de gloire, d'implémentations matérielles dédiées sous la forme de serveurs spécialisés.

Ce paradigme est aussi omniprésent dans les travaux ayant trait à la phylogénie, champs disciplinaire s'attachant à comprendre les mécanismes de l'évolution en inférant des relations d'ancestralité entre espèces, gènes et autres entités biologiques. Ici, l'un des challenges est la reconstruction d'une phylogénie, c'est-à-dire l'inférence d'un arbre de la vie cohérent avec les données contemporaines, tout en induisant un nombre minimal d'événements évolutifs. Par exemple, si l'on souhaite baser la reconstruction sur la séquence associée à un gène (ARN, protéine...) présent dans une famille d'organismes, on pourra choisir une phylogénie induisant un nombre minimal de mutations, insertions et délétions. Des principes similaires gouvernent les algorithmes de prédiction des réarrangements génomiques, l'inférence des réseaux biologiques, la reconstruction des séquences ancestrales, les études d'association génotype/phénotype...

2 Algorithmique des ARN

Les travaux effectués depuis les années 1970 sur la structure secondaire de l'ARN constituent un exemple d'interaction pluridisciplinaire particulièrement fructueuse, à l'interface entre mathématiques discrètes, biologie, physique, chimie et informatique.

2.1 L'ARN, une molécule polyvalente à l'origine de la vie ?

Les Acides RiboNucléiques (ARN) sont des molécules composées de nucléotides Adénine, Cytosine, Guanine et Uracile, assimilables à des séquences sur un alphabet $\{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{U}\}$. Ils sont transcrits comme des copies de portions de l'ADN, les Thymines de ce dernier y étant remplacées par l'Uracile ($\mathsf{T} \rightarrow \mathsf{U}$). Les molécules d'ARN sont de tailles très variables, et sont encodées par des séquences de tailles variant de 20 nucléotides (nts) à 3 000 nts environ dans la cellule, et pouvant même atteindre jusqu'à 30 000 nts pour les génomes entiers, constitués d'ARN, de certains virus comme le coronavirus SARS-CoV. Cette diversité de taille est la conséquence directe de la grande variété de rôles joués par l'ARN au sein de la cellule, parmi lesquels :

— Médiateur de l'information génétique : les ARN messagers encodent les protéines, synthétisées comme des chaînes d'acides aminés, chaque acide aminé étant encodé par un triplet de nucléotides. De façon assez remarquable, un unique ARN messager peut parfois encoder simultanément plusieurs protéines fonctionnelles. Plusieurs mécanismes (épissage alternatif, décalage de phase, translecture...) contribuent à ce non-déterminisme, influencé à la fois par des motifs présents dans la séquence,

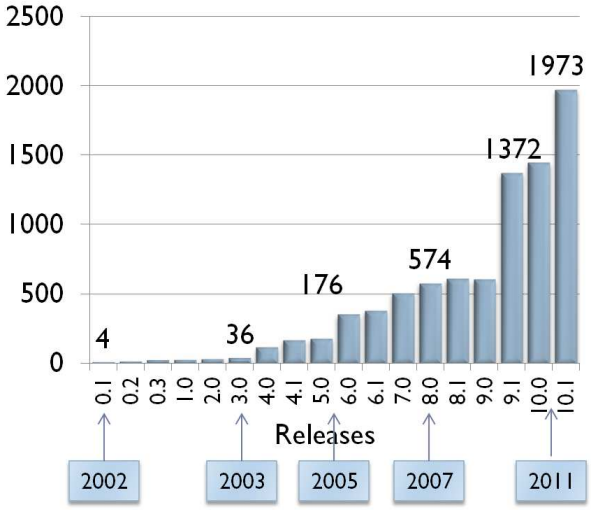


FIGURE 1: Evolution du nombre de familles d'ARN répertoriées au sein de la base de donnée RFAM [17]. Cette croissance est loin d'être ralentie, et la version 11.0 de la base, disponible depuis août 2012, s'est élargie à 2208 familles d'ARN.

la structure, ou encore les concentrations relatives des différents acteurs du monde cellulaire ;

— Partie-prenante de la machinerie traductionnelle : outre leur rôle de *code source*, les ARN sont aussi acteurs de la traduction des ARN en protéines. Ils participent ainsi au ribosome, un énorme complexe moléculaire composé de nombreuses protéines et d'ARN ribosomaux, qui décodera les ARN messagers en protéines. Des ARN de transfert fournissent aussi la matière première au ribosome, en lui apportant les acides aminés, véritables *briques de base* qui, assemblés linéairement, formeront les protéines ;

— Acteur de la régulation : non seulement les ARN rendent possible la synthèse des protéines, mais ils en gouvernent aussi l'expression quantitative. Par exemple, des petits ARN simple-brin sont au cœur de l'interférence à ARN, un mécanisme de régulation au cours duquel ils se fixent sur des ARN messagers et y empêchent la fixation du ribosome, inhibant *in fine* la synthèse des protéines encodées.

Cette liste est loin d'être exhaustive, et la base de données RFAM [17], qui recense et organise en familles fonctionnelles les ARN identifiés dans la littérature, a connu une croissance explosive depuis sa création en 2002, comme l'illustre la Figure 1. De plus, l'initiative ENCODE [10], qui a analysé systématiquement 1% du génome humain, a révélé en 2007 qu'une grande majorité (environ 93%) de celui-ci est transcrit en ARN. En extrapolant sur l'ensemble du génome, et en excluant les 30% de gènes codant pour des protéines, dont 2% directement pour les acides aminés, il subsiste une foule d'ARN synthétisés, la plupart sans fonction connue à l'heure actuelle. Si certaines de ces séquences sont sans doute le produit de mécanismes stochastiques assimilables à un *bruit*

de fond, il est difficile d'imaginer l'avantage sélectif procuré par une synthèse coûteuse d'ARN à une si grande échelle. Il est donc raisonnable d'anticiper la découverte, dans les décennies qui viennent, de nouvelles fonctions, modes d'action et familles pour l'ARN, à l'instar des 197 nouvelles familles de longs ARN non-codants récemment identifiées [3] dans des régions auparavant ignorées car ne codant pas pour des gènes identifiés.

La diversité fonctionnelle de l'ARN est telle que celui-ci constitue actuellement le support biochimique le plus probable à l'origine de la vie ! En effet, les mécanismes du vivant reposent sur une dichotomie entre gènes qui encodent les éléments fonctionnels et enzymes actrices responsables de la réplication du matériel génétique. Chez les organismes contemporains, les gènes sont codés par l'ADN, et les protéines assument une grande partie des rôles enzymatiques. Or, les capacités enzymatiques de l'ADN sont trop limitées pour permettre une réplication efficace, et les protéines se prêtent mal à une auto-réplication fidèle. Cette situation soulève donc une question cruciale :

> Comment *démarrer* un système (la vie) dont chacun des deux acteurs principaux (ADN et protéines) semble avoir besoin de l'autre pour exister et se reproduire ?

L'ARN résout cet apparent *paradoxe de l'œuf et de la poule* en cumulant capacité de stockage, les génomes de certains virus étant par exemple entièrement constitués d'ARN, et fonctions enzymatiques autorisant l'auto-réplication. Un *monde prébiotique à ARN* constitue donc actuellement l'hypothèse la plus séduisante pour l'origine de la vie, et fait l'objet des travaux de toute une communauté mêlant biologistes moléculaires et biochimistes [4].

2.2 Comprendre la structure de l'ARN pour en comprendre la fonction

À la différence des ADN, les ARN sont synthétisés sous la forme d'une copie simple, aussi appelée simple-brin, et n'adoptent donc pas nécessairement la structure en double-hélice qui caractérise l'ADN. Au contraire, l'ARN adopte des formes complexes en se repliant sur lui-même. Il est ainsi soumis à des fluctuations à échelle nanométrique, et se retrouve stabilisé dans des conformations par l'appariement de certains de ces nucléotides via la formation de liaisons hydrogènes. Bien que presque toutes les paires de nucléotides soient susceptibles de former un appariement, les plus stabilisatrices sont les paires dites canoniques, parmi lesquelles on distingue les paires Watson-Crick (G–C ou A–U) et Wobble (G–U). Il résulte de ce processus une (parfois plusieurs) forme tridimensionnelle complexe essentielle, chez la plupart des ARN, pour la réalisation d'une fonction spécifique.

Il est généralement admis que le processus du repliement des ARN s'effectue de façon hiérarchique [35] comme l'illustre la Figure 2. Initialement, l'ARN se replie sur lui-même, établissant un ensemble d'appariements canoniques sans croisement, donnant lieu à une structure arborescente appelée la structure secondaire. Cette dernière peut être représentée de nombreuses façons équivalentes. Dans un deuxième temps, la forme tri-dimensionnelle adoptée par l'ARN rend possible la formation de motifs plus faiblement stabilisateurs. Parmi ceux-ci, on trouve des liaisons non-canoniques et des motifs topologiques complexes appelés pseudo-nœuds, constitués de couples d'appariements en

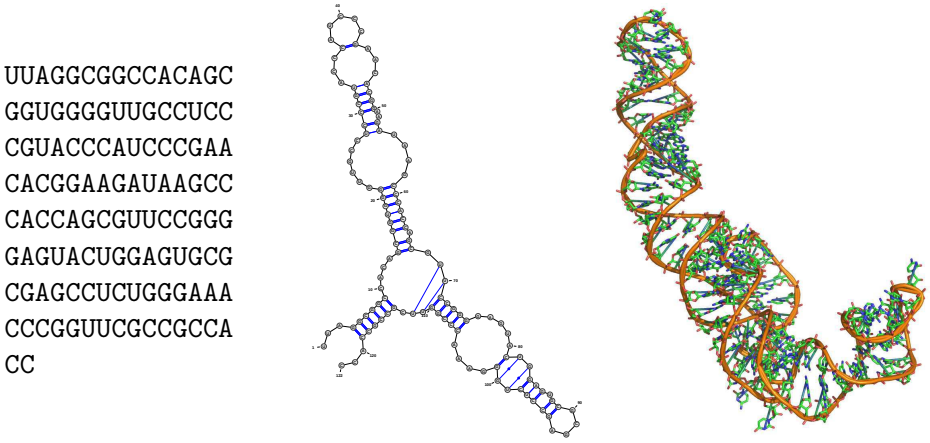
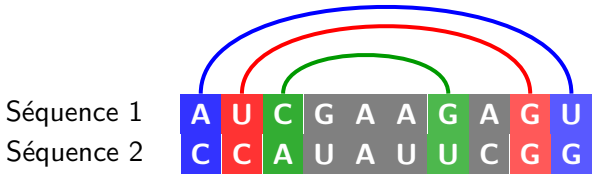


FIGURE 2: Trois principaux niveaux de représentation pour un ARN ribosomal (Id. PDB : 1K73 :B) correspondant aux étapes de structuration dans l'hypothèse d'un repliement hiérarchique.

situation de croisement quand dessinés sur le demi-plan supérieur. La prédiction de la forme adoptée par un ARN à l'issu de son repliement commence donc souvent par la prédiction d'une (ou plusieurs) structure(s) secondaire(s) candidate(s) pour celui-ci. Cette prédiction initiale est ensuite complétée par la modélisation des éléments supplémentaires, et enfin par l'agencement tri-dimensionnel de celle-ci.

Chez les ARN non-codants, qui ne sont pas traduits en protéines, la structure secondaire constitue souvent le principal déterminant de la fonction. La structure est en effet tout autant, voire parfois plus, préservée par l'évolution que la séquence précise des nucléotides. Comme le montre l'*exemple jouet* ci-dessous, deux séquences très différentes peuvent servir de support à une même structure secondaire, à travers des substitutions locales préservant les appariements :



La structure secondaire est donc au cœur de nombreuses approches basées sur la conservation d'une structure commune parmi une famille de molécules homologues. Par exemple, elle aide à rechercher de nouvelles occurrences d'un ARN dans un ou plusieurs génomes, ou encore à prédire le repliement commun à un ensemble de séquences d'ARN. La prédiction de la structure secondaire fonctionnelle d'un ARN constitue donc une première étape indispensable pour comprendre son (ou ses) mode(s) d'action(s), et le

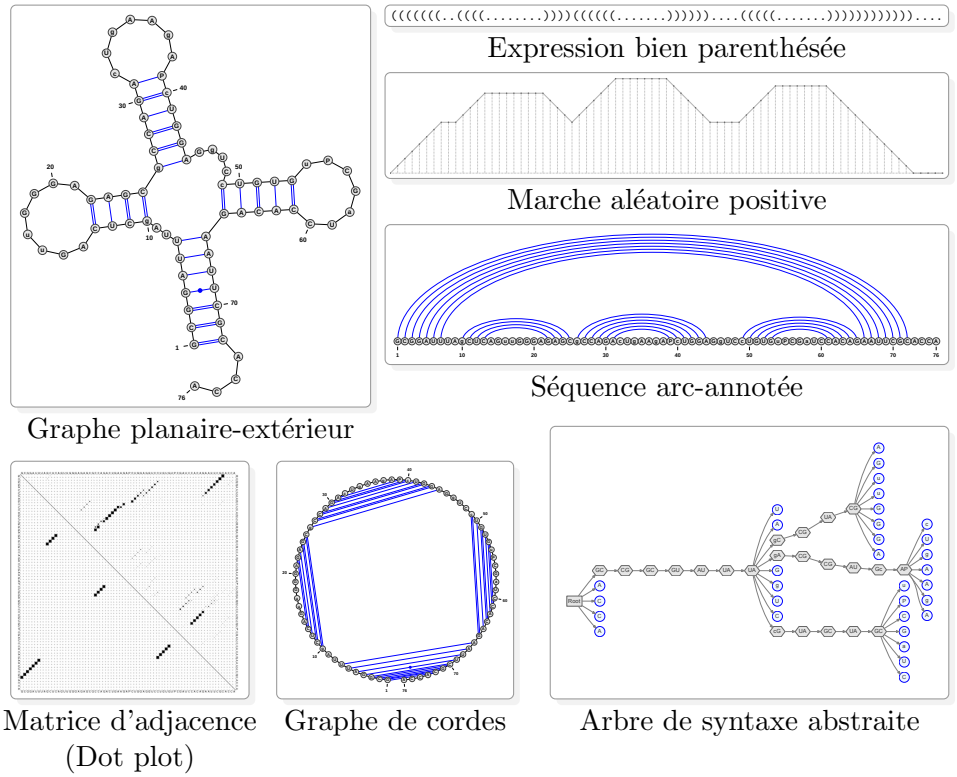


FIGURE 3: Représentations alternatives (équivalentes) de la structure secondaire d'un ARN de transfert.

replacer dans le contexte plus large des systèmes biologiques.

2.3 Représentations et propriétés combinatoires de la structure secondaire d'ARN

Du fait de la contrainte de non-croisement, la structure secondaire est analogue à un arbre, une propriété fort sympathique d'un point de vue algorithmique. Par ailleurs, des contraintes à la fois géométriques et biochimiques interdisent l'appariement canonique de positions consécutives dans un ARN, matérialisant une difficulté des biopolymères à effectuer des *demi-tours complets* dans leur trajectoire tridimensionnelle. Il en résulte donc une contrainte de distance minimale $\theta \geq 1$ entre des positions appariées.

La structure arborescente sous-jacente à la structure secondaire autorise de nombreuses représentations et caractérisations équivalentes (voir Figure 3), fournissant une interface naturelle avec de nombreuses communautés de recherche en informatique et en mathématiques discrètes :

1. un mot de Motzkin, c'est-à-dire une expression bien parenthésée sur l'alphabet $\{(,), \bullet\}$, dans laquelle on interdit le motif `()` ;
2. un graphe planaire extérieur biconnexe de degré maximal $\Delta \leq 3$;

3. un arbre unaire-binaire, dont les pères des feuilles gauches sont unaires ;
4. un arbre de syntaxe abstraite associé à la grammaire hors-contexte

$$S \to (\ S^{\overline{\varepsilon}}\ )\ S \mid \bullet\, S \mid \varepsilon \qquad\qquad S^{\overline{\varepsilon}} \to (\ S^{\overline{\varepsilon}}\ )\ S \mid \bullet\, S\ ;$$

5. un ensemble stable de sommets (non-adjacents) d'un graphe de cordes, défini comme le graphe d'intersection d'un ensemble de cordes dessinées sur un cercle ;
6. une marche (aléatoire) positive à pas dans l'ensemble

$$\{\mathtt{U} : (+1, +1),\ \mathtt{D} : (+1, -1),\ \mathtt{H} : (+1, 0)\},$$

commençant et finissant à l'ordonnée 0, et excluant le motif *pyramidal* `UD`...

Ces représentations diverses permettent la formation d'intuitions algorithmiques complémentaires, ainsi que l'utilisation d'approches classiques dans différentes disciplines, afin de mieux comprendre et analyser ces objets. Par exemple, à l'interface des mathématiques discrètes et de la biologie, l'arsenal de la combinatoire énumérative/analytique peut être utilisé pour prouver que le nombre de structures secondaires compatibles avec un ARN est, en moyenne, exponentiel sur la taille de celui-ci [2].

2.4 Prédiction du repliement par maximisation du nombre d'appariements

Une première question algorithmique se pose naturellement : Comment prédire la structure secondaire la plus stable pour une séquence d'ARN donnée ? À l'équilibre thermodynamique, celle-ci est en effet la plus probablement observée, et il est donc raisonnable d'imaginer que l'évolution ait favorisé celle-ci comme conformation fonctionnelle. Le problème d'optimisation combinatoire associé est appelé prédiction de la structure secondaire par minimisation de l'énergie libre, notion sur laquelle nous reviendrons dans la section suivante. Il exclut *a priori* toute approche gloutonne, et l'explosion du nombre de structures secondaires compatibles avec une séquence d'ARN interdit une énumération exhaustive des structures secondaires candidates.

À la fin des années 70, un premier algorithme est proposé par Nussinov *et al* [29] dans un modèle où le nombre d'appariements canoniques est considéré comme indicateur de la stabilité. La minimisation de l'énergie libre y revient donc à maximiser le nombre de paires de bases. Le problème informatique associé peut alors être résolu efficacement par une technique appelée programmation dynamique : on commence par calculer efficacement le nombre maximal d'appariements pour l'ARN par le biais d'une récurrence, puis on reconstruit une structure secondaire possédant de tels appariements en retraçant les étapes de la récurrence ayant permis d'obtenir le score maximal.

Plus précisément, l'algorithme de Nussinov part de l'observation que toute structure secondaire, compatible avec une séquence d'ARN donnée, possède (au moins) une des propriétés suivantes :

2. Plus précisément, le nombre de structures secondaires compatibles avec un ARN composé de n nucléotides admet un équivalent asymptotique de l'ordre de $1.87^n/n\sqrt{n}$ [42].

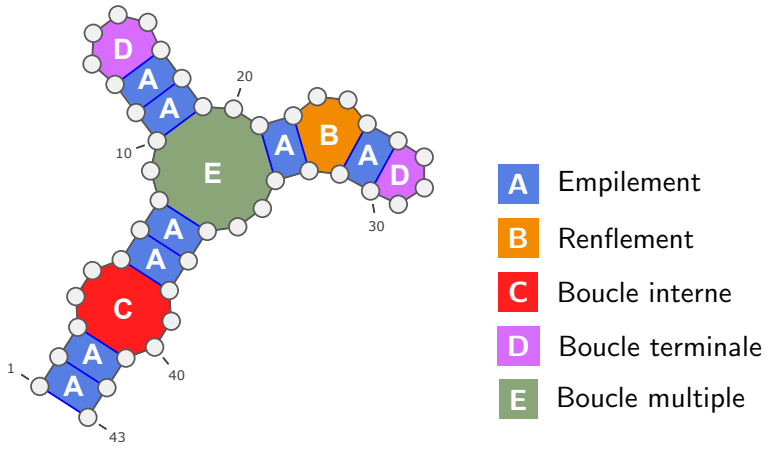


FIGURE 4: Principaux éléments de la décomposition en k-boucles d'une structure secondaire d'ARN. Un appariement *virtuel* peut être ajouté entre les premières et dernières positions afin de prendre en compte d'éventuels nucléotides non-appariés sur la face extérieure.

— les premier et dernier nucléotides sont appariés, sous réserve de compatibilité des nucléotides présents à ces positions ;

— le premier (resp. dernier) nucléotide est non-apparié ;

— la structure secondaire se partitionne en deux sous-structures indépendantes, c'est-à-dire ne partageant pas d'appariement.

Il est alors possible d'étendre ce raisonnement à toute sous-séquence définie sur un intervalle de la séquence d'ARN. On en déduit une récurrence permettant de calculer, en temps/mémoire $\Theta(n^3)/\Theta(n^2)$, le nombre maximal de paires de bases réalisables sur chaque intervalle d'une séquence d'ARN composée de n nucléotides. Une fois ces valeurs calculées, il est possible de retracer les étapes du calcul pour retrouver l'ensemble de paires de bases ayant contribué à maximiser le nombre d'appariements, et d'en déduire ainsi une structure secondaire optimalement stable.

De façon intéressante, l'algorithme proposé par Nussinov est extrêmement similaire à celui proposé indépendamment par Gavril [15] pour la recherche d'un ensemble stable de cardinalité maximale dans les graphes de cordes. On y retrouve aussi une spécialisation de l'algorithme CYK [20] permettant l'analyse syntaxique à partir de grammaires algébriques, dans sa version pondérée populaire en traitement automatique du langage. Ce lien entre bioinformatique des ARN et traitement du langage naturel est aussi à l'origine d'avancées récentes.

2.5 Au-delà du discret : modèle de Turner et MFold

Dans le modèle de stabilité considéré par l'algorithme de Nussinov, seules les paires de bases contribuent à la stabilité d'un repliement. Or la réalité biochimique des ARN

est bien plus complexe, et l'ARN est stabilisé par de nombreux motifs structurels. Ceux-ci interviennent dans le calcul de l'énergie libre, définie de façon relative, une structure étant d'autant plus stable que son énergie est négative.

En particulier, le modèle de Turner décompose la structure secondaire en occurrences de motifs structure/séquence, associés à des contributions énergétiques, comme l'illustre la Figure 4. Ces dernières sont typiquement mesurées empiriquement, au travers d'ingrates et exhaustives expériences systématiques de spectrophotométrie [37]. On obtient alors l'énergie libre d'une structure secondaire en additionnant les contributions des motifs y apparaissant.

Un fossé menace alors de se creuser entre les biochimistes, qui utilisent en pratique le modèle de Turner de façon heuristique et largement manuelle, et les proto-bioinformaticiens (informaticiens et mathématiciens de la première heure) de l'ARN, qui privilégient une optimisation exacte dans un modèle désormais considéré comme naïf. Heureusement, en 1981, Zuker et Stiegler [43] réconcilient le domaine, en remarquant que le modèle de Turner peut être complètement pris en compte via une modification subtile de l'algorithme de Nussinov.

Plus précisément, ces auteurs proposent une décomposition canonique des structures secondaires en un ensemble couvrant de k-boucles, illustrée par la Figure 4. Ils remarquent alors que les contributions énergétiques primitives du modèle de Turner peuvent être associées aux occurrences des k-boucles dans la décomposition. Ils en déduisent alors un ensemble de récurrences permettant de calculer l'énergie libre minimale sur chaque intervalle, puis un algorithme de programmation dynamique permettant la reconstruction de la structure d'énergie libre minimale dans le modèle de Turner. En imposant des restrictions réalistes sur les tailles de certains motifs, le calcul des récurrences peut être effectué en temps/mémoire $\Theta(n^3)/\Theta(n^2)$. On obtient alors l'algorithme `MFold` [43], un des plus grands succès de la bioinformatique, dont les cinq principales publications totalisent plus de 15 000 citations[3], témoignant d'un usage quasi-quotidien du logiciel par les biologistes et bioinformaticiens de l'ARN.

2.6 Structures sous-optimales et paradigme thermodynamique

Malgré la grande popularité de `MFold`, largement méritée de part sa rapidité et la fiabilité générale de ses prédictions — allant de 60% [14] à 73% [25] de sensibilité en moyenne selon les bases de séquences considérées —, la méthode rencontre des difficultés pour prédire certaines classes d'ARN :

— Certaines de ces difficultés sont de nature intrinsèque, et liées à un modèle d'énergie établi expérimentalement, mettant en jeu des valeurs extrapolées à partir de plus petits motifs ou de températures/concentrations ioniques différentes des conditions biologiques ;

— Par ailleurs, l'ARN se replie *in vivo* en interaction avec son environnement, constitué de protéines, d'ARN et de petites molécules, qui peuvent *guider* son repliement vers des structures *a priori* peu stables. Certaines catégories d'ARN voient même

3. Source : Google Scholar, avril 2014.

certains de leur nucléotides modifiés par des enzymes, perturbant ainsi leurs propriétés physico-chimiques, ce qui peut déstabiliser certaines structures ;

— Enfin, la structure secondaire, en ignorant les couples d'appariements en situation de croisement (pseudo-nœuds), néglige des contributions potentiellement stabilisatrices.

Tous ces facteurs sont susceptibles de renverser l'ordre relatif de stabilité, perçu par l'algorithme, pour deux structures secondaires compatibles avec une séquence.

Une première approche pour contourner ce problème consiste à considérer les structures sous-optimales. Zuker propose en 1989 un premier algorithme [40] pour la génération des structures P-optimales, l'ensemble des structures stables faisant apparaître certains appariements jugés crédibles. Cette définition possède comme mérite manifeste d'éviter une explosion du nombre de structures candidates. Elle interdit cependant une étude systématique de l'ensemble des sous-optimaux, et ne permet pas de déterminer, par exemple, si la structure d'énergie minimale est isolée ou, au contraire, si elle est concurrencée par d'autres structures. Des travaux de l'école Viennoise de biochimie théorique viendront, une décennie plus tard, compléter cette première définition en proposant un algorithme de génération exhaustive des structures secondaires sous-optimales à différence d'énergie [39].

Une alternative très prisée actuellement propose un changement de paradigme complet, basé sur un concept central de la physique statistique, l'équilibre de Boltzmann. En effet, en focalisant sur la structure d'énergie minimale, on oublie que le phénomène du repliement est un phénomène intrinsèquement stochastique. La stabilité d'un repliement est certes importante, mais la probabilité d'observer un ARN dans une topologie donnée dépend aussi du nombre de conformations de stabilités comparables, présentes au sein des structures possibles pour l'ARN. En postulant un équilibre thermodynamique de Boltzmann, on suppose que l'ARN est resté suffisamment longtemps dans des conditions stables, et que la probabilité d'observer une structure secondaire $S \in \mathcal{S}_\omega$ pour un ARN ω est donnée par

$$\mathbb{P}(S \mid \omega) = \frac{e^{\frac{-E_\omega(S)}{kT}}}{\mathcal{Z}_\omega} \qquad \{\text{Distribution de Boltzmann}\}$$

$$\text{où} \quad \mathcal{Z}_\omega = \sum_{S \in \mathcal{S}_\omega} e^{\frac{-E_\omega(S)}{kT}}, \qquad \{\text{Fonction de partition de } \omega\}$$

où k est la constante de Boltzmann et T la température en Kelvins. L'équilibre de Boltzmann peut aussi être interprété comme la distribution stationnaire d'un processus Markovien modélisant la dynamique du repliement.

En pondérant exponentiellement les structures selon leur énergie, cette distribution opère un compromis subtil entre la stabilité des molécules et l'explosion combinatoire du nombre de conformations sous-optimales. La structure d'énergie minimale reste certes la plus probable, mais sa probabilité décroît exponentiellement sur la taille des séquences considérées. Il en résulte une diversité de paysages thermodynamiques bien plus riche que ne l'aurait laissé présager le simple examen de la structure d'énergie minimale. Une

telle distribution permet alors de donner un sens mathématique rigoureux aux questions suivantes :

— Le processus du repliement est-il *bien défini* ? [41]

— Combien y a-t-il de conformations *principales* dans l'ensemble ? [9]

— Quels sont les appariements récurrents dans l'ensemble de Boltzmann ? [26]

— Fournissent-ils une indication de qualité pour les prédictions ? [25]

— Quelles mutations sont les plus susceptibles de perturber la fonction ? [38]

— Quelle distribution de distance à une structure de référence ? [13]

Le verrou algorithmique principal pour répondre à ces questions concerne le calcul de la fonction de partition, *a priori* complexe car défini ci-dessus comme la somme d'un nombre exponentiel de termes. Or, John McCaskill remarque en 1990 [26] que, sous certaines hypothèses techniques, un schéma de programmation dynamique pour la minimisation d'énergie peut être transformé *syntaxiquement*, à travers un simple changement d'algèbre, en un algorithme d'efficacité comparable pour calculer la fonction de partition. Des réponses algorithmiques aux questions posées ci-dessus sont alors obtenues par des modifications de ce schéma, ou à travers des méthodes d'échantillonnage exact elles-aussi dérivées du calcul de la fonction de partition.

2.7 L'évolution comme alliée : approches comparatives

Malgré les gains substantiels, à la fois en qualité des prédictions et en quantification de la confiance des prédictions [25], induits par ce changement de paradigme, la recherche d'un repliement hautement précis pour un nouvel ARN restait un problème ardu. Heureusement, à l'heure où les approches énergétiques atteignaient leurs limites, des *approches comparatives*, l'exploitation des propriétés évolutives de l'ARN, allaient permettre un nouveau bond en avant. En effet, l'hypothèse d'une relation forte entre structure et fonction motive les études structurelles mais, en renversant le point de vue, permet aussi de postuler que des ARN *homologues*, ayant une fonction commune, partagent très probablement une structure commune.

On peut donc déjà exploiter cette remarque pour améliorer notre capacité de prédiction, dès lors que des ARN homologues sont identifiés. On se pose alors le problème de la prédiction d'une structure secondaire commune à deux ou plusieurs séquences. Si celles-ci sont déjà alignées, alors l'exécution d'un algorithme de type `MFold`, légèrement étendu pour incorporer les contributions simultanées des séquences, permet la prédiction d'une structure consensus pour une famille d'ARN préalablement alignés [18]. Cependant, cette situation est assez rare, et l'absence de contraintes fonctionnelles fortes s'exprimant sur la séquence entraîne souvent une évolution rapide de celle-ci. La prédiction d'un alignement est alors délicate tant que la structure commune reste inconnue, or la prédiction de celle-ci est justement notre objectif premier !

Pour sortir de ce paradoxe de l'œuf et de la poule, David Sankoff propose en 1985 [32] un algorithme pour le repliement/alignement simultané d'une famille de séquences optimisant la somme des scores d'alignement et des énergies de repliement. De complexité

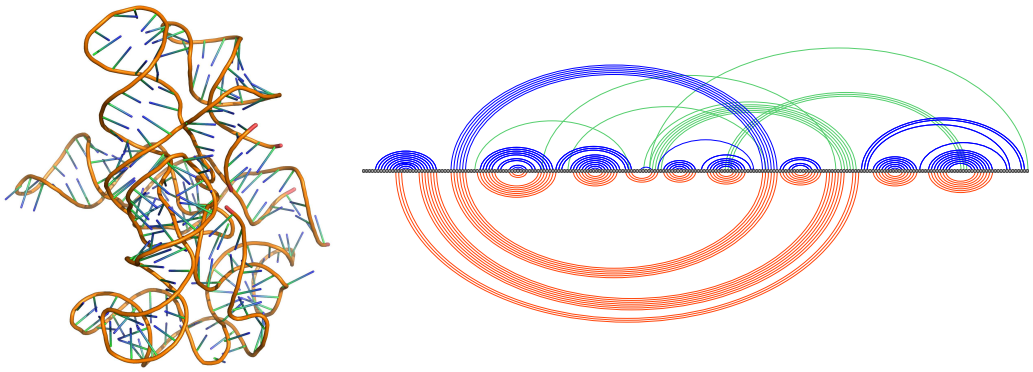


FIGURE 5: Pour ce ribozyme du groupe I (PDB [5] ID : 1Y0Q :A, gauche), les performances d'une prédiction par minimisation d'énergie sont affectées par l'omission des appariements en situation de croisement. Un logiciel de type `MFold` prédira donc des appariements (bas) substantiellement différents, à la fois de ceux lisibles dans la structure 3D (haut), et de leur sous-ensemble sans croisement de cardinalité maximale (haut, trait gras).

exponentielle pour un nombre non-borné de séquences, il se simplifie en un algorithme en $\Theta(n^6)/\Theta(n^4)$ pour deux séquences. Malgré sa complexité élevée, cette catégorie d'approches permet d'obtenir un tel gain de qualité sur les prédictions réalisées qu'elle a aujourd'hui la faveur des bioinformaticiens de l'ARN [14]. Cette popularité a d'ailleurs motivé de nombreux compromis algorithmiques afin de rendre réaliste l'exécution de cette approche à grande échelle [36].

Cette approche comparative est aussi utilisée afin de rechercher des homologues, dans les génomes, à des ARN issus d'une famille déja identifiée. Une fois un alignement obtenu, il est alors possible de construire une structure secondaire consensus pour la famille. En conjonction avec l'alignement multiple, cette dernière permet alors la construction d'un modèle de covariation [28], c'est-à-dire une grammaire stochastique grâce à laquelle de nouvelles séquences peuvent être rapatriées des bases de données publiques. Ces nouvelles séquences viennent alors enrichir l'alignement, amorçant un cercle vertueux qui a rendu possible la caractérisation de nombreuses familles d'ARN, et constituent donc l'un des piliers de la base de données RFAM [17].

2.8 Enrichir les espaces de conformations

Parallèlement à ces développements, une direction de recherche orthogonale s'est développée, visant à explorer des espaces de recherche plus riches. L'hypothèse d'un repliement hiérarchique, initialement sans croisement, n'est en effet pas une règle absolue, et de nombreux ARN contiennent des nœuds, pseudo-nœuds et autres motifs topologiques complexes (voir Figure 5). Ces motifs, qui sont représentés par des ensembles d'appariements en croisement, sont parfois indispensables pour qu'un ARN remplisse sa fonction. De plus, ils influencent fortement la modélisation 3D en limitant ses degrés

de liberté, et leur prédiction constitue donc une étape préliminaire essentielle vers des prédictions de structure à résolution atomique.

Cependant, autoriser des appariements en croisement soulève des problèmes algorithmiques importants. Dans le modèle de Nussinov maximisant le nombre de paires de bases, ou leur poids total, le problème de prédiction d'une structure minimisant l'énergie avec pseudo-nœuds peut être résolu en temps polynomial comme une instance particulière de la recherche, dans un graphe, d'un ensemble stable (pondéré) maximal [8]. Cependant, le problème devient NP-difficile [24] dès qu'un modèle d'énergie, analogue à celui de Turner, est pris en considération, et le reste même dans un modèle se limitant aux empilements d'appariements [23]. Enfin, les modèles d'énergie les plus expressifs semblent même exclure (à moins que $\mathsf{P} = \mathsf{NP}$) tout schéma d'approximation polynomial de ratio constant [33].

Cette situation, en apparence désespérée, a suscité des travaux où l'espace de recherche se prête *par construction* à la programmation dynamique, le tout en couvrant la plus grande proportion possible des structures rencontrées dans la littérature. Akutsu [1] propose ainsi un algorithme exact en $\Theta(n^4)$ permettant de prédire des pseudo-nœuds simples. S'ensuivent alors de nombreux travaux cherchant à optimiser le compromis expressivité/efficacité. En particulier, des contributions élégantes, issues de la physique théorique, proposent de considérer le genre topologique, un paramètre associé à la carte obtenue en dessinant tous les appariements de la structure secondaire dans le demi-plan supérieur [6]. On constate alors que les structures connues ont une valeur faible pour le genre, motivant la conception d'algorithmes pour le repliement en genre limité. Le problème est cependant loin d'être résolu, et fait encore l'objet de développements supplémentaires, tant sur l'algorithmique qu'au niveau des modèles d'énergie et espaces de conformations sous-jacents, dans un dialogue fructueux à l'interface de la biochimie et de l'optimisation combinatoire.

Des travaux très similaires permettent aussi de prédire les interactions ARN-ARN. En effet, celles-ci se modélisent assez naturellement comme la conjonction de deux repliements *internes* à chacun des ARN, et d'un ensemble d'appariements *interactions*, reliant des positions laissées libres dans les deux structures. Malheureusement, le problème de la prédiction simultanée des repliements et interactions devient alors NP-difficile, même dans sa version purement combinatoire, où l'on cherche à maximiser le nombre d'appariements, où chacune des deux structures *internes* est une simple structure secondaire, et où les appariements en interaction sont eux-mêmes sans croisement [2]. De même que pour la prédiction des pseudo-nœuds, la communauté recherche actuellement une stratégie de contournement pour ce résultat négatif, explorant des approches heuristiques [7] et des résolutions exactes sur des espaces de recherche restreints [27].

Une autre direction pour l'extension des espaces de conformations consiste à considérer, non seulement les traditionnels appariements G-C, A-U et G-U, mais aussi les appariement non-canoniques [21]. Bien que ces derniers soient plus rares, et *a priori* moins contributifs à la stabilité d'une conformation, ils s'organisent en modules bien identifiés, récurrents, conservés par l'évolution, et pour la plupart fonctionnels. De plus, leur prédiction fournirait un réseau de contraintes plus riches, permettant d'éliminer des degrés de liberté

dans la recherche d'une conformation 3D. Un modèle de conformation étendant la notion de k-boucle a en effet été proposé en 2008 par Major et Parisien [30], permettant alors de reconstruire la conformation 3D d'ARN connus à des résolutions record. Ces travaux prometteurs peinent cependant à être étendus et systématisés, malgré l'existence d'un cadre algorithmique unifié [19]. Ce retard tient principalement au fait que l'absence de données thermodynamiques associées, en conjonction avec une population de modèles 3D déterminés expérimentalement bien plus modeste pour l'ARN que pour les protéines, rend délicate, voire irréaliste à l'heure actuelle, la mise en œuvre d'approches basées sur des potentiels statistiques.

3 Conclusion et discussion

Pour terminer ce petit tour d'horizon (non exhaustif) des problématiques et avancées algorithmiques en bioinformatique des ARN, je voudrais rappeler la constante confrontation interdisciplinaire qui jalonne l'histoire de la bioinformatique des ARN, et conclure sur l'importance des contributions, passées et à venir, de l'algorithmique au sein d'une démarche de modélisation.

Comme ont pu l'illustrer les nombreuses références qui émaillent ce texte, la bioinformatique des ARN a fait l'objet de nombreuses contributions aux interfaces entre l'informatique et plusieurs domaines scientifiques. Des biochimistes expérimentaux y ont collaboré avec des bioinformaticiens pour produire un modèle d'énergie entièrement paramétré, le modèle de Turner, à un niveau de granularité permettant une algorithmique de prédiction efficace. Des physiciens y ont contribué par des concepts permettant une classification topologique des structures déterminée expérimentalement, plus tard transformés en algorithmes efficaces pour l'optimisation dans des espaces de conformations expressifs. Des spécialistes de l'évolution ont étudié la relation structure/séquence chez l'ARN comme un modèle de la théorie de l'évolution neutre, résultant un peu par hasard en les premiers outils et algorithmes pour la conception d'ARN synthétiques (Design d'ARN). La biologie moléculaire y suscite et utilise des outils bioinformatiques pour comprendre de nouveaux mécanismes de régulation impliquant l'ARN, parmi lesquels l'interférence par ARN, à l'origine de nombreuses promesses et objet du prix Nobel de physiologie et médecine attribué à Craig et Mello en 2006.

Toutes ces avancées ont été rendues possibles par de nombreuses contributions, issues de champs très variés de l'informatique. Historiquement, l'utilisation de langages formels a permis la recherche de nouveaux ARN dans les génomes, ou encore la modélisation d'espaces de repliements expressifs. Des concepts issus du traitement automatique du langage ont fourni un cadre probabiliste, alternatif à l'approche thermodynamique, dont s'inspirent encore toute une classe d'approches en vogue pour l'étude du repliement. La théorie algorithmique des graphes a fourni un langage pour une reformulation claire, mathématiquement bien définie, de nombreux problèmes. La combinatoire a rendu possible une analyse en moyenne des structures d'ARN aléatoires et des algorithmes œuvrant sur celles-ci, et a fourni un cadre privilégié pour la conception des algorithmes de programmation dynamique omniprésents en bioinformatique. L'optimisation combi-

natoire, et plus particulièrement la programmation mathématique, est de plus en plus fréquemment utilisée pour contourner les résultats de difficulté algorithmique apparaissant, entre autres, dans la prédiction de structure/appariements, l'étude de la cinétique ou le design d'ARN fonctionnels. Enfin, des techniques de fouille de données, reposant sur des structures d'indexation succinctes issues de l'algorithmique du texte, jouissent d'une popularité croissante pour la détection de nouveaux gènes ARN à travers plusieurs génomes. La recherche en bioinformatique des ARN est donc une recherche profondément pluridisciplinaire, et d'impact démontré sur les pratiques des biologistes, y compris à un niveau fondamental.

En effet, éminemment empiriques, les sciences du vivant reposent de façon cruciale sur la notion de modèle, permettant d'exprimer en des termes testables et réfutables une explication avancée pour un phénomène d'intérêt. Des algorithmes exacts y jouent un rôle essentiel dans la validation/invalidation des modèles comme explication de la réalité. En effet, un modèle n'est qu'une construction logique tant que ses conséquences, les propriétés dérivées sous l'hypothèse de celui-ci, n'ont pas été confrontées à une réalité observable. Si les conséquences du modèle sont compatibles avec l'observation empirique, alors la force d'explication du modèle, et notre foi en celui-ci, s'en trouvent renforcées. En cas de contradiction apparente, le modèle doit être remis en question, raffiné, ou tout simplement abandonné.

C'est ici qu'une algorithmique exacte ou, à défaut, offrant des garanties théoriques, devient indispensable : Quelle conclusion tirer d'une incompatibilité entre la réalité observable et les conséquences, estimées bien approximativement et sans garanties théoriques, d'un modèle ? S'agit-il juste d'un artefact lié à un calcul incorrect ? En l'absence d'algorithmes bien conçus, et calculables en temps raisonnable, les modèles formulés en biologie deviennent donc bien difficiles à valider et réfuter. Malheureusement, certains modèles discrets s'avèrent extrêmement difficiles à analyser, en particulier quand ceux-ci sont associés à des problèmes algorithmiques difficilement résolubles au sens de la théorie de la complexité. Les choix de modélisation devraient donc être guidés, non seulement par une connaissance intime des phénomènes étudiés et des données disponibles, mais aussi par une culture algorithmique suffisamment large pour guider la recherche d'un *modèle compromis*, conjuguant expressivité suffisante et espoir de résolution exacte. Une telle recherche justifie, à elle seule, la poursuite et le renforcement d'un dialogue pluridisciplinaire, d'ores et déjà fructueux, à l'interface entre l'informatique et les sciences du vivant.

4 Remerciements

L'auteur tient à remercier chaleureusement Christine Froidevaux et Alain Denise pour leurs relectures attentives de versions préliminaires de ce document.

Références

[1] T. Akutsu. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Appl. Math.*, 104(1-3) :45–62, 2000.

[2] C. Alkan, E. Karakoc, J. H. Nadeau, S C. Sahinalp, and K. Zhang. RNA-RNA interaction prediction and antisense RNA target search. *J Comput Biol*, 13(2) :267–282, March 2006.

[3] P. P. Amaral, M. B. Clark, D. K. Gascoigne, M. E. Dinger, and J. S. Mattick. lncRNAdb : a reference database for long noncoding RNAs. *Nucleic Acids Res*, 39(Database issue) :D146–D151, January 2011.

[4] J. Atkins, R. Gesteland, and T. Cech. *RNA Worlds : From Life's Origins to Diversity in Gene Regulation*. Cold Spring Harbor Perspectives in Biology. Cold Spring Harbor Laboratory Press, 2011.

[5] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res*, 28(1) :235–242, January 2000.

[6] M. Bon, G. Vernizzi, H. Orland, and A. Zee. Topological classification of RNA structures. *J Mol Biol*, 379(4) :900–911, June 2008.

[7] A. Busch, A. S. Richter, and R. Backofen. IntaRNA : efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, 24(24) :2849–2856, 2008.

[8] R. B. Cary and G. D. Stormo. Graph-theoretic approach to RNA modeling using comparative data. *Proceedings International Conference on Intelligent Systems for Molecular Biology*, 3 :75–80, 1995.

[9] Y. Ding, C. Y. Chan, and C. E. Lawrence. RNA secondary structure prediction by centroids in a boltzmann weighted ensemble. *RNA*, 11(8) :1157–1166, August 2005.

[10] ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146) :799–816, June 2007.

[11] T. W. Fawcett and A. D. Higginson. Heavy use of equations impedes communication among biologists. *Proceedings of the National Academy of Sciences*, 109(29) :11735–11739, 2012.

[12] T. W. Fawcett and A. D. Higginson. Reply to chitnis and smith, fernandes, gibbons, and kane : Communicating theory effectively requires more explanation, not fewer equations. *Proceedings of the National Academy of Sciences*, 109(45) :E3058–E3059, 2012.

[13] E. Freyhult, V. Moulton, and P. Clote. Boltzmann probability of RNA structural neighbors and riboswitch detection. *Bioinformatics*, 23(16) :2054–2062, August 2007.

[14] P. P. Gardner and R. Giegerich. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, 5 :140, September 2004.

[15] F. Gavril. Algorithms for a maximum clique and a maximum independent set of a circle graph. *Networks*, 3(3) :261–273, 1973.

[16] J. Gibbons. Do not throw equations out with the theory bathwater. *Proceedings of the National Academy of Sciences*, 109(45) :E3054, 2012.

[17] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy. Rfam : an RNA family database. *Nucleic Acids Res*, 31(1) :439–441, January 2003.

[18] I. L. Hofacker, M. Fekete, and P. F. Stadler. Secondary structure prediction for aligned RNA sequences. *J Mol Biol*, 319(5) :1059–1066, June 2002.

[19] C. Höner zu Siederdissen, S. H. Bernhart, P. F. Stadler, and I. L. Hofacker. A folding algorithm for extended RNA secondary structures. *Bioinformatics*, 27(13) :i129–i136, July 2011.

[20] T. Kasami. An efficient recognition and syntax-analysis algorithm for context-free languages. Scientific report AFCRL-65-758, Air Force Cambridge Research Lab, Bedford, MA, 1965.

[21] N. B. Leontis and E. Westhof. Geometric nomenclature and classification of RNA base pairs. *RNA*, 7(4) :499–512, April 2001.

[22] V. I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10 :707–10, 1966.

[23] R. B. Lyngsø. Complexity of pseudoknot prediction in simple models. In *Proceedings of ICALP*, pages 919–931, 2004.

[24] R. B. Lyngsø and C. N. Pedersen. RNA pseudoknot prediction in energy-based models. *J Comput Biol*, 7(3-4) :409–427, 2000.

[25] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol*, 288(5) :911–940, May 1999.

[26] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7) :1105–1119, 1990.

[27] U. Mückstein, H. Tafer, J. Hackermüller, S. H. Bernhart, P. F. Stadler, and I. L. Hofacker. Thermodynamics of RNA-RNA binding. *Bioinformatics*, 22(10) :1177–1182, May 2006.

[28] E. P. Nawrocki, D. L. Kolbe, and S. R. Eddy. Infernal 1.0 : inference of RNA alignments. *Bioinformatics*, 25(10) :1335–1337, May 2009.

[29] R. Nussinov, G. Pieczenik, J. Griggs, and D. Kleitman. Algorithms for loop matchings. *SIAM Journal on Applied Mathematics*, 35(1) :68–82, 1978.

[30] M. Parisien and F. Major. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, 452(7183) :51–55, March 2008.

[31] Yann Ponty. Bio-algorithmique des ARN : petite promenade aux interfaces. In Eric Sopena, editor, *1024 - Bulletin de la société informatique de France*, volume 4, pages 23–53. SIF - Institut Henri Poincaré, 11 rue Pierre et Marie Curie, 75231 Paris Cedex 05, Octobre 2014.

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

[33] S. Sheikh, R. Backofen, and Y. Ponty. Impact of the energy model on the complexity of RNA folding with pseudoknots. In J. Krkkinen and J. Stoye, editors, *Combinatorial Pattern Matching*, volume 7354 of *Lecture Notes in Computer Science*, pages 321–333. Springer Berlin Heidelberg, 2012.

[34] T. Smith and M. Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1) :195–197, 1981.

[35] I J Tinoco, O. C. Uhlenbeck, and M. D. Levine. Estimation of secondary structure in ribonucleic acids. *Nature*, 230(5293) :362–367, April 1971.

[36] H. Touzet and O. Perriquet. CARNAC : folding families of related RNAs. *Nucleic Acids Res*, 32(Web Server issue) :W142–W145, July 2004.

[37] D. H. Turner and D. H. Mathews. NNDB : the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res*, 38(Database issue) :D280–D282, January 2010.

[38] J. Waldisphl and Y. Ponty. An unbiased adaptive sampling algorithm for the exploration of RNA mutational landscapes under evolutionary pressure. *J Comput Biol*, 18(11) :1465–1479, November 2011.

[39] S. Wuchty, W. Fontana, I. L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2) :145–165, February 1999.

[40] M. Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 244(4900) :48–52, 1989.

[41] M. Zuker and A. B. Jacobson. Using reliability information to annotate RNA secondary structures. *RNA*, 4(6) :669–679, 1998.

[42] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bull Math Biol*, 46(4) :591–621, 1984.

[43] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9(1) :133–148, January 1981.

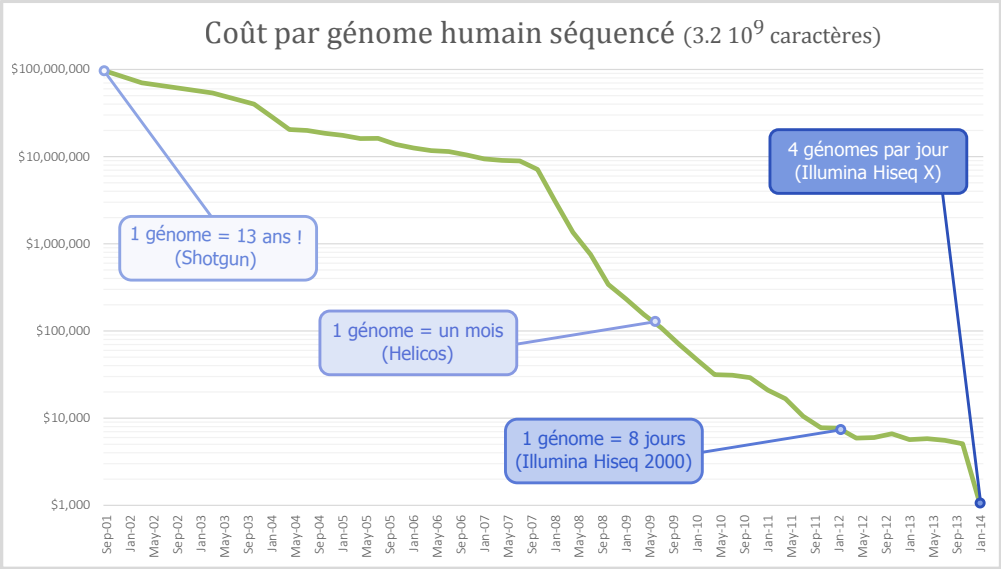


FIGURE 6: Évolution (échelle logarithmique) du coût associé au (re)séquençage d'un génome humain entier (Source : NHGRI [68]).

A Séquençage haut-débit et algorithmique du texte

Des algorithmes efficaces sont aussi indispensables afin de faire face au déluge de données produites, à un débit sans précédent, par des techniques expérimentales en constante évolution. Les pratiques quotidiennes de pans entiers de la recherche en biologie ont ainsi été révolutionnés par l'arrivée, la pérennisation puis la démocratisation des techniques de séquençage haut-débit.

Au début des années 2000, le séquençage shotgun, initié par Sanger et perfectionné dans le cadre du séquençage du génome humain, permettait de produire des fragments d'un ADN à un débit relativement faible et pour un coût important. La dernière décennie a vu se succéder des techniques de séquençage haut-débit de plus en plus rapides et complexes pour des coûts de plus en plus faibles. En particulier, comme illustré par la Figure 6, le constructeur Illumina aurait récemment franchi une barre symbolique en annonçant une machine capable de séquencer un génome individuel (couverture moyenne 30×) en une journée pour un coût total inférieur à 1 000 $, prix incluant les consommables et le travail des techniciens [4].

Un tel coût, dérisoire au regard du niveau de vie des pays développés, permet désormais d'envisager un recours généralisé au séquençage, et ouvre la porte à des applications au potentiel révolutionnaire, telle une médecine personnalisée selon le contenu précis de notre génome ou de notre flore intestinale. Cette situation soulève aussi des problématiques de stockage des données et de traitement algorithmique des données produites, atteignant par exemple de l'ordre de 300 Go par reséquençage de génome humain

4. Ce chiffre exclut cependant l'investissement initial, qui s'élève tout de même à 10 000 000$!

dans le contexte du diagnostic médical. Elle soulève aussi de très concrètes et pressantes questions d'ordre éthique : Qui disposera des données des patients ? Sous quelle forme ? Comment empêcher des séquençages *pirates* ? Plusieurs thèmes de l'informatique (cryptographie distribuée, confidentialité différentielle...) pourront fournir des réponses techniques partielles, mais ne sauront se substituer à un débat démocratique, ainsi qu'à une volonté politique forte pour en appliquer les conclusions.

Reséquençage de génomes connus. Quelque révolutionnaires que soient les nouvelles technologies de séquençage, elles ne produisent jamais que des petits fragments de notre ADN. Deux approches différentes peuvent être utilisées pour les transformer en séquences génomiques, en fonction de la disponibilité ou non d'un génome de référence, appartenant à un individu suffisamment proche, au sens de l'évolution, de l'individu séquencé.

Si un génome de référence est disponible, alors il est possible d'effectuer un reséquençage, c'est-à-dire une projection des fragments sur celui-ci, par exemple afin d'identifier ou reconnaître des mutations induisant un phénotype. Cette technique est aussi largement utilisée dans d'autre types d'expériences, parmi lesquelles les expériences RNA-Seq, où l'on cherche à quantifier la production d'ARN messagers à partir du nombre de fragments chevauchant les régions génomiques leur étant associées. Une telle entreprise soulève typiquement d'importants problèmes de renormalisation, certaines régions génomiques étant plus *couvertes* que d'autres par le séquençage, pour des raisons intrinsèques ou relevant d'un biais expérimental. En conséquence, on utilise souvent ces expériences de façon *différentielle*, en comparant les profils de couverture obtenus dans différences conditions, afin par exemple d'identifier l'effet d'un changement environnemental sur l'expression des gènes, ou identifier des comportements pathologiques. Afin de faire face à l'explosion des données produites, des structures de données succinctes et efficaces (Filtres de Bloom [48]), ainsi que des techniques d'encodage et de compression dédiées (Transformée de Burrows/Wheeler [62]) sont indispensables.

Assemblage de novo. En l'absence de génome de référence, on souhaite alors effectuer un assemblage de novo, c'est-à-dire reconstruire une grande séquence d'ADN compatible avec un ensemble de fragments. L'approche dominante passe par la reconstruction d'un grand graphe dirigé, appelé graphe de De Bruijn, dont la construction est illustrée par la Figure 7. La reconstruction de la séquence d'ADN initiale s'apparente alors à la recherche d'un chemin eulérien, un chemin passant par chacune des arêtes du graphe. Un tel chemin n'est pas forcément unique, par exemple si des motifs répétés apparaissent dans la séquence initiale, telle la composante **C** de la Figure 7. Ici, les données disponibles ne permettent pas de distinguer la séquence initiale (correspondant au chemin $\mathbf{A.B.C.C.D_2.B.C.D_1.C}$) parmi l'ensemble des séquences alternatives (dont le chemin $\mathbf{A.B.C.D_1.C.C.D_2.B.C}$). Certaines portions peuvent aussi n'être recouvertes par aucun fragment. Dans de telles situations, des expériences complémentaires visant la production de plus grands fragments, ou encore l'utilisation d'un modèle de maximisation de vraisemblance, permettront de départager les séquences candidates.

Au-delà de la génomique. Le champ d'application de ces techniques ne se limite plus désormais au séquençage des génomes et à l'analyse des niveaux de transcription. Elles pénètrent aussi la biologie moléculaire et structurale, où elles apparaissent comme une

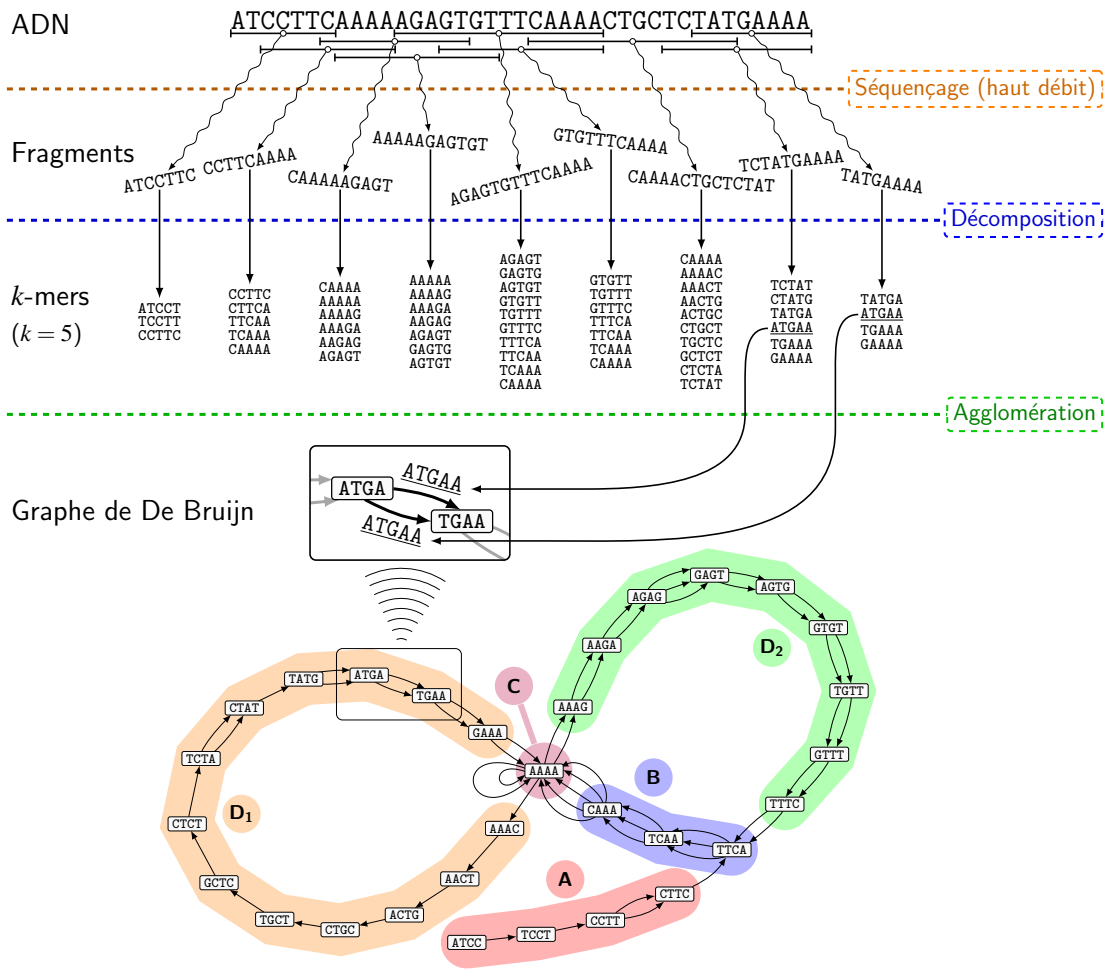


FIGURE 7: Graphe de De Bruijn pour l'assemblage. Les fragments obtenus par séquençage de l'ADN sont décomposés en k-mers, des petits fragments de taille k homogènes. On reconstruit alors le graphe de De Bruijn, dont les nœuds sont les $(k-1)$-mers, et les arêtes relient deux $(k-1)$-mers apparaissant dans un k-mer séquencé. La reconstruction de la séquence initiale s'apparente alors à la recherche d'un chemin eulérien, passant par toutes les arêtes du graphe.

étape-clé au sein de méthodes permettant, à haut-débit :

— une caractérisation de la structure de la chromatine, l'architecture 3D de l'ADN, à travers une détermination de ses contacts inter et intra-chromosomiques (méthodes 5C, Hi-C), qui permet de comprendre son influence sur les mécanismes cellulaires, ou encore de déterminer l'architecture 3D des génomes fortement structurés tel celui de la levure [49] ;

— une détermination des interactions ARN-Protéines (Clip-SEQ) [64], permettant de mieux comprendre l'articulation de ces deux acteurs du monde cellulaire, et leur insertion dans les réseaux de régulation ;

— une estimation des profils d'accessibilité dans les ARN structurés (`FragSeq` [65] ou `SHAPE-Seq` [45]), aidant à la prédiction de structure.

Ces données ne fournissent souvent qu'une information incomplète, assimilable à une *projection* du phénomène d'intérêt. Elles sont donc souvent incorporées dans des algorithmes d'optimisation multiobjectif, et la conception d'algorithmes de prédiction et d'annotation tirant efficacement partie de ces données est aujourd'hui un des challenges algorithmiques majeurs en bioinformatique. En particulier, les nombreux biais systématiques apparaissant dans les données haut-débit soulèvent des problèmes, particulièrement ardus, d'inférence de paramètres cachés couplés à une maximisation de la vraisemblance, souvent résolus en pratique par des heuristiques propres aux méthodes Bayésiennes.

B Petit détour par la combinatoire

Remontons maintenant en 1978, à l'origine des travaux théoriques sur la structure secondaire de l'ARN, et effectuons comme Waterman et Smith [67] un petit exercice de combinatoire énumérative/analytique afin d'énumérer les structures secondaires d'ARN de taille n. On rappelle que celles-ci sont engendrées par la grammaire hors-contexte

$$S \to (\, S^{\overline{\varepsilon}} \,)\, S \mid \bullet\, S \mid \varepsilon \qquad\qquad S^{\overline{\varepsilon}} \to (\, S^{\overline{\varepsilon}} \,)\, S \mid \bullet\, S \;;$$

Considérons alors les séries génératrices

$$S(z) = \sum_{n\geq 0} s_n z^n \qquad \text{et} \qquad S^{\overline{\varepsilon}}(z) = \sum_{n\geq 0} s^{\overline{\varepsilon}} z^n,$$

où s_n (resp. $s_n^{\overline{\varepsilon}}$) est le nombre de structures secondaires (resp. non vide), et z est une variable formelle. On peut alors remarquer que la grammaire ci-dessus est non-ambiguë, et qu'elle peut donc être traduite *automatiquement* en le système d'équations fonctionnelles

$$S(z) \to z\, S^{\overline{\varepsilon}}(z)\, z\, S(z) + z\, S(z) + 1 \qquad \Rightarrow \qquad S^{\overline{\varepsilon}}(z) = z\, S^{\overline{\varepsilon}}(z)\, z\, S(z) + z\, S(z) \;;$$

En résolvant ce système, on trouve deux solutions conjuguées, dont une seulement admet un développement à coefficient positifs :

$$S(z) = \frac{1 - z + z^2 - \sqrt{1 - 2z - z^2 + z^4}}{2z^2}. \tag{1}$$

Cette série génératrice est, par construction, algébrique et admet une singularité dominante, unique et de type *racine*, en $z = \rho := \frac{3-\sqrt{5}}{2}$. Une analyse de singularité *automatique* (on trouvera plus de détails dans la *bible* de Flajolet/Sedgewick [50]) nous donne alors un équivalent asymptotique précis pour le nombre de structures secondaires de taille n

$$s_n := [z^n]\, S(z) = \kappa \cdot \frac{(1/\rho)^n}{n\sqrt{n}}(1 + \mathcal{O}(1/n)), \text{ où } 1/\rho \approx 2.618\ldots$$

où κ est une constante.

On peut aussi, de façon similaire, s'intéresser au nombre moyen de structures compatibles avec une séquence. On remarque alors que l'ensemble des couples structure/séquence compatibles, c'est-à-dire telles que les positions appariées dans la structure correspondent à des paires de nucléotides canoniques, est engendré par la grammaire

$$T \to \begin{matrix} (_{\mathsf{A}}\, T^{\overline{\varepsilon}}\,)_{\mathsf{U}}\, T \\ (_{\mathsf{U}}\, T^{\overline{\varepsilon}}\,)_{\mathsf{A}}\, T \\ (_{\mathsf{G}}\, T^{\overline{\varepsilon}}\,)_{\mathsf{C}}\, T \\ (_{\mathsf{C}}\, T^{\overline{\varepsilon}}\,)_{\mathsf{G}}\, T \\ (_{\mathsf{U}}\, T^{\overline{\varepsilon}}\,)_{\mathsf{G}}\, T \\ (_{\mathsf{G}}\, T^{\overline{\varepsilon}}\,)_{\mathsf{U}}\, T \end{matrix} \mid \begin{matrix} \bullet_{\mathsf{A}}\, T \\ \bullet_{\mathsf{U}}\, T \\ \bullet_{\mathsf{G}}\, T \\ \bullet_{\mathsf{C}}\, T \end{matrix} \mid \varepsilon \qquad\qquad S^{\overline{\varepsilon}} \to \begin{matrix} (_{\mathsf{A}}\, T^{\overline{\varepsilon}}\,)_{\mathsf{U}}\, T \\ (_{\mathsf{U}}\, T^{\overline{\varepsilon}}\,)_{\mathsf{A}}\, T \\ (_{\mathsf{G}}\, T^{\overline{\varepsilon}}\,)_{\mathsf{C}}\, T \\ (_{\mathsf{C}}\, T^{\overline{\varepsilon}}\,)_{\mathsf{G}}\, T \\ (_{\mathsf{U}}\, T^{\overline{\varepsilon}}\,)_{\mathsf{G}}\, T \\ (_{\mathsf{G}}\, T^{\overline{\varepsilon}}\,)_{\mathsf{U}}\, T \end{matrix} \mid \begin{matrix} \bullet_{\mathsf{A}}\, T \\ \bullet_{\mathsf{U}}\, T \\ \bullet_{\mathsf{G}}\, T \\ \bullet_{\mathsf{C}}\, T \end{matrix}.$$

On la transforme alors en système d'équation, qu'on résout pour obtenir la série génératrice

$$T(z) := \sum_{n\geq 0} t_n z^n = \frac{1 - 4\,z + 6\,z^2 - \sqrt{1 - 8\,z + 4\,z^2 - 48\,z^3 + 36\,z^4}}{12\,z^2},$$

où les t_n comptent désormais les couples structures/séquences compatibles de taille n, asymptotiquement équivalents à

$$t_n := [z^n]\, T(z) = \kappa' \cdot \frac{(1/\rho')^{-n}}{n\sqrt{n}}(1 + \mathcal{O}(1/n)), \text{ où } 1/\rho' \approx 8.164\ldots$$

où κ' est une constante. On en conclut donc qu'un ARN de taille n, tiré uniformément parmi les 4^n séquences possibles sur $\{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{U}\}$, est en moyenne compatible avec $\Theta(2.04^n/n\sqrt{n})$ structures secondaires. Cette croissance exponentielle chute à $\Theta(1.87^n/n\sqrt{n})$ en imposant une distance minimale, souvent utilisée dans la littérature, de $\theta = 3$ entre les nucléotides appariés [42].

C C'était demain... résultats récents et problèmes ouverts

Outre les questions algorithmiques laissées ouvertes sur la prédiction du repliement par les approches *ab initio* et comparatives, plusieurs questions bien définies animent actuellement la communauté, et seront très probablement l'objet de nombreuses contributions futures à l'interface informatique/sciences du vivant.

Approches génériques. On l'imagine aisément à la lecture des équations de programmation dynamique de la littérature (s'étalant parfois sur plusieurs pages), l'implémentation des algorithmes évoqués ici s'avère parfois concrètement problématique. Leur conception elle-même est complexe, avec des problèmes de bornes sur les sommes, bornes prudemment omises dans ce document. Comme l'écrivait au début des années 2000 un relecteur anonyme, fréquemment cité par Robert Giegerich : « *Le développement d'une récurrence de programmation dynamique réussie est une question d'expérience, de talent et de chance* ! »[5] Plusieurs formalismes déclaratifs ont ainsi été proposés pour abstraire un schéma de programmation dynamique de son équation, de façon à limiter au maximum les erreurs liées à son implémentation.

Au cours des années 1990, Fabrice Lefebvre propose d'utiliser des grammaires attribuées comme formalisme unificateur [57]. Ces travaux sont ensuite poursuivis et étendus dans les années 2000 par le groupe de Robert Giegerich, qui introduit le formalisme de la programmation dynamique algébrique [53, 63, 54], couplé avec des compilateurs efficaces vers des langages *bas-niveau*. L'utilisation, même par des informaticiens, de tels formalismes pour exprimer leurs algorithmes paraissait appartenir à la science-fiction il y a une petite dizaine d'années. Pourtant, nombreux sont ceux désormais, parmi lesquels des bioinformaticiens de formation, qui utilisent ce type de formalisme pour tester des nouvelles hypothèses/algorithmes à faible coût.

Optimisations de la programmation dynamique. Le problème du repliement de l'ARN est représentatif d'une grande classe de problèmes se prêtant à la programmation dynamique sur des structures arborescentes. À ce titre, ils ont été rencontrés dans d'autres communautés, où des améliorations ont été apportées à leur complexité. Par exemple, le calcul de la fonction de partition et les probabilités de paires de bases à pu être amélioré à $\Theta(n^{2.38})$ [70] par utilisation de la méthode de Valiant [66]. Cette dernière consiste à ramener le calcul de la matrice de programmation dynamique au produit d'une famille de matrices, permettant alors de profiter des dernières optimisations dans le domaine. L'astuce des *quatre-russes* [44] permet aussi de gagner un facteur logarithmique sur la complexité du repliement [51], mais les constantes importantes apparaissant dans son implémentation limitent sa portée en pratique.

De façon plus spectaculaire, plusieurs auteurs ont utilisé un principe de calcul clairsemé (*sparsification*) appliqué à la programmation dynamique prédictive pour les ARN. L'idée est de calculer, lors des étapes initiales de la récurrence, non seulement un score optimal mais aussi une liste de candidats (indices des boucles) non-trivialement dominés, et méritant une exploration lors des étapes ultérieures. Le calcul est alors toujours exact

5. *The development of successful dynamic programming recurrences is a matter of experience, talent and luck.* Anonymous Reviewer.

mais un gain en temps n'est pas toujours garanti. Sous certaines hypothèses (distribution des paires de bases en loi de puissance), la cardinalité de ces listes est bornée par une constante. Il en résulte alors un gain algorithmique linéaire, et le repliement est alors résoluble en temps $\Theta(n^2)$ [46]. Une caractérisation des classes de séquences garantissant un tel gain reste cependant ouverte.

Design d'ARN. Le design d'ARN consiste à concevoir une séquence se repliant, au sens de la minimisation d'énergie, en une structure secondaire donnée en entrée. Il est en ce sens le problème inverse du repliement, et plusieurs travaux y font référence sous le nom de *repliement inverse* de l'ARN. De façon assez surprenante, et malgré deux décennies de recherches sur le sujet [56], la complexité théorique du problème n'est toujours pas établie, une situation totalement exceptionnelle en bioinformatique.

Mimant en un certain sens l'évolution, plusieurs auteurs ont proposé des approches basées sur la recherche locale [47], l'algorithmique génétique [59] ou encore une résolution exponentielle exacte basée sur de la programmation par contraintes [52]. Cependant, le comportement de ce type d'approches est très peu prévisible, et celles-ci s'avèrent difficilement adaptables (en temps raisonnable) à des ensembles de contraintes plus expressives, nécessaires dans les contextes biologiques. Une étude plus approfondie de ce problème est donc à la fois motivée par son intrigante (absence de) structure exploitable algorithmiquement, et par des applications potentielles fascinantes en biologie synthétique [61].

Considérations cinétiques. En étudiant l'ARN par une approche probabiliste de type *fonction de partition* ou, *a fortiori*, en prédisant le repliement par minimisation d'énergie, on travaille implicitement sous l'hypothèse que le processus du repliement a atteint un régime stationnaire, l'équilibre de Boltzmann. Or, le temps de mélange de la chaîne de Markov associée est loin d'être indépendant des propriétés de la séquence d'ARN considérée et peut, dans certains cas, dépasser la durée de vie de l'ARN dans la cellule. On peut donc rencontrer des pièges cinétiques, des minimas locaux de l'énergie libre, dont un ARN aura bien du mal à sortir avant sa dégradation finale par les enzymes. L'ARN peut alors n'être jamais observable dans un état potentiellement extrêmement stable, et l'hypothèse d'un équilibre de Boltzmann donne une vision faussée de sa fonction.

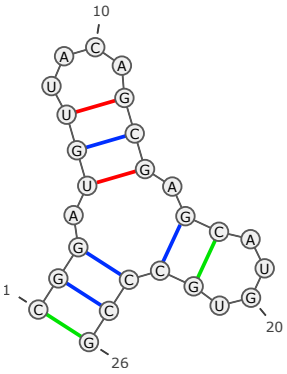
Malheureusement, on rencontre rapidement plusieurs difficultés quand on cherche à étudier les propriétés cinétiques fines d'un ARN. D'une part, il existe un nombre exponentiel de pièges cinétiques potentiels parmi les conformations compatibles avec une séquence [58]. De plus, le calcul de la barrière d'énergie, c'est-à-dire la quantité minimale d'énergie devant être fournie pour passer d'une conformation à une autre, est NP-complet même dans un modèle de Nussinov, et pour des conformations sans pseudo-nœuds [60].

Les approches existantes reposent donc sur une simulation [69], ou sur des heuristiques exécutées sur un sous-ensemble de structures sous-optimales stables [55]. Cependant cette situation, où l'on recherche les propriétés dynamiques de processus stochastiques associés à de très grands espaces d'états, semble faire écho à des problématiques rencontrées en vérification. Il est donc permis d'espérer des contributions issues de ce domaine dans un avenir proche.

D Modèles d'énergie

Plusieurs modèles d'énergie libre ont été considérés pour la prédiction de structure par minimisation d'énergie. Ceux-ci diffèrent principalement par le choix des motifs structurels correspondant à des contributions énergétiques élémentaires. L'énergie libre d'une structure est alors, en général, approchée par une simple somme des contributions de ses motifs.

En illustration, voici les différentes énergies libres associées à une structure secondaire dans plusieurs modèles d'énergie issus de la littérature :



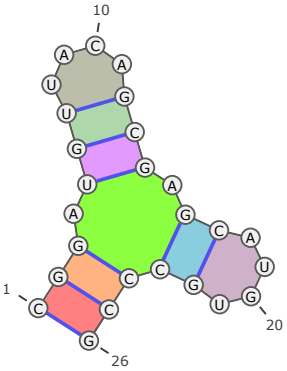
— #Paires de bases

$$E(S) = -8 \text{ kcal.mol}^{-1}$$

— #Liaisons hydrogènes

$$E(S) = 4\,\Delta_G\left(\text{GC}\right) + 2\,\Delta_G\left(\text{CG}\right) + 2\,\Delta_G\left(\text{UG}\right)$$
$$= -(4 \times 3 + 2 \times 2 + 2 \times 1) = -18 \text{ kcal.mol}^{-1}$$

— Empilements (Turner 2004)



$$E(S) = \Delta_G\left(\text{CG/GC}\right) + \Delta_G\left(\text{GG/CC}\right) + \Delta_G\left(\text{UG/GC}\right)$$
$$+ \Delta_G\left(\text{UG/GC}\right) + \Delta_G\left(\text{UG/GC}\right)$$
$$= -(2.4 + 3.3 + 1.4 + 2.5 + 3.4) = -13 \text{ kcal.mol}^{-1}$$

— Plus proche voisin (Turner 2004)

$$E(S) = \Delta_G\left(\text{CG/GC}\right) + \Delta_G\left(\text{GG/CC}\right) + \Delta_G\left(\text{AG/UC}\right) + \Delta_G\left(\text{UG/GC}\right)$$
$$+ \Delta_G\left(\text{UG/GC}\right) + \Delta_G\left(\text{GA}\right) + \Delta_G\left(\text{UG/GC}\right) + \Delta_G\left(\text{CA/UG}\right)$$
$$= -2.4 - 3.3 + 3.6 - 1.4 - 2.5 + 5.4 - 3.4 + 4.1$$
$$= 0.1 \text{ kcal.mol}^{-1}$$

Remarquons que la variabilité des énergies estimées par ces différents modèles n'est pas nécessairement problématique, l'énergie libre étant définie de façon relative. En revanche, l'énergie positive de la structure dans le modèle *plus proche voisin* indique que la structure *ouverte*, sans aucun appariement et ayant une énergie nulle par convention, lui sera préférée par un algorithme de prédiction.

Références supplémentaires

[44] V. L. Arlazarov, E. A. Dinic, M. A. Kronrod, and I. A. Faradev, *On economical construction of the transitive closure of a directed graph*, Soviet Mathematics—Doklady **11** (1970), no. 5, 1209–1210.

[45] S. Aviran, C. Trapnell, J. B. Lucks, S. A. Mortimer, S. Luo, G. P. Schroth, J. A. Doudna, A. P. Arkin, and L. Pachter, *Modeling and automation of sequencing-based characterization of RNA structure.*, Proc Natl Acad Sci U S A **108** (2011), no. 27, 11069–11074 (eng).

[46] R. Backofen, D. Tsur, S. Zakov, and M. Ziv-Ukelson, *Sparse RNA folding : Time and space efficient algorithms*, Journal of Discrete Algorithms **9** (2010), no. 1, 12–31.

[47] A. Busch and R. Backofen, *INFO-RNA–a fast approach to inverse RNA folding.*, Bioinformatics **22** (2006), no. 15, 1823–1831 (eng).

[48] R. Chikhi and G. Rizk, *Space-efficient and exact de bruijn graph representation based on a bloom filter*, Algorithms for Molecular Biology **8** (2013), no. 1, 22.

[49] Z. Duan, M. Andronescu, K. Schutz, C. Lee, J. Shendure, S. Fields, W. S. Noble, and C., *A genome-wide 3C-method for characterizing the three-dimensional architectures of genomes.*, Methods **58** (2012), no. 3, 277–288 (eng).

[50] P. Flajolet and R. Sedgewick, *Analytic combinatorics*, Cambridge University Press, 2009.

[51] Y. Frid and D. Gusfield, *A simple, practical and complete O(n3/log n)-time algorithm for RNA folding using the four-russians speedup.*, Algorithms Mol Biol **5** (2010), 13 (eng).

[52] J. A. Garcia-Martin, P. Cote, and I. Dotu, *RNAiFOLD : A constraint programming algorithm for RNA inverse folding and molecular design*, J Bioinform Comput Biol **11** (2013), no. 02, 1350001, PMID : 23600819.

[53] R. Giegerich and C. Meyer, *Algebraic dynamic programming*, AMAST, 2002, pp. 349–364.

[54] R. Giegerich and H. Touzet, *Modeling dynamic programming problems over sequences and trees with inverse coupled rewrite systems*, Algorithms **7** (2014), no. 1, 62–144.

[55] I. L. Hofacker, C. Flamm, C. Heine, M. T. Wolfinger, G. Scheuermann, and P. F. Stadler, *Barmap : RNA folding on dynamic energy landscapes.*, RNA **16** (2010), no. 7, 1308–1316 (eng).

[56] I. L. Hofacker, W. Fontana, P. Stadler, L. Bonhoeffer, M. Tacker, and P. Schuster, *Fast folding and comparison of RNA secondary structures*, Monatshefte für Chemie / Chemical Monthly **125** (1994), no. 2, 167–188 (English).

[57] F. Lefebvre, *An optimized parsing algorithm well suited to RNA folding*, ISMB, 1995, pp. 222–230.

[58] W. A. Lorenz and P. Clote, *Computing the partition function for kinetically trapped RNA secondary structures.*, PLoS One **6** (2011), no. 1, e16178 (eng).

[59] R. B. Lyngsø, J. W J. Anderson, E. Sizikova, A. Badugu, T. Hyland, and J. Hein, *Frnakenstein : multiple target inverse RNA folding.*, BMC Bioinformatics **13** (2012), 260 (eng).

[60] J. Manuch, C. Thachuk, L. Stacho, and A. Condon, *NP-completeness of the energy barrier problem without pseudoknots and temporary arcs*, Natural Computing **10** (2011), no. 1, 391–405.

[61] G. Rodrigo, T. E. Landrain, and A. Jaramillo, *De novo automated design of small RNA circuits for engineering synthetic riboregulation in living cells.*, Proc Natl Acad Sci U S A **109** (2012), no. 38, 15271–15276 (eng).

[62] M. Salson, T. Lecroq, M. Léonard, and L. Mouchard, *A four-stage algorithm for updating a burrows-wheeler transform*, Theoretical Computer Science **410** (2009), no. 43, 4350–4359.

[63] G. Sauthoff, S. Janssen, and R. Giegerich, *Bellman's GAP : a declarative language for dynamic programming*, PPDP, 2011, pp. 29–40.

[64] J. Ule, K. Jensen, A. Mele, and R. B. Darnell, *CLIP : a method for identifying protein-RNA interaction sites in living cells.*, Methods **37** (2005), no. 4, 376–386 (eng).

[65] J. G. Underwood, A. V. Uzilov, S. Katzman, C. S. Onodera, J. E. Mainzer, D. H. Mathews, T. M. Lowe, S. R. Salama, and D. Haussler, *FragSeq : transcriptome-wideRNA structure probing using high-throughput sequencing.*, Nat Methods **7** (2010), no. 12, 995–1001 (eng).

[66] L. G. Valiant, *General context-free recognition in less than cubic time*, J. Comput. Syst. Sci. **10** (1975), no. 2, 308–314.

[67] M. S. Waterman and T. F. Smith, *Secondary structure of single stranded nucleic acids*, Math Biosci **42** (1978), 257–266.

[68] K A Wetterstrand, *DNA sequencing cost : Data from the NHGRI genome sequencing program (GSP)*, February 2014.

[69] A. Xayaphoummine, T. Bucher, F. Thalmann, and H. Isambert, *Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations.*, Proc Natl Acad Sci U S A **100** (2003), no. 26, 15310–15315 (eng).

[70] S. Zakov, D. Tsur, and M. Ziv-Ukelson, *Reducing the worst case running times of a family of RNA and CFG problems, using valiant's approach.*, Algorithms Mol Biol **6** (2011), no. 1, 20 (eng).

[71] M. Zuker and D. Sankoff, *RNA secondary structures and their prediction*, Bull Math Biol **46** (1984), no. 4, 591–621.