Université de Montréal

Quantification de la relation séquence-activité de l'ARN par prédiction de structure tridimensionnelle

par
Karine St-Onge

Département d'Informatique et de Recherche Opérationnelle
Faculté des Arts et des Sciences

Thèse présentée à la Faculté des Arts et des Sciences
en vue de l'obtention du grade de Philosophiæ Doctor (Ph. D.)
en informatique

Août, 2011

Université de Montréal
Faculté des Arts et des Sciences

Cette thèse intitulée :

Quantification de la relation séquence-activité de l'ARN par prédiction de structure tridimensionnelle

Présentée par :
Karine St-Onge

a été évaluée par un jury composé des personnes suivantes :

Nadia El-Mabrouk, président-rapporteur
François Major, directeur de recherche
Sylvie Hamel, co-directeur
Nicolas Moitessier, membre du jury
Alain Laederach, examinateur externe
Richard Giasson, représentant du doyen de la FAS

Résumé

Dans un premier temps, nous avons modélisé la structure d'une famille d'ARN avec une grammaire de graphes afin d'identifier les séquences qui en font partie. Plusieurs autres méthodes de modélisation ont été développées, telles que des grammaires stochastiques hors-contexte, des modèles de covariance, des profils de structures secondaires et des réseaux de contraintes. Ces méthodes de modélisation se basent sur la structure secondaire classique comparativement à nos grammaires de graphes qui se basent sur les motifs cycliques de nucléotides. Pour exemplifier notre modèle, nous avons utilisé la boucle E du ribosome qui contient le motif Sarcin-Ricin qui a été largement étudié depuis sa découverte par cristallographie aux rayons X au début des années 90.

Nous avons construit une grammaire de graphes pour la structure du motif Sarcin-Ricin et avons dérivé toutes les séquences qui peuvent s'y replier. La pertinence biologique de ces séquences a été confirmée par une comparaison des séquences d'un alignement de plus de 800 séquences ribosomiques bactériennes. Cette comparaison a soulevée des alignements alternatifs pour quelques unes des séquences que nous avons supportés par des prédictions de structures secondaires et tertiaires. Les motifs cycliques de nucléotides ont été observés par les membres de notre laboratoire dans l'ARN dont la structure tertiaire a été résolue expérimentalement. Une étude des séquences et des structures tertiaires de chaque cycle composant la structure du Sarcin-Ricin a révélé que l'espace des séquences dépend grandement des interactions entre tous les nucléotides à proximité dans l'espace tridimensionnel, c'est-à-dire pas uniquement entre deux paires de bases adjacentes. Le nombre de séquences générées par la grammaire de graphes est plus petit que ceux des méthodes basées sur la structure secondaire classique. Cela suggère l'importance du contexte pour la relation entre la séquence et la structure, d'où l'utilisation d'une grammaire de graphes contextuelle plus expressive que les grammaires hors-contexte.

Les grammaires de graphes que nous avons développées ne tiennent compte que de la structure tertiaire et négligent les interactions de groupes chimiques spécifiques avec des éléments extra-moléculaires, comme d'autres macromolécules ou ligands. Dans un deuxième temps et pour tenir compte de ces interactions, nous avons développé un modèle qui tient compte de la position des groupes chimiques à la surface des structures tertiaires. L'hypothèse étant que les groupes chimiques à des positions conservées dans des séquences prédéterminées actives, qui sont déplacés dans des séquences inactives pour une fonction précise, ont de plus grandes chances d'être impliqués dans des interactions avec des facteurs. En poursuivant avec l'exemple de la boucle E, nous avons cherché les groupes de cette boucle qui pourraient être impliqués dans des interactions avec des facteurs d'élongation. Une fois les groupes identifiés, on peut prédire par modélisation tridimensionnelle les séquences qui positionnent correctement ces groupes dans leurs structures tertiaires. Il existe quelques modèles pour adresser ce problème, telles que des descripteurs de molécules, des matrices d'adjacences de nucléotides et ceux basé sur la thermodynamique. Cependant, tous ces modèles utilisent une représentation trop simplifiée de la structure d'ARN, ce qui limite leur applicabilité.

Nous avons appliqué notre modèle sur les structures tertiaires d'un ensemble de variants d'une séquence d'une instance du Sarcin-Ricin d'un ribosome bactérien. L'équipe de Wool à l'université de Chicago a déjà étudié cette instance expérimentalement en testant la viabilité de 12 variants. Ils ont déterminé 4 variants viables et 8 létaux. Nous avons utilisé cet ensemble de 12 séquences pour l'entraînement de notre modèle et nous avons déterminé un ensemble de propriétés essentielles à leur fonction biologique. Pour chaque variant de l'ensemble d'entraînement nous avons construit des modèles de structures tertiaires. Nous avons ensuite mesuré les charges partielles des atomes exposés sur la surface et encodé cette information dans des vecteurs. Nous avons utilisé l'analyse des composantes principales pour transformer les vecteurs en un ensemble de variables non corrélées, qu'on appelle les composantes principales. En utilisant la distance Euclidienne pondérée et l'algorithme du plus proche voisin, nous avons appliqué la technique du « Leave-One-Out Cross-Validation » pour choisir les meilleurs paramètres pour prédire

l'activité d'une nouvelle séquence en la faisant correspondre à ces composantes principales. Finalement, nous avons confirmé le pouvoir prédictif du modèle à l'aide d'un nouvel ensemble de 8 variants dont la viabilité à été vérifiée expérimentalement dans notre laboratoire.

En conclusion, les grammaires de graphes permettent de modéliser la relation entre la séquence et la structure d'un élément structural d'ARN, comme la boucle E contenant le motif Sarcin-Ricin du ribosome. Les applications vont de la correction à l'aide à l'alignement de séquences jusqu'au design de séquences ayant une structure prédéterminée. Nous avons également développé un modèle pour tenir compte des interactions spécifiques liées à une fonction biologique donnée, soit avec des facteurs environnants. Notre modèle est basé sur la conservation de l'exposition des groupes chimiques qui sont impliqués dans ces interactions. Ce modèle nous a permis de prédire l'activité biologique d'un ensemble de variants de la boucle E du ribosome qui se lie à des facteurs d'élongation.

Mots-clés : ARN, Structure, 3D, Séquence, Fonction, Grammaire, QSAR, Sarcin-Ricin

Abstract

Initially, we modeled the structure of an RNA family with a graph grammar to identify sequences that correspond to it. Several other modeling approaches have been developed to derive sequences, such as stochastic context-free grammars, covariance models, secondary structures profiles and constraint networks. These modeling methods are based on secondary structure compared to our graph grammars which are based on the nucleotide cyclic motifs. To exemplify our graph grammar model, we used the loop E of the ribosome that contains the Sarcin-Ricin motif that has been widely studied since its discovery by X-ray crystallography in the early 90s.

We built a graph grammar for the structure of the Sarcin-Ricin motif and derived the sequences that correspond to it. The biological relevance of these sequences is supported by an alignment of 800 bacterial ribosomal sequences. This comparison raised alternative alignments for some of the sequences that we supported by predictions of secondary and tertiary structures. According to a new tertiary structure, those alternative alignments accommodate the new derived sequences.

The nucleotide cyclic motifs used in the grammar were observed by members of our laboratory in RNA tertiary structures that were solved experimentally. We study the sequences and tertiary structures of the nucleotide cyclic motifs of the Sarcin-Ricin motif. This study suggests that the space of sequences depends heavily on interactions between all nucleotides in the nearby three-dimensional space and not only between two adjacent base pairs. We compare the number of sequences generated by the graph grammar with non contextual methods and our graph grammar generates less sequences. This suggests the importance of context for the relationship between sequence and structure, hence the use of a contextual graph grammar is more expressive than context-free grammars.

The graph grammars we used include the tertiary structure but neglect the interactions with extra-molecular factors, such as other macromolecules or ligands. In a second stage and to take into account these interactions, we developed a model incorporating the positioning of chemical groups on the surface of the tertiary structures. The assumption being that the chemical groups that are conserved on the surface of the RNA in active sequences are more likely to be involved in interactions with extra-molecular factors. Continuing with the example of the loop E, we searched the groups that could be involved its interactions with elongation factors. Knowledge of the groups involved in the important interactions serves to predict by three-dimensional modeling new sequences that have potentials to realize these interactions and thus the same function. There are few models that have been developed to address this problem: molecular descriptors, nucleotide adjacency matrices and others based on thermodynamics. These models use an oversimplified representation of the RNA structure, which limits their applicability.

We applied our model to the tertiary structures of a set of variants of a sequence of one instance of the Sarcin-Ricin motif from a bacterial ribosome. Wool and coworkers at the University of Chicago studied this proceeding experimentally by testing the viability of twelve variants. They identified four viable variants and eight lethal. We used this set of twelve sequences for training our model and we identified a set of essential properties to their biological function. For each variant of the training set we built models of tertiary structures. We then measured the partial charges of exposed atoms on the surface and we encoded this information into vectors. We used principal component analysis to transform the vectors into a set of uncorrelated variables, called principal components. Using the weighted Euclidean distance and a nearest neighbor algorithm, we applied the technique of "Leave-One-Out Cross-Validation" to choose the best parameters to predict the activity of a new sequence to match these principal components. Finally, we validated the predictive model using a new set of eight variants whose viability has been verified experimentally in our laboratory.

In conclusion, graph grammars are used to model the relationship between sequence and structure of an RNA structural element, such as the ribosomal loop E containing the Sarcin-Ricin motif. Applications range from the correction of sequence alignment to sequence design with a predetermined structure. We also developed a model to take into account the specific interactions related to a specific biological function. Our model is based on the retention of the exposure of chemical groups that are involved in these interactions. This model has allowed us to predict the biological activity of a set of variants of the loop E that binds to elongation factors.

Keywords: RNA, Structure, Sequence, Function, Grammar, QSAR, Sarcin-Ricin

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et abréviations

A
Adénine 31

ADN
Acide DésoxyriboNucléique 27

ARN
Acide RiboNucléique 27

ARNr
ARN ribosomique 29

ARNt
ARN de transfert 52

BFGS
Broyden–Fletcher–Goldfarb–Shanno xi

C
Cytosine 31

CSP
Constraint Satisfaction Problem 46

ERPIN
Easy RNA Profile IdentificatioN 57

G
Guanine 31

H
Face Hoogsteen 36

HIV
Human Immunodeficiency Virus 65

L-BFGS
Limited memory Broyden–Fletcher–Goldfarb–Shanno xi

LOOCV
Leave-One-Out Cross-Validation xv

MC-Sym

Such a lonely day,
and it's mine,
the most loneliest day of my life.
Such a lonely day,
and it's mine,
it's a day that I'm glad I survived.
Lonely Day - System of a down

Mon amour, vois-tu,
la vie nous attend,
elle s'offre à nous,
et à ses enfants?
Qui tout comme nous,
à chaque Printemps,
cueillerons l'amour,
au hasard du temps.
Pour cet amour – Roger Dumas

Remerciements

Je voudrais remercier M^me^ Volonté qui m'a permis de débuter mon doctorat en 2005 sous la direction de François Major et de Sylvie Hamel. Merci à vous deux de m'avoir dirigé pendant ces 7 années.

Je souhaite aussi remercier tous les gens que j'ai côtoyés et qui m'ont aidé durant ces années : Véronique, Paul, Marc-Frédérick, Maria, Karine, Romain, Emmanuelle, Philippe, Sébastien, Éric, Fabrice, Rym, mais surtout M^me^ Entraide et sans oublier tous ceux de la communauté *Famille*.

Je tiens également à remercier M. Courage qui m'a épaulé et soutenu afin de passer à travers un triste et pénible événement. « Rien n'arrive pour rien dans la vie », c'est ce que M^me^ Espérance m'a toujours dit. Merci Martin d'avoir passé dans ma vie.

Il ne faudrait pas passer sous silence la contribution de M^me^ Patience et M^me^ Persévérance qui ont été indispensable pendant les 3 dernières années et sans qui, je ne serais jamais parvenue à mes fins. Un gros merci mesdames.

Finalement, je voudrais remercier M^me^ Béatitude qui m'a présenté Yanick, mon amoureux. Merci Yanick d'être à mes côtés. Avec toi, la vie prend tout son sens. Je t'aime.

Introduction

Une séquence d'acide ribonucléique (ARN) se replie dans l'espace pour adopter une structure tertiaire qui lui confère sa fonction. Lors du repliement, des interactions chimiques intra et extra moléculaires se forment. De ces interactions, les interactions intra moléculaires stabilisent la structure et les interactions extra moléculaires stabilisent les arrimages avec d'autres molécules, par exemple d'autres ARN, des métabolites, des protéines, de l'acide désoxyribonucléique (ADN), etc. Une fois l'ARN repliée, les structures tertiaires d'ARN sont constituées d'un ensemble de motifs structuraux de bases. Plusieurs de ces motifs ont été prédits par analyse comparative de séquences [1]. Donc, il existe une relation entre la séquence de nucléotides d'un ARN et sa structure. En effet, les structures d'ARN résolues par cristallographie aux rayons X montrent la présence de motifs structuraux [2]. De plus, par loi de thermodynamique, la récurrence des motifs suggère qu'ils se replient dans des conformations stables. Les rôles des motifs structuraux sont multiples : maintenir la structure de l'ARN comme telle, participer à des interactions avec d'autres molécules et induire une fonction catalytique en sont des exemples [3][4]. En conséquence, il est maintenant reconnu que les motifs sont des éléments cruciaux, voir fondateurs, de la structure de l'ARN et de leurs fonctions biologiques [2][5][6][7][8].

Une famille d'ARN est un ensemble de séquences qui partagent la même fonction et qui, conséquemment, se replient dans des structures similaires. En informatique, les familles d'ARN ont été représentées par des grammaires stochastiques hors-contexte [9], des modèles de covariance [10][11][12][13], des profils de structures secondaires [14][15] et des réseaux de contraintes [16][17]. Ces modèles peuvent générer les séquences de la famille qu'ils représentent, aligner et reconnaître leurs séquences et donc chercher de nouvelles séquences d'une famille donnée dans des génomes.

Cependant, les représentations énumérées plus haut se basent uniquement sur la structure secondaire classique, et ne prennent en considération que les interactions de paires

de bases classiques (c'est-à-dire de type Watson-Crick entre CG et AU) alors qu'une structure d'ARN se compose également de paires non classiques, d'interactions tertiaires et d'empilement de bases [18]. De plus, la structure secondaire classique ne tient compte que de tandems de paires de bases et non pas du contexte dans lequel ces paires se trouvent. En particulier, les paires de bases et les nucléotides avoisinants qui influent sur leur formation ne sont pas pris en compte. Pourtant, ce contexte a une influence sur l'espace des séquences d'une paire de bases donnée et cela au-delà de sa géométrie locale [19].

Notre laboratoire a découvert les motifs cycliques de nucléotides qui permettent de capturer l'information du contexte [2]. Ces cycles sont composés de tous les types de paires de bases (classiques et non classiques), d'interactions tertiaires et d'empilements de bases. Dans une situation de prédiction de structure secondaire et tertiaire, les cycles considèrent toutes les interactions adjacentes, préservant ainsi le contexte de chacune [20].

Lorsqu'on doit en plus tenir compte d'une fonction biologique spécifique, il faut considérer les interactions extra moléculaires. Deux molécules interagissent *via* des interactions impliquant certains de leurs groupes chimiques, le plus souvent situés à leurs surfaces. Pour considérer les groupes chimiques qui participent à ces interactions, nous devons considérer la structure tertiaire.

Pour prédire la structure tertiaire, nous utilisons le programme *MC-Sym* [21]. Avec les structures tertiaires prédites d'un ensemble de séquences d'une famille, nous pouvons déterminer les positions conservées des groupes chimiques qui peuvent être impliqués dans des interactions nécessaires à une fonction donnée. Toute structure qui aurait des groupes de même nature positionnées aux bons endroits devrait pouvoir réaliser la même fonction.

Au Chapitre 1, j'introduis la structure des ARN et les techniques que j'ai utilisées pour quantifier la relation entre la séquence et l'activité d'un ARN.

Au Chapitre 2, je présente des méthodes de représentation informatiques pour des familles d'ARN, et au Chapitre 3, les méthodes pour quantifier la relation entre la structure et l'activité.

Au Chapitre 4, j'introduis une nouvelle méthode de représentation pour des familles d'ARN avec des grammaires de graphes. Finalement, au Chapitre 5, je présente un modèle pour déterminer les groupes chimiques nécessaires à la réalisation d'une fonction biologique donnée et comment on peut utiliser ce modèle pour prédire si une structure tertiaire donnée et issue d'une nouvelle séquence peut remplir cette fonction.

À l'aide du modèle de représentation de familles d'ARN avec des grammaires de graphes, nous avons généré des séquences du motif Sarcin-Ricin qui ont été comparées avec celles d'un alignement de l'ARN ribosomique (ARNr) bactérien. À l'aide du modèle tertiaire, nous avons pu prédire la viabilité de bactéries de la plupart des variants testés de la boucle E qui se lie à des facteurs d'élongation. Nous avons testé ces séquences expérimentalement.

Chapitre 1 Contexte et définitions

1.1 Structure d'ARN

1.1.1 Structure tertiaire

L'ARN est un polymère linéaire constitué d'un enchaînement de nucléotides, Adénine (A), Cytosine (C), Guanine (G) et Uracile (U), reliés par des liaisons phosphodiester. La plupart des ARN sont formés d'un simple brin dans la cellule et se replient sur eux-mêmes (voir **Figure 1A**), formant une structure stabilisée par des interactions intramoléculaires. La formation de paires de bases intramoléculaires, entre bases complémentaires (A avec U, G avec C et, parfois, G avec U) est le fondement de cette structure. Ces paires de bases permettent le repliement de l'ARN, ce qui produit des hélices (série de paires de bases) et des boucles (absence de paire de bases). En plus de ces paires de bases classiques, l'ARN peut former des paires de bases non-canoniques et des paires de bases à longue distance, appelées aussi tertiaires.

L'existence de structures tertiaires bien définies dans les ARN est un élément important de la fonction. Ces structures permettent à l'ARN de former des sites de liaison précis (pour des petites molécules ou des protéines) et lui permettent d'assurer des fonctions catalytiques. Par cette relation avec la fonction biologique, l'analyse et la prédiction de la structure tertiaire est un champ de recherche très actif à la fois dans le domaine de la biologie moléculaire et de la bio-informatique.

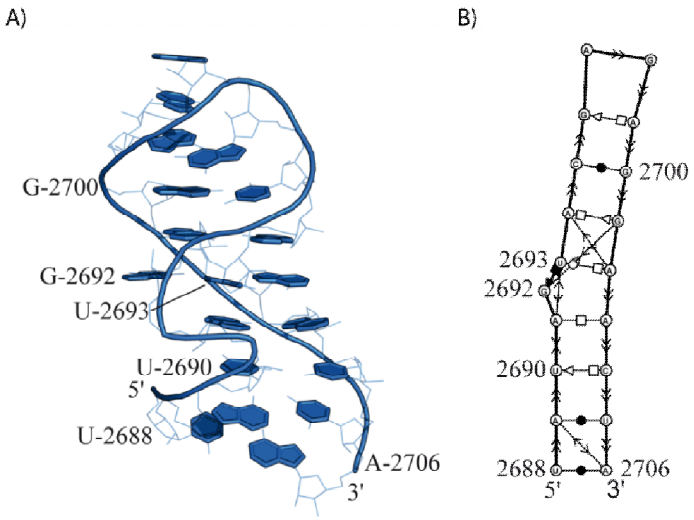


Figure 1. Motif Sarcin-Ricin. Les nucléotides 2688 à 2706 du 23S de l'ARNr de *Haloarcula marismortui*. A) Structure tertiaire. Les nucléotides sont illustrés par des formes planaires. Les liaisons phosphodiesters sont illustrées par un cylindre. B) Graphe d'interactions. Les liaisons phosphodiesters (gras), les paires de bases (●, ●■, ●▶, ■●, ■, ■▶, ◀●, ◀■, ◀, ○, ○□, ○▷, □○, □, □▷, ◁○, ◁□ et ◁), les empilements (>>, <<, >< et <> et la séquence (A, C, G et U) provenant de la structure en (A).

1.1.1.1 Motif

Un motif est un fragment d'une structure d'ARN qui est répété dans une ou plusieurs structures tertiaires et qui est donc conservé à travers l'évolution et souvent relié à une fonction biologique. Par exemple, la **Figure 2** illustre la molécule 23S de l'ARNr de *H. marismortui* contenant sept instances du motif Sarcin-Ricin montrées en bleu.

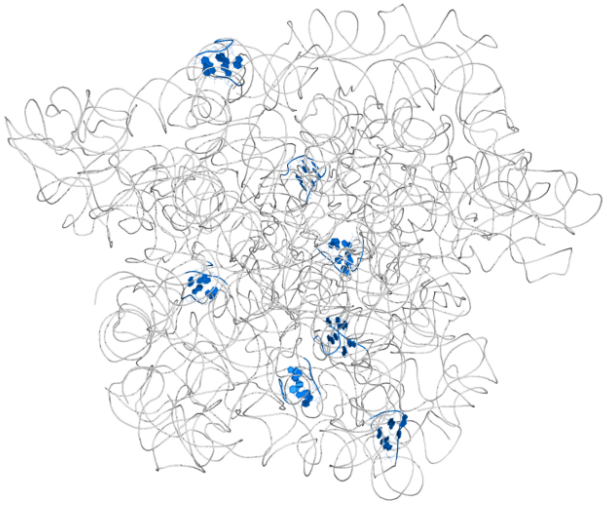


Figure 2. 23S de l'ARNr de *H. marismortui*. Les liaisons phosphodiesters sont représentées par un cylindre. Les nucléotides et les liaisons phosphodiesters des motifs Sarcin-Ricin sont illustrés en bleu. Les autres nucléotides ne sont pas montrés.

1.1.1.2 Famille

Généralement, les ARN fonctionnels ont une meilleure conservation au niveau de la structure que de la séquence. Une famille d'ARN est donc un ensemble de séquences qui partagent la même fonction et qui se replient dans des structures similaires. Rfam « RNA families » [11], est une base de données qui regroupe près de 2000 familles (version 10.1). Ces familles sont représentées, entre autres, par un alignement de séquences et une structure consensus (structure qui représente la majorité d'un ensemble de structures).

1.1.2 Graphe d'interactions

Le graphe d'interactions représente informatiquement la structure tertiaire d'un ARN afin de développer des outils d'annotation automatique dans le but de mieux caractériser, classifier et comparer ces structures entre elles. Nous utilisons l'outil *MC-*

Annotate [22][23], sur la structure tertiaire pour obtenir le graphe d'interactions (voir **Figure 1B**).

Un graphe d'interactions et un graphe $G = \{V, E\}$ où V est l'ensemble des nucléotides (sommets ou nœuds) et E est l'ensemble des interactions observées (liaisons phosphodiesters, paires de bases et empilement) entre deux nucléotides (arêtes ou arcs). À noter que plus d'une interaction par arête peut être présente simultanément entre deux nucléotides. Les types d'arêtes doivent avoir la capacité de refléter ce genre de combinaison (c'est-à-dire liaison phosphodiester-empilement, liaison phosphodiester-paire de bases, etc.). Par exemple, dans la **Figure 1B**, tous les empilements le long des liaisons phosphodiesters (combinaison liaison phosphodiester et empilement) et la paire de bases entre les nucléotides 2692 et 2693 (combinaison de liaison phosphodiester et paire de bases) sont des arêtes représentant plusieurs interactions.

1.1.2.1 Types d'interactions

1.1.2.1.1 Liaison phosphodiester

La liaison phosphodiester correspond au lien entre deux nucléotides. Le phosphate du groupe phosphate relie les nucléotides par les carbones 3' et 5' (voir **Figure 3**). La représentation d'une liaison phosphodiester dans le graphe d'interactions, est un trait (voir **Figure 4**).

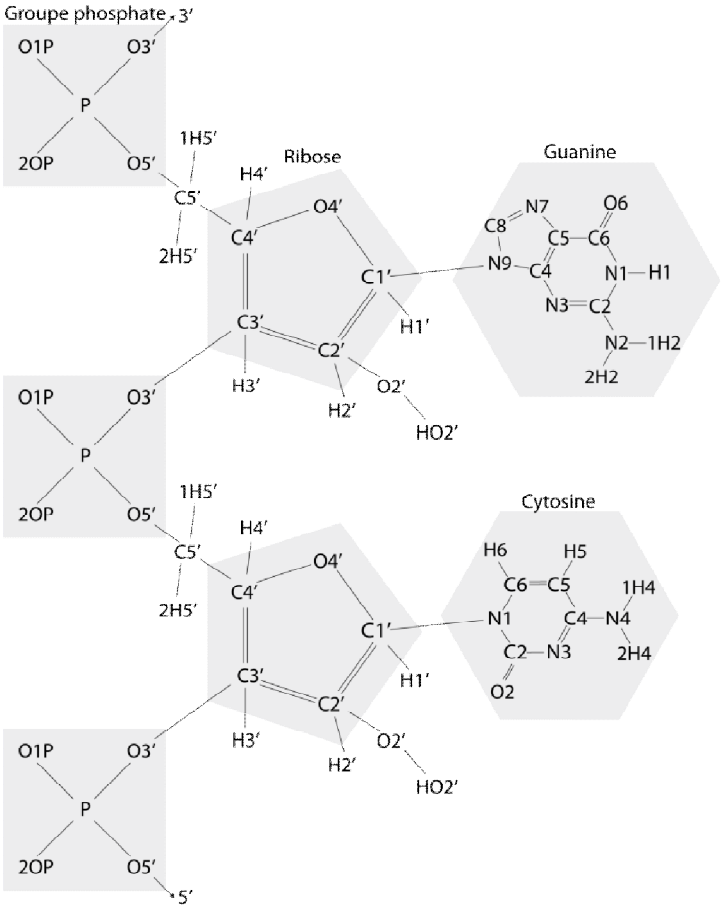


Figure 3. Deux nucléotides. La structure chimique d'un nucléotide C suivi d'un nucléotide G. Le groupe phosphate est représenté par un carré, le ribose par un pentagone et le nucléotide par un hexagone.

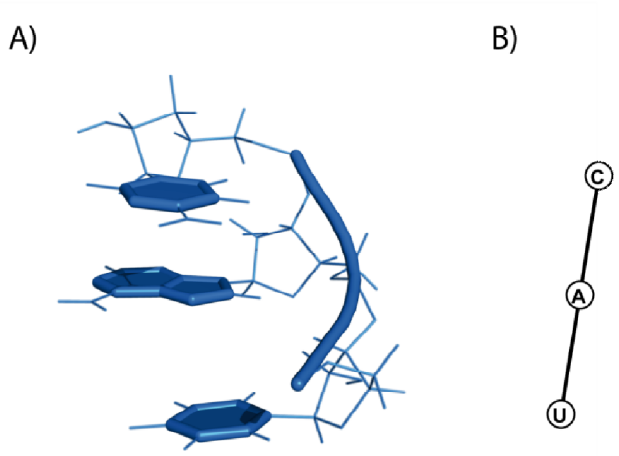


Figure 4. Liaisons phosphodiesters d'un trinucléotide. A) Structure tertiaire. Une liaison phosphodiester dans la structure tertiaire, où cette la trace des phosphates est illustrée par un cylindre. Les nucléotides sont illustrés par des formes planaires. B) Graphe d'interactions. Une liaison phosphodiester dans le graphe d'interactions, où cette liaison est représentée par une arête. La séquence est représentée par A, C, G et U.

1.1.2.1.2 Paire de bases

Nous utilisons la nomenclature de Leontis et Westhof pour décrire les types des paires de bases [24], et pour indiquer les côtés des bases impliqués dans les liaisons hydrogène (voir **Figure 5**). Les noms et symboles suivants ont été définis dans le but de représenter chacun des trois côtés d'une base : le côté Watson-Crick (W), ● (*cis*) ; ❍ (*trans*), le côté Hoogsteen (H), ■ (*cis*) ; □ (*trans*), et le côté Sucre (S), ◀▶ (*cis*) ; ◁▷ (*trans*) [24]. Lorsque deux bases interagissent par le même côté, seulement un symbole est utilisé. Par exemple, X□□Y s'écrit X□Y. L'orientation *cis*/*trans* reflète l'orientation relative de la liaison phosphodiester par rapport à la médiane du plan formé par les deux bases de la paire (voir **Figure 6**).

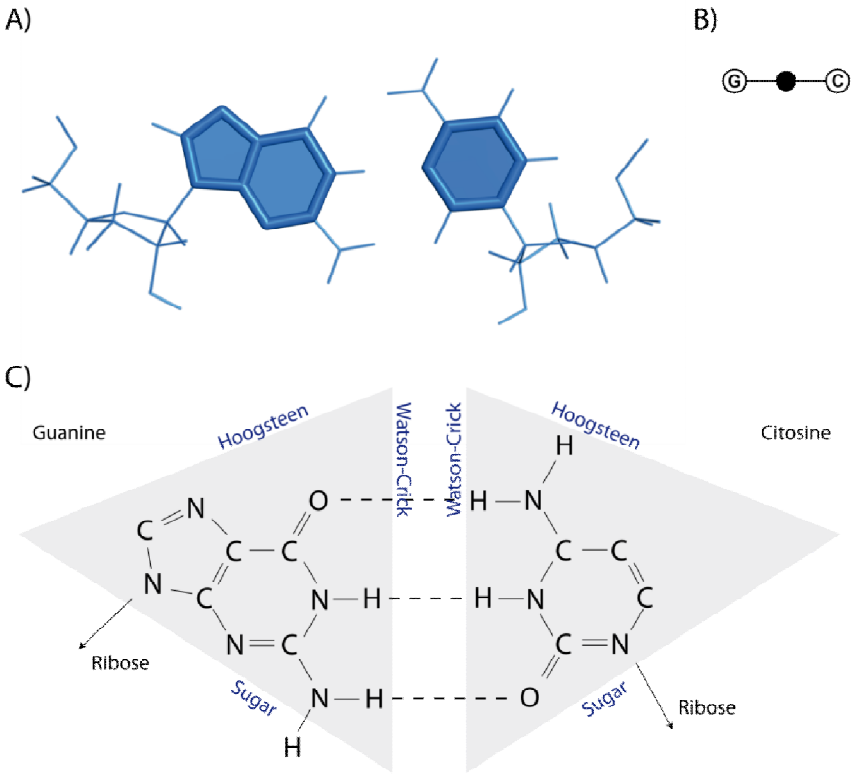


Figure 5. Paire de bases G●C. A) Structure tertiaire. Une paire de bases dans la structure tertiaire. Les nucléotides sont illustrés par des formes planaires. B) Graphe d'interactions. Une paire de bases dans le graphe d'interactions, où la paire de bases est représentée par le symbole ●. La séquence est représentée par A, C, G et U. C) Schéma. Chaque base possède trois côtés : Watson-Crick, (W), Hoogsteen, (H), et Sucre (S). Les liaisons hydrogène sont représentées par des lignes pointillées.

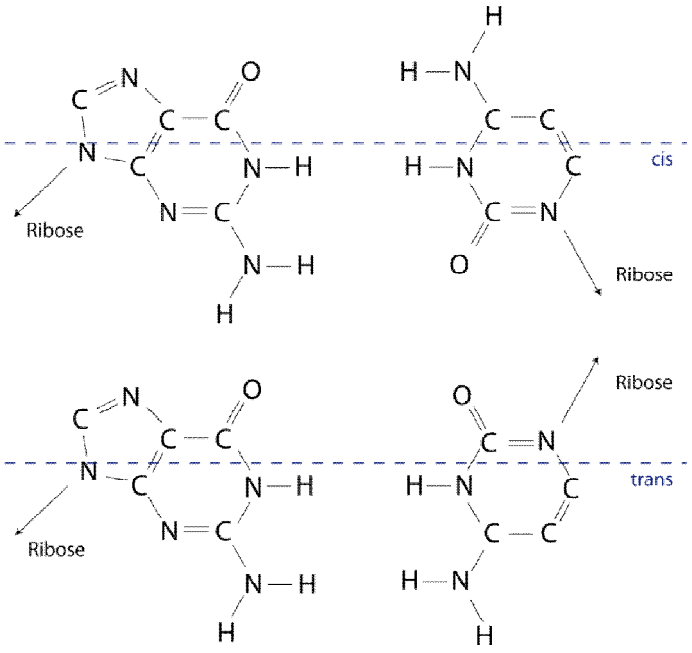


Figure 6. Orientations des paires de bases. À partir du plan formé par la paire de bases, si les deux riboses sont orientés du même côté de la médiane (ligne pointillée), l'orientation de la paire de bases est *cis*. Si les riboses sont dirigés de chaque côté, l'orientation de la paire de bases est *trans*.

1.1.2.1.2.1 Matrice d'isostéricité

Les matrices d'isostéricité représentent un classement de la géométrie des paires de bases, où chaque classe se compose d'un ensemble de séquences. Les séquences au sein d'une même classe d'isostéricité ont la même géométrie et sont donc interchangeables dans une structure tertiaire.

Les matrices d'isostéricité peuvent être utilisées afin de caractériser des motifs d'ARN, de définir de nouvelles variantes pour ces motifs, et d'analyser leurs covariations dans les alignements dans le but de les corriger [3].

Pour bâtir les matrices d'isostéricité, Leontis *et al.* [25] ont superposé la géométrie pour chaque type de paires de bases (X●Y, X●■Y, etc.) des séquences possibles (A, C, G et U) et les ont classées selon l'orientation de leur atome C_1'-*base* (voir **Figure 7**). Par exemple, une paire de bases C❍□A est isostérique à une paire de bases U❍□A. En considérant les 6 types de paires de bases (combinaison de ●, ■ et ▶) et leurs 2 orientations chacune (*cis* et *trans*), il y a 12 matrices d'isostéricité (voir **Figure 8**). [25]

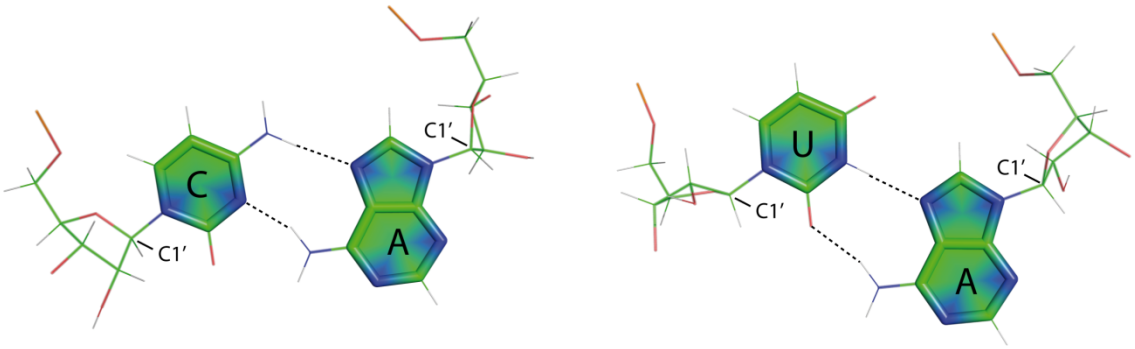


Figure 7. Paires de bases isostériques. Exemple de paires de bases isostériques : C❍□A et U❍□A. L'orientation relative des deux C_1' dans chacune des paires de bases est semblable. Les nucléotides sont illustrés par des formes planaires. Les atomes d'azote sont en bleu ; oxygène en rouge ; carbone en vert ; hydrogène en gris.

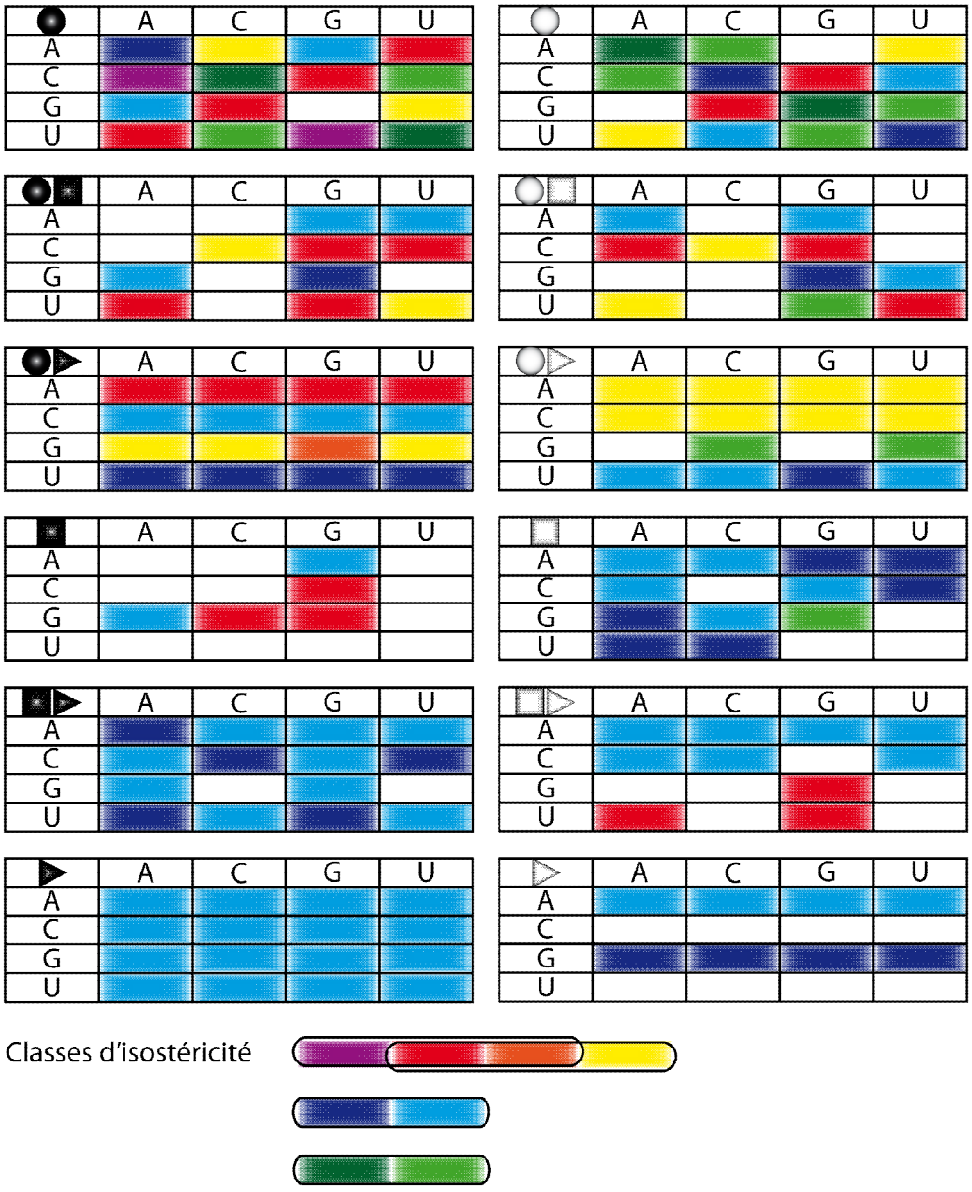


Figure 8. Matrices d'isostéricité. Chaque matrice correspond à un type de paires de bases (Watson-Crick ● ; Hoogsteen ■ et Sucre ▶, d'orientation *cis* ● ou *trans* ❍). Les couleurs dans les cellules sont des sous-classes d'isostéricité au sein d'une même matrice (la relation d'isostéricité ne s'applique pas d'une matrice à l'autre, par exemple une paire de bases A●U n'est pas isostérique à une paire de bases C❍G). Les cellules blanches représentent des paires de bases impossibles.

1.1.2.1.3 Empilement de bases

Des pointes de flèches sont utilisées pour indiquer l'orientation d'une base, indépendamment de la direction de la liaison phosphodiester (voir **Figure 9** et **Figure 10**). Les bouts des flèches indiquent la normale du plan des bases pyrimidines (C ou U), telles que définie dans une double-hélice d'ARN classique, où les vecteurs normaux sont orientés vers l'extrémité 3' [26]. Pour les pyrimidines, ce vecteur normal est obtenu par la règle de la main droite suivant le système d'axes définie par N1 à C6 autour de l'anneau de la pyrimidine. Cet anneau chez les purines (A ou G) est inversé par rapport aux pyrimidines.

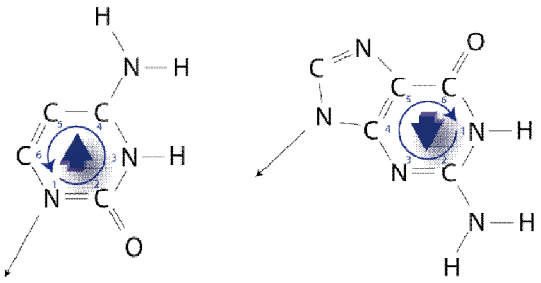


Figure 9. Vecteurs normaux. Identification du vecteur normal chez une pyrimidine (à gauche et dans ce cas-ci un C), où le vecteur normal sort de l'image, et chez une purine (à droite et dans ce cas-ci un G), où le vecteur normal entre dans l'image.

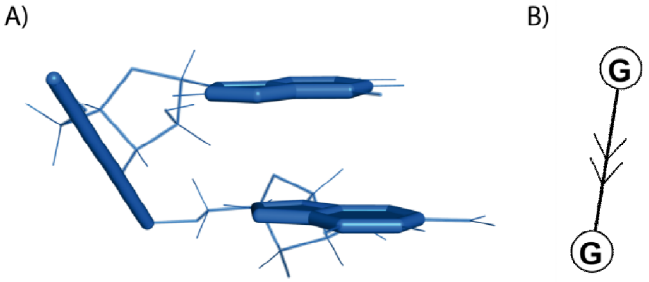


Figure 10. Empilement. A) Structure tertiaire. Un empilement dans la structure tertiaire, où les liaisons phosphodiesters sont illustrées par un cylindre. Les nucléotides sont illustrés par des formes planaires. B) Graphe d'interactions. Un empilement dans le graphe

d'interactions, où les liaisons phosphodiesters sont représentées par une arête et l'empilement par des flèches (>>). La séquence est représentée par A, C, G et U.

Les deux orientations possibles des vecteurs normaux pour chacune des deux bases empilées produit quatre types d'empilements de bases : upward (>>), downward (<<), outward (<>) et inward (><). Les deux pointes de flèches pointant dans la même direction (downward et upward) correspondent au type d'empilement dans une double-hélice d'ARN classique. Upward et downward sont choisit dépendamment de quelle base est référée en premier (c'est-à-dire A>>B signifie que B est empilé upward à A, ou A est empilé downward à B.) Les deux autres types d'empilements sont moins fréquents, respectivement inward (A><B; A ou B est empilé inward à, respectivement B ou A) et outward (A<>B; A ou B est empilé outward à, respectivement B ou A.

1.1.2.2 Cycle

Dans un graphe, un cycle est une suite d'arêtes consécutives dont le sommet de départ et de fin sont identiques. Un cycle est minimal s'il ne contient pas d'autres cycles. Dans un cycle minimal, le degré de tous les sommets est égal à deux. Par exemple, la **Figure 11** illustre les cycles minimaux du motif Sarcin-Ricin, qui sont identifiés de C_1 à C_5. Horton [27] a établi un algorithme polynomial qui trouve tous les cycles d'un graphe.

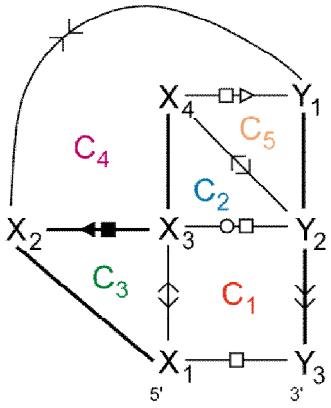


Figure 11. Graphe d'interactions du Sarcin-Ricin. Les cinq cycles minimaux : C_1 à C_5. Les liaisons phosphodiesters sont représentées en gras, les paires de bases par les symboles ●,■, etc., les empilements pas les symboles >>, ><, etc. et les nucléotides par X_1 à X_4 et Y_1 à Y_3.

1.1.2.3 Motifs cycliques de nucléotides (NCM)

Les motifs cycliques de nucléotides (NCM) « Nucleotide Cyclic Motif » sont comme les cycles, mais ne tiennent pas compte des interactions d'empilement. En fait, ils forment des cycles minimaux à partir des interactions de paires de bases et des liaisons phosphodiesters seulement. En ignorant les interactions d'empilement, les NCM adjacents partagent généralement une paire de bases commune.

Il existe deux sortes de NCM, ceux à double brins qui composent les hélices et ceux à un seul brin qui composent les boucles (voir **Figure 12**). Les NCM à double brins sont bornés par une paire de bases à chaque extrémité et les NCM à un seul brin sont bornés par une paire de bases à l'extrémité de la boucle.

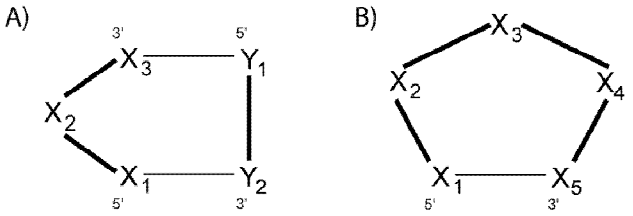


Figure 12. NCM. Les nucléotides sont étiquetés par des X_i et Y_i. Les liaisons phosphodiesters sont illustrées par des traits gras. Les interactions de paires de bases sont illustrées par des traits fins. A) NCM à double brins B) NCM à un seul brin.

1.1.3 Structure secondaire

La structure secondaire est un graphe d'interactions représentant la liaison phosphodiester et les paires de bases canoniques Watson-Crick (AU, GC et GU). Toutefois, seules les paires de bases qui composent les hélices y sont présentes.

La structure secondaire est souvent représentée en utilisant la notation d'expression parenthésée. Cette notation est un mot composé de points « . » et de parenthèses « (,) ». Chaque symbole correspond à un nucléotide de la structure secondaire où un point représente un nucléotide non apparié et une parenthèse représente un nucléotide apparié (l'association d'une parenthèse ouvrante avec une fermante représente une paire de bases dans la structure secondaire). Puisque le nombre de nucléotides qui participent à une paire de bases est pair (chaque nucléotide à son partenaire), les parenthèses sont balancées. Les parenthèses ouvrantes sont associées aux nucléotides en 3' des paires de bases et les fermantes sont associées aux nucléotides en 5' des paires de bases (voir **Figure 13**).

1.1.3.1 Prédiction de structure secondaire avec NCM

Pour prédire la structure secondaire à partir d'une séquence, nous utilisons l'outil *MC-Fold* [20] qui prend en entrée une séquence et prédit une liste ordonnée de structures secondaires. Comparativement aux autres outils de prédiction de structures secondaires qui

ne considèrent que les paires de bases classiques (AU, CG et GU), *MC-Fold* inclut les paires de bases non canoniques.

MC-Fold utilise une base de données de NCM pour générer un ensemble de structures secondaires sous-optimales pour une séquence donnée (voir **Figure 13**). Les structures sont évaluées et triées selon leur stabilité et en fonction de la séquence. Chaque structure se voit attribuée une valeur qui est transformée en énergie selon une distribution de Boltzmann [28]. Les structures secondaires générées sont ensuite représentées par des expressions parenthésées.

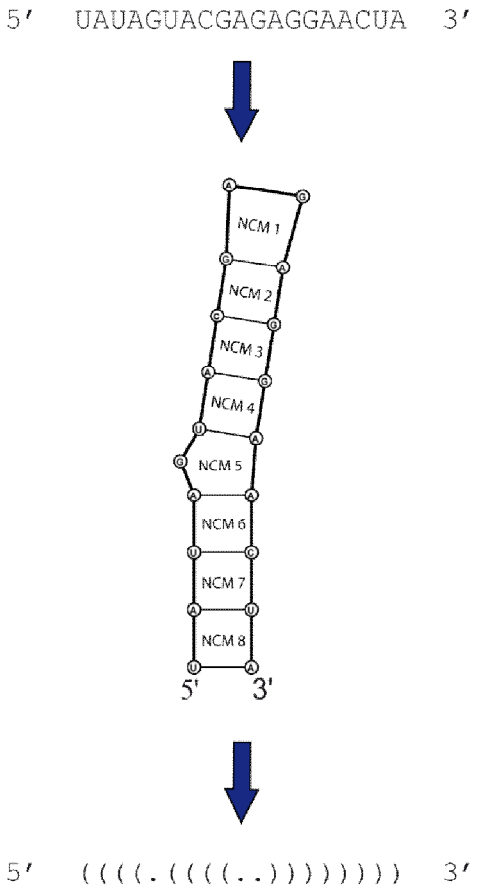


Figure 13. Principe de *MC-Fold*. *MC-Fold*, prend une séquence en entrée et utilise la notion de NCM pour produire un ensemble de structures secondaires possibles, représentées par des expressions parenthésées, pour lesquelles la séquence d'entrée a un

potentiel de repliement. Les liaisons phosphodiesters sont représentées par des traits gras, les paires de bases par des traits et la séquence par A, C, G et U.

1.1.4 Modélisation tridimensionnelle avec NCM

Dans le but de modéliser des structures tertiaires à partir d'un graphe d'interactions, nous utilisons l'outil *MC-Sym* « Macromolecular Conformations by SYMbolic programming » [21]. Cet outil utilise l'information contenue dans le graphe d'interactions à modéliser afin de construire des structures tertiaires correspondant à ce graphe.

Le procédé par lequel *MC-Sym* cherche des structures tertiaires cohérentes avec un graphe d'interactions est le même que de chercher des objets satisfaisant un certain nombre de contraintes. Il s'agit de la résolution de satisfaction de contraintes (CSP) « Constraint Satisfaction Problem » [29]. *MC-Sym* implémente ce procédé pour modéliser des structures tertiaires qui satisfont toutes les interactions d'un graphe d'interactions donné.

Pour y arriver, *MC-Sym* utilise une librairie de fragments tertiaires qui correspond à chaque NCM de la structure à modéliser. L'assemblage des NCM explore toutes les possibilités d'agencement des fragments tertiaires de NCM. Cet assemblage est obtenu par la superposition des deux copies de la paire de bases commune en trois dimensions (voir **Figure 14**).

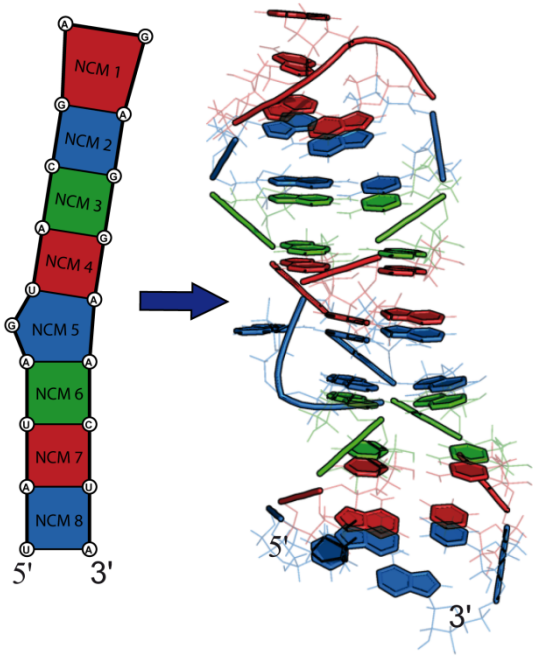


Figure 14. Principe de *MC-Sym*. En utilisant des fragments de NCM disponibles dans une librairie, *MC-Sym* les assemble afin de construire en trois dimensions le graphe d'interactions demandé. À gauche, les liaisons phosphodiesters sont représentées par des traits gras, les paires de bases par des traits et la séquence par A, C, G et U. À droite, les liaisons phosphodiesters sont illustrées par un cylindre, les nucléotides sont illustrés par des formes planaires. Les couleurs correspondent aux NCM.

1.2 QSAR « Quantitative Structure-Activity Relationship »

Afin de déterminer les propriétés d'une structure tertiaire nécessaires à la réalisation d'une fonction biologique donnée, nous nous sommes inspiré des méthodes de QSAR « Quantitative Structure-Activity Relationship ». Le QSAR est un modèle mathématique qui quantifie la relation entre la structure d'une molécule et son activité. Comme plusieurs propriétés différentes (nombre de carbones, poids moléculaire, présence de groupes rattachés aux carbones, etc) peuvent être considérées dans la relation entre une structure et son activité et que ces propriétés sont plus ou moins reliées à l'activité d'intérêt, le choix des propriétés pour exprimer la relation est important.

Toutes les propriétés considérées dans la relation entre une structure et son activité sont introduites dans un système d'équations linéaires. Résoudre ce système correspond à quantifier la contribution des propriétés pour l'activité biologique. La formulation QSAR la plus commune se traduit par l'équation **Eq. (1)**,

$$\begin{aligned} Activité_1 &= C_1P_{11} + C_2P_{12} + \dots + C_mP_{1m} \\ Activité_2 &= C_1P_{21} + C_2P_{22} + \dots + C_mP_{2m} \\ &\dots \\ Activité_n &= C_1P_{n1} + C_2P_{n2} + \dots + C_mP_{nm} \end{aligned} \qquad \textbf{(1)}$$

où chaque équation du système représente une structure, dont l'activité $Activité_i$ et les propriétés P_{ij} ont été mesurées et les contributions C_j ont été attribuées par la résolution du système d'équations linéaires.

1.2.1 Analyse en composantes principales (PCA)

Afin d'appliquer le modèle mathématique QSAR, nous utilisons l'analyse en composantes principales (PCA) « Principal Component Analysis », un modèle d'apprentissage, pour nous aider à choisir les propriétés reliées à l'activité.

Le PCA a été inventé en 1901 par Karl Pearson[30]. C'est une procédure mathématique qui utilise les transformations orthogonales pour convertir un ensemble d'observations (possiblement dépendantes les unes des autres) en un ensemble de variables indépendantes appelées composantes principales (voir **Figure 15**). Le nombre de composantes principales est inférieur ou égal au nombre d'observations originales. Cette transformation est définie de sorte que la première composante principale a la plus grande variance possible (elle représente le mieux l'ensemble des observations) et chacune des composantes suivantes a la plus grande variance possible (sous la contrainte d'être orthogonale aux composantes précédentes).

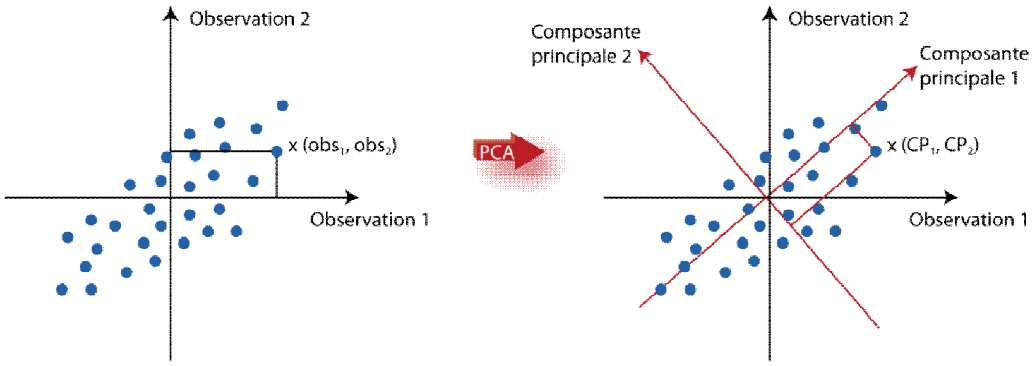


Figure 15. PCA. Un exemple de PCA à 2 dimensions. Gauche) Des observations (points) représenté dans un graphique où la coordonnée de chaque point est le couple ($observation_1$, $observation_2$). Droite) Les mêmes observations qu'à gauche. La composante principale 1 représente l'axe ayant le plus de variance. La composante 2 est l'axe orthogonal à la composante 1 ayant le plus de variance. Un point dans le graphique est le couple ($composante_1$, $composante_2$).

Chapitre 2 Représentation des familles d'ARN

2.1 Grammaire stochastiques hors-contexte

2.1.1 Grammaire hors-contexte

Un langage hors-contexte est un ensemble de séquences de caractères tiré d'un alphabet et spécifique à une représentation appelée grammaire. Formellement, une grammaire hors-contexte, $H = \{ N, \sum, P\}$, est formée d'un ensemble fini de symboles non-terminaux (ou variables), d'un ensemble fini de symboles terminaux et d'un ensemble fini de règles de productions. L'ensemble P des règles de production spécifie comment les séquences contenant des symboles non-terminaux peuvent être réécrites en changeant les symboles non-terminaux pour obtenir de nouvelles séquences. Le langage représenté par une grammaire est l'ensemble des séquences de symboles terminaux qui peuvent être obtenues à partir du symbole de départ $S_0 \in N$ en appliquant les règles de productions P [9].

Soit S un symbole non-terminal, a un symbole terminal et α une séquence de symboles terminaux et non-terminaux. Chaque production de P a la forme suivante : $S \longrightarrow \alpha$, indiquant qu'un symbole non-terminal S peut être remplacé par la séquence α. Pour des séquences d'ARN, l'alphabet des symboles terminaux $\sum$ correspond aux quatre nucléotides : A, C, G et U. Une règle de production $S \longrightarrow GSC$ représente une paire de bases G●C et une règle de production de la forme $S \longrightarrow SS$ représente un embranchement dans une structure secondaire (voir **Figure 16**). Une dérivation débute avec le symbole de départ S_0 et se termine lorsque la séquence contient uniquement des symboles terminaux.

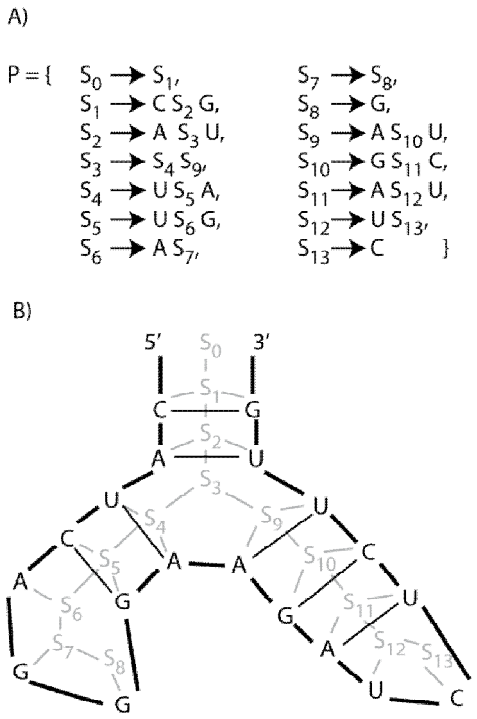


Figure 16. Grammaire hors-contexte. Une grammaire hors-contexte simple qui peut être utilisée pour dériver un ensemble de séquences d'ARN, de la forme CAUCNNNGAAGANNUCUUG. A) Règles de productions. Un ensemble de règles de production P qui génère des séquences d'ARN ayant certaines restrictions sur la structure. S_0 (départ), $S_1, \ldots, S_{13}$ sont les symboles non-terminaux ; A, C, G et U sont les symboles terminaux. Une application des règles de production P qui génère la séquence CAUCNNNGAAGANNUCUUG par les dérivations indiquées. Par exemple, la règle $S_1 \longrightarrow \mathrm{C}\, S_2\, \mathrm{G}$ transforme la séquence S_1 par $\mathrm{C}\, S_2\, \mathrm{G}$. B) Structure secondaire. La structure secondaire de l'ARN associée à la dérivation des règles de production.

2.1.2 Grammaire stochastique hors-contexte

Sakakibara *et al.* [9] ont développé une grammaire stochastique hors-contexte pour modéliser les ARN de transfert (ARNt) afin de discriminer les séquences de différentes familles. Ils ont montré que cette méthode de modélisation a la flexibilité et l'efficacité

requise pour résoudre les problèmes d'appartenance à une famille d'ARN, d'alignement multiple et de prédiction de structure secondaire.

Dans une grammaire stochastique hors-contexte, chaque règle de production a une probabilité. Ces probabilités sont utilisées pour définir la probabilité d'une dérivation (produit des probabilités des productions).

2.1.3 Faiblesses des grammaires stochastiques hors-contexte

Les grammaires stochastiques hors-contexte ont permis de discriminer 96% des séquences provenant de plusieurs familles d'ARNt. Toutefois, une des difficultés majeures de cette application est le développement de la grammaire initiale appropriée, c'est-à-dire de définir chacune des règles de production ainsi que les probabilités qui leurs sont associées. Pour faciliter le développement de la grammaire, Sakakibara *et al.* réfèrent, entre autres, aux modèles de covariances développés par Eddy et Durbin [10]. Une autre difficulté est le temps d'exécution (n^3) pour traiter les longues séquences, qui pourrait se résoudre en utilisant une heuristique, par exemple. De plus, une grammaire hors-contexte, comme son nom l'indique, manque de contexte entre les dérivations, c'est-à-dire qu'une paire de bases est générée sans tenir compte des autres paires de bases qui l'entourent.

2.2 Modèles de covariance

Eddy et Durbin [10] ont appliqué les modèles de covariance à la prédiction d'une structure secondaire consensus, à l'alignement de séquences et à la recherche dans les données génomiques.

Un modèle de covariance est un modèle probabiliste qui décrit la flexibilité des familles de structures secondaires. Étant donné qu'un arbre peut représenter les paires de bases d'une structure secondaire, le modèle de covariance est basé sur un arbre ordonné.

Cependant, un arbre ne peut pas représenter les autres interactions du graphe d'interactions, telles que les empilements par exemple. [10]

La **Figure 17** illustre un arbre ordonné représentant une séquence d'ARN. Plusieurs arbres différents peuvent représenter la même séquence, mais le meilleur arbre, dans le cas des modèles de covariance, est celui dont chaque nœud d'une paire de bases est associé à une paire de bases dans la structure secondaire. Un arbre représente donc, à la fois, une séquence et une structure secondaire. La séquence peut être retrouvée en parcourant l'arbre de manière infixe.

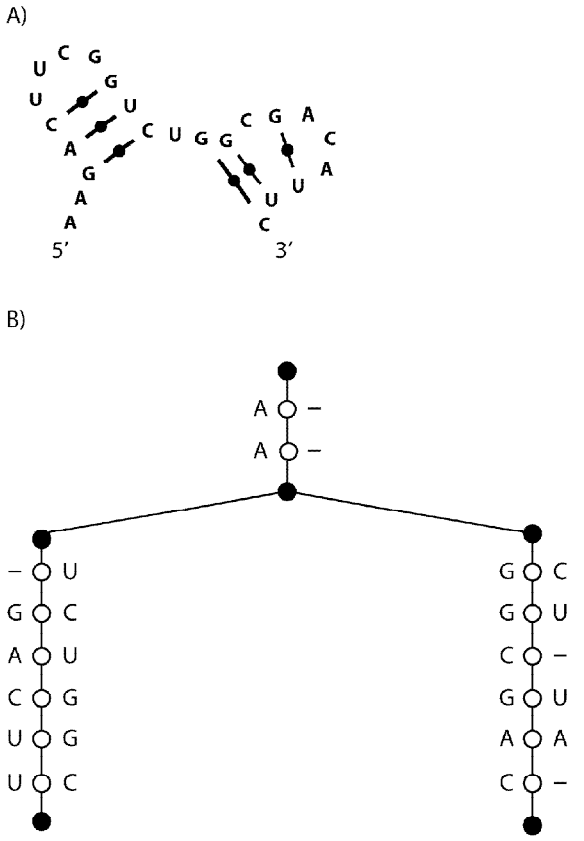


Figure 17. Arbre ordonnée. (A) Un exemple d'une structure secondaire d'ARN. (B) Un arbre binaire ordonné représentant la structure en (A). L'arbre contient des nœuds de début, de fin et de branchement (bifurcation) ● ainsi que des nœuds de paires de bases et de singleton ○ pour accommoder la séquence.

Un tel arbre est une description inflexible d'un ARN. Pour représenter une famille d'ARN, il faut pouvoir décrire des insertions, suppressions et substitutions (« mismatches »). Il suffit d'utiliser un arbre où chaque nœud décrit deux colonnes d'un alignement multiple au lieu d'un nucléotide provenant d'une seule séquence. La **Figure 18** illustre les différents états, qui remplacent les nœuds, permettant de décrire les variations entre les structures secondaires. Les états d'identité (MATP, MATL, MATR) représentent les colonnes conservées d'un alignement par rapport à une structure consensus. Les états d'insertions (INSL, INSR) représentent des insertions relatives à la structure consensus. L'état de suppression (DEL) n'émet aucun nucléotide et permet les suppressions par rapport à la structure consensus. Les deux états MATL et MATR permettent une suppression d'un côté ou de l'autre de la paire de bases consensus, produisant ainsi un « bulge ». Les états sont reliés entre eux par des transitions probabilistes.

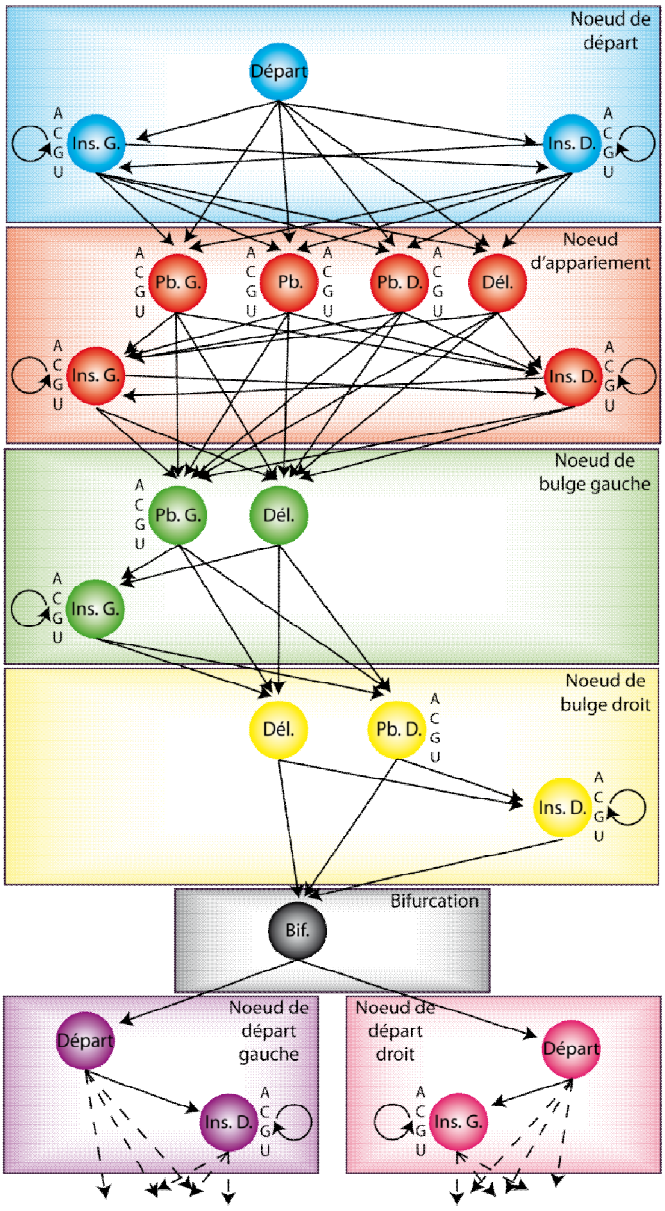


Figure 18. États d'un modèle de covariance. Les sept types de nœuds de la **Figure 17** sont divisés en états. Il y a sept états différents : bifurcation (Bif.), départ (Départ), insertion à gauche (Ins. G.), insertion à droite (Ins. D.), paire de bases (Pb.), paire de bases à gauche (Pb. G.), paire de bases à droite (Pb. D.) et délétion (Dél.). Les transitions sont indiquées par les flèches. Les états qui représentent des nucléotides sont indiqués en ayant ACGU à côté.

2.2.1 Faiblesses des modèles de covariance

Les modèles de covariance produisent des alignements multiples ayant une précision supérieure à 90%, toutefois les régions imprécises sont causées par l'information du graphe d'interactions non considérée. Étant donné que les modèles de covariances ignorent les covariations du graphe d'interactions, cette méthode de modélisation n'est qu'une aide et non une solution complète afin de produire des alignements très précis. Eddy et Durbin [10] les justifient en affirmant que la majorité de l'information importante pour un ARN provient de sa séquence et de sa structure secondaire. L'information contenue dans le graphe d'interactions, est toutefois essentielle afin d'augmenter la spécificité des méthodes de modélisation.

Ils ont recherché des ARNt dans les données génomiques et ont comparé leur méthode de modélisation à *tRNAscan* [31], un outil spécialisé dans la recherche d'ARNt dans les séquences, obtenant une sensibilité meilleure de 2%. Cette méthode de modélisation permet donc de rechercher des ARN ayant des structures secondaires similaires malgré des séquences possiblement très différentes. Par contre, leur méthode de modélisation est limitée à la recherche de séquences inférieures à 200 nucléotides.

De plus, les modèles de covariance ne tiennent pas compte du contexte autour des paires de bases. C'est-à-dire que la covariation d'une paire de base ne dépend pas des nucléotides avoisinants.

2.3 Profils de structure secondaire

Lambert *et al.* [15] ont utilisé les profils de structure secondaire afin de développer un outil de recherche de motifs dans un alignement de séquences. Cet outil, *ERPIN* « Easy RNA Profile IdentificatioN » [14] prend en entrée un alignement et un choix d'une structure secondaire (parmi une cinquantaine dont les profils sont préalablement construits) et recherche cette structure choisie dans l'alignement fourni.

Un profil de structure secondaire, basé sur un alignement multiple, contient la fréquence d'apparition de chaque nucléotide dans chacune des colonnes de l'alignement et est représenté par une forêt (ensemble d'arbres ordonnés). La **Figure 19** illustre la représentation d'une structure secondaire en une forêt. Les paires de bases de la structure secondaire correspondent aux nœuds internes de la forêt étiquetés des nucléotides qui forment les paires. Les nucléotides non appariés de la structure secondaire correspondent aux feuilles de la forêt étiquetées du nucléotide. [32]

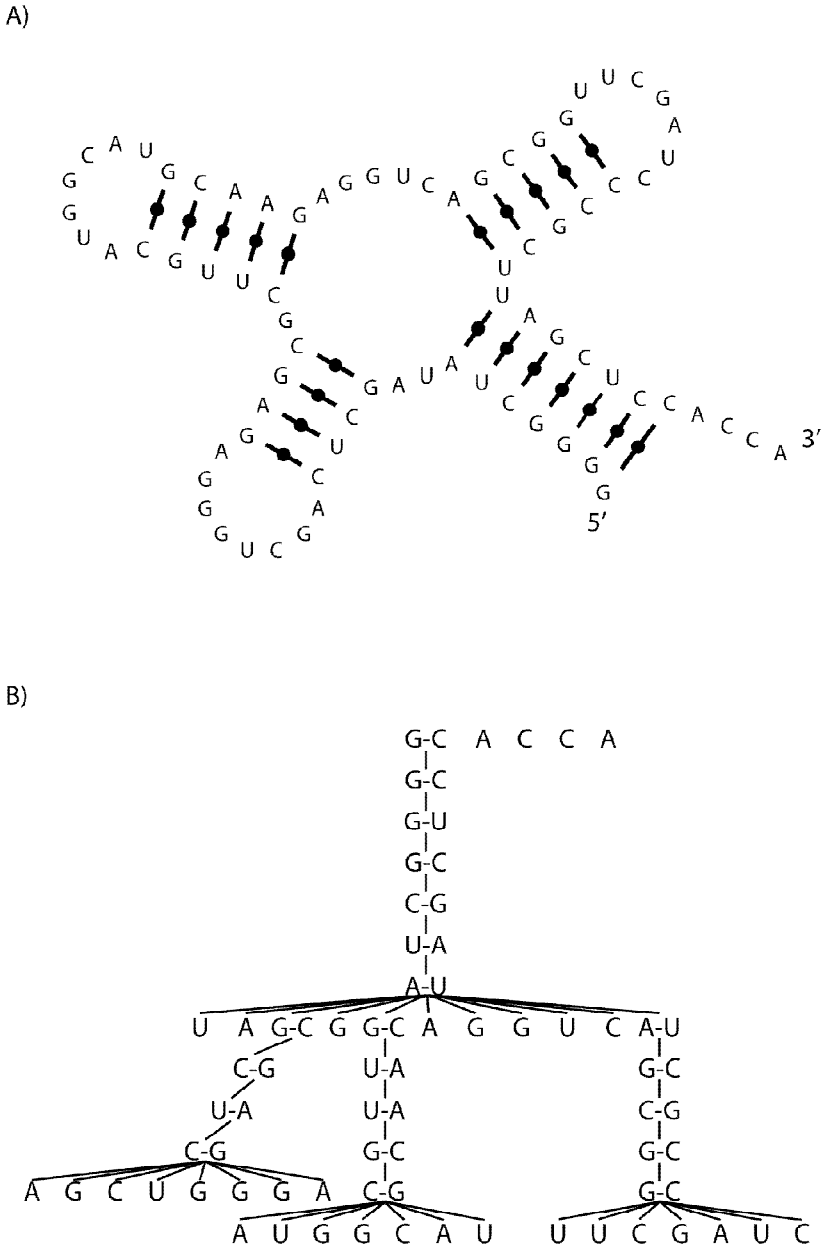


Figure 19. Forêt. A) Structure secondaire. La structure secondaire de l'ARNt de *Escherichia coli*. B) Représentation en forêt. La représentation en forêt de (A). Les paires de bases correspondent aux nœuds internes étiquetés des nucléotides qui forment les paires. Les nucléotides non appariés correspondent aux feuilles étiquetées du nucléotide.

Un profil d'un alignement de séquences est donc une forêt dont les nœuds sont étiquetés par des fréquences, tel qu'illustré à la **Figure 20**. Les colonnes de fréquences de la forêt, sont associées aux fréquences des nucléotides dans l'alignement.

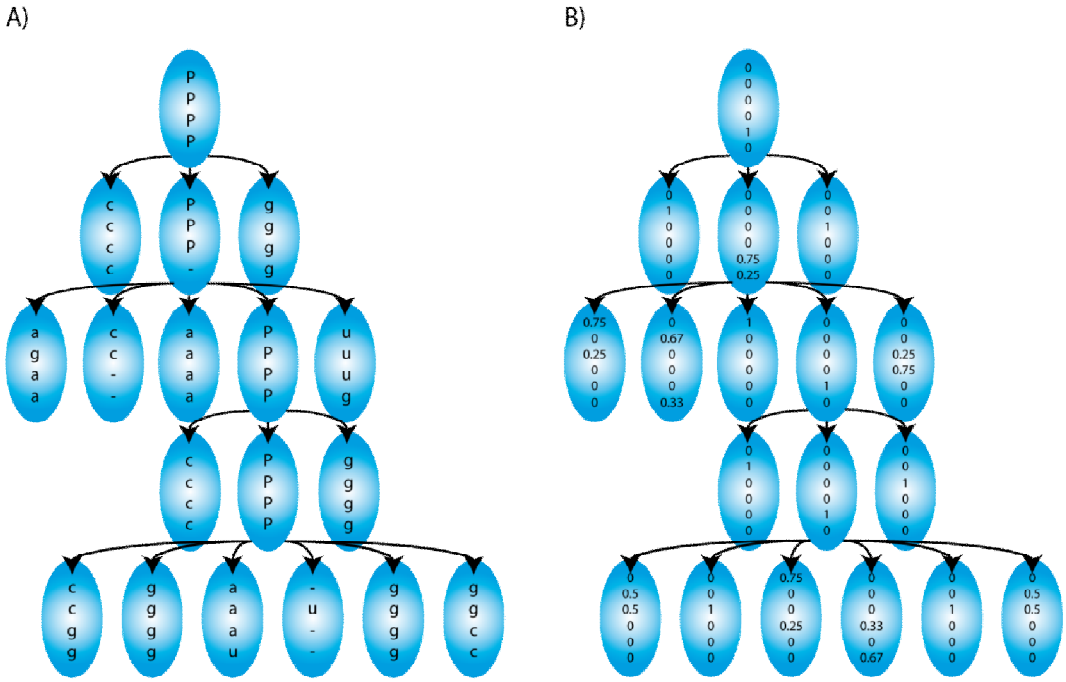


Figure 20. Profil d'un alignement. A) Alignement multiple. La représentation d'un alignement multiple de structures secondaires. B) Profil. Le profil correspondant de (A). Les colonnes de fréquences, du haut au bas, sont associées aux fréquences des nucléotides A, C, G, U, *P* (pour apparié) et – (non disponible).

2.3.1 Faiblesses des profils de structure secondaire

La limitation de cet outil est la restriction des motifs prédéfinis, toutefois pour ces motifs, ils obtiennent généralement une sensibilité de plus de 80%. Pour des familles de motifs moins structurés ou trop variables, les profils de structure secondaire ont une sensibilité de 60 à 70%.

Lambert *et al.* ne mentionnent pas l'impact des interactions du graphe d'interactions sur cette méthode de modélisation. Étant donné que les profils sont basés sur des structures secondaires, ils sont limités et ne tiennent pas compte des informations du graphe d'interactions. L'ajout des paires de bases non canoniques serait envisageable pour cette méthode, mais l'information des empilements de bases demeurerait absente pour la modélisation de motifs.

2.4 Réseaux de contraintes

Thebault *et al.* [33] ont recherché des ARNt dans les données génomiques à l'aide de réseaux de contraintes afin de comparer leur rapidité d'exécution et leur efficacité à d'autres outils.

Un réseau de contraintes est un triplet (V, D, C), où :

- $V = x_1, \ldots, x_n$ est un ensemble de n variables
- $D = d_1, \ldots, d_n$ est l'ensemble des n domaines associés à chaque variable
- $C = c_1, \ldots, c_s$ est l'ensemble des contraintes appliquées aux domaines et variables

Une solution d'un réseau de contraintes est l'association des valeurs des domaines aux variables, satisfaisant l'ensemble des contraintes [33][17].

Dans le cadre des motifs d'ARN, où un motif est composé de boucles, d'hélices, de boucle internes, etc., il est possible de caractériser un motif en définissant les relations entre ces différents éléments. Par la suite, une occurrence d'un tel motif dans les données génomiques est identifiée par les coordonnées relatives de ses éléments. Par exemple, la **Figure 21** illustre la recherche dans les données génomiques d'un motif à 5 paires de bases et ayant une boucle variant entre 5 et 22 nucléotides. L'ensemble des variables contient l'hélice (paires de bases) et la boucle, les domaines sont des sous-séquences et les contraintes visent à associer des sous-séquences compatibles à l'hélice et à la boucle.

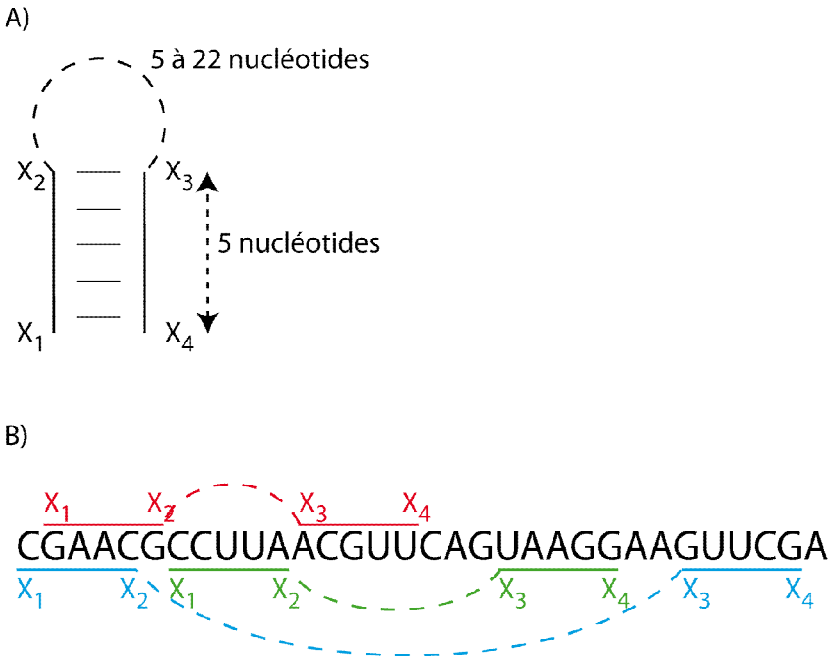


Figure 21. Recherche d'un motif. A) Motif. Illustration d'un motif ayant une hélice à 5 paires de bases et une boucle entre 5 et 22 nucléotides. B) Occurrences. Exemple de trois occurrences du motif en (A) trouvés dans une séquence.

2.4.1 Faiblesses des réseaux de contraintes

Leur méthode de modélisation a une sensibilité de plus de 95%, mais une spécificité maximale de 15% lors de la recherche d'ARNt, c'est-à-dire que leur méthode permet beaucoup de faux positifs et très peu de faux négatifs.

Leur caractérisation des éléments composant un motif (hélices, boucles, boucles internes) ne tient compte que de la structure secondaire. Il serait possible d'ajouter l'information des paires de bases non canoniques à leur méthode, mais pas celle des empilements de bases. De plus, l'information des paires de bases est prise hors contexte, c'est-à-dire qu'une séquence pour une paire de base n'influence pas les possibilités pour la séquence d'une paire de bases voisine.

Chapitre 3 Méthodes de quantification de la relation entre la structure et l'activité

3.1 Descripteur bidimensionnel de molécule

Morales Helguera *et al.*[34] ont appliqué le logiciel *DRAGON* [35] à un ensemble de 96 analogues de nucléosides (produits antiviraux) actifs et 165 analogues de nucléosides non-actifs. Parmi les 3224 descripteurs bidimensionnels disponibles de *DRAGON*, Morales Helguera *et al.* ont sélectionnés 259 descripteurs pour construire leur modèle QSAR. Parmi les descripteurs sélectionnés, il y a la topologie chimique (représentation planaire des atomes et des liens entre eux) ainsi que plusieurs informations provenant de cette topologie (nombre de chemins, indice de connectivité, etc.) Leur modèle permet de distinguer l'activité de 83,8% des analogues de nucléotides.

Le concept de descripteurs moléculaires est défini par Todeschini et Consonni [36]. C'est un procédé par laquelle les molécules sont analysées afin de quantifier un ensemble de propriétés. L'information d'une propriété est encodée par une valeur numérique, ce qui permet de représenter une molécule (ensemble de propriétés) par en ensemble de valeurs. L'ensemble des propriétés pouvant être considérées par les descripteurs moléculaires est très grand, plus de 4000 propriétés [35] et sont répertoriées en différentes classes : descripteurs 0D (nombre d'atomes, poids moléculaire, etc.), descripteurs 1D (ensemble de descripteurs 0D représentants des fragments de la molécule), descripteurs 2D (nombre de chemins, indice de connectivité, etc.), descripteurs 3D (volume, taille, surface, etc.) et descripteurs 4D (potentiel électrostatique, structure stérique, etc.)

Parmi toutes les propriétés qui peuvent être utilisées par les descripteurs moléculaires, les descripteurs bidimensionnels moléculaires ou 2D ne considèrent que les propriétés se rattachant à la structure secondaire des molécules.

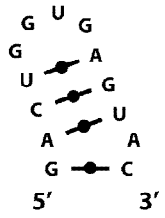
3.2 Matrice d'adjacences de nucléotides

Marrero Ponce *et al.* [37] ont utilisé les matrices d'adjacences de nucléotides dans le but de représenter l'affinité avec laquelle la paromomycin se lie à l'ARN HIV-1. Ils ont appliqué le modèle mathématique QSAR pour mettre en relation les matrices d'adjacences et la constante d'équilibre de la paromomycin lorsqu'elle se lie à l'ARN HIV-1. Leurs résultats démontrent qu'ils ont été en mesure de représenter l'affinité (r^2 = 0,87) avec laquelle la paromomycin se lie à l'ARN HIV-1.

En mathématiques, une matrice d'adjacences pour un graphe fini G à n sommets est une matrice de dimension $n \times n$ dont l'élément non-diagonal a_{ij} est le nombre d'arêtes liant le sommet i au sommet j. L'élément diagonal a_{ii} est le nombre de boucles au sommet i.

La **Figure 22B** illustre la matrice d'adjacences associée à la structure de la **Figure 22A**. Dans ce cas-ci, le nombre d'arêtes liant un sommet i à un sommet j est la somme des liaisons hydrogène et des liaisons phosphodiesters.

A)



B)

	G	A	C	U	G	G	U	G	A	G	U	A	C
G	0	1	0	0	0	0	0	0	0	0	0	0	3
A	1	0	1	0	0	0	0	0	0	0	2	0	0
C	0	1	0	1	0	0	0	0	0	3	0	0	0
U	0	0	1	0	1	0	0	0	2	0	0	0	0
G	0	0	0	1	0	1	0	0	0	0	0	0	0
G	0	0	0	0	1	0	1	0	0	0	0	0	0
U	0	0	0	0	0	1	0	1	0	0	0	0	0
G	0	0	0	0	0	0	1	0	1	0	0	0	0
A	0	0	0	2	0	0	0	1	0	1	0	0	0
G	0	0	3	0	0	0	0	0	1	0	1	0	0
U	0	2	0	0	0	0	0	0	0	1	0	1	0
A	0	0	0	0	0	0	0	0	0	0	1	0	1
C	3	0	0	0	0	0	0	0	0	0	0	1	0

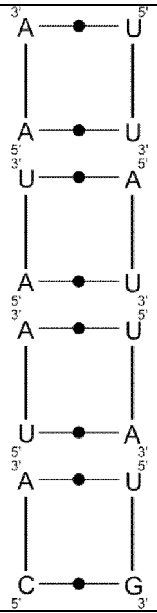
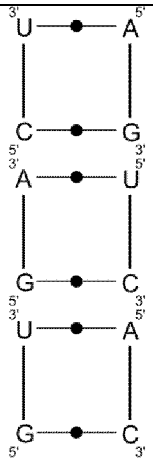
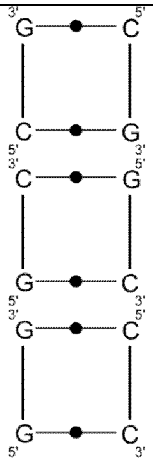
Figure 22. Matrice d'adjacences. A) Structure secondaire. Une structure secondaire d'ARN. B) Matrice d'adjacences. La matrice d'adjacences correspondant à (A). La séquence est indiquée sur la première ligne et la première colonne. Chaque case représente le nombre de liaisons (hydrogènes+ phosphodiesters) entre deux nucléotides de la structure.

3.3 Thermodynamique

Gonzalez-Diaz *et al.* [38] ont utilisé 623 micro ARN (miRNA) et 2000 séquences de contrôles négatifs générées aléatoirement. Ils ont utilisé le logiciel *Quickfold* afin d'obtenir les valeurs d'énergie libre, d'enthalpie, d'entropie et de température de fusion. Ils ont ensuite appliqué le modèle mathématique QSAR sur ces données. Leur méthode est capable de reconnaître 87% des miRNA et 93% des contrôles négatifs.

La thermodynamique est une science qui a pour but l'étude des formes d'énergie et de leurs transformations notamment transformations de chaleur en travail et *vice versa*. Dans le cas des structures secondaires, des règles ont été établies permettant de calculer *a priori* l'énergie libre associée à la formation de chaque paire de bases [39][40]. Par exemple, la **Table 1** énumère l'énergie libre associée à chaque NCM canonique lors de la formation d'une double hélice. En connaissant ces règles, il est possible de calculer l'énergie libre d'une structure secondaire d'ARN. À partir de l'énergie libre associée à chaque NCM, il est possible de calculer l'énergie libre de la double hélice de la **Figure 23** en additionnant les valeurs de l'énergie libre ajoutée des NCM qui composent la double hélice.

Table 1 Énergie libre ajouté des NCM canoniques. La première (troisième et cinquième) colonne illustre un NCM. La deuxième (quatrième et sixième) colonne indique l'énergie libre ajoutée (kcal/mol) du NCM dans une double hélice [40].

NCM	Énergie libre ajoutée	NCM	Énergie libre ajoutée	NCM	Énergie libre ajoutée
	-6.6 -5.7 -8.1 -10.5		-7.6 -13.3 -10.2		-8.0 -14.2 -12.2

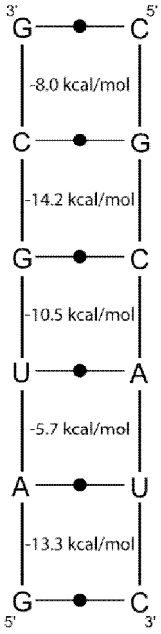


Figure 23 Double hélice. L'énergie libre de cette double hélice est de -51.7 kcal/mol (-8.0kcal/mol -14.2 kcal/mol -10.5 kcal/mol -5.7 kcal/mol -13.3 kcal/mol). Les liaisons phosphodiesters sont représentées par des traits gras, les paires de bases par le symbole ● et la séquence par A, C, G et U.

3.4 Faiblesses des méthodes existantes

Les descripteurs bidimensionnels, les matrices d'adjacences et la thermodynamique se réfèrent à des propriétés relatives à la structure secondaire et au graphe d'interactions, afin d'appliquer le modèle mathématique QSAR pour relier l'activité à la structure. Par contre, ces méthode ne peuvent pas s'appliquer dans le cadre qui nous intéresse, c'est-à-dire celui d'identifier des propriétés chimiques au niveau de la structure tertiaire, puisqu'elles se limitent à la structure secondaire et au graphe d'interactions.

Bien qu'il existe des méthodes QSAR qui se basent sur des descripteurs 3D, c'est-à-dire aux propriétés au niveau de la structure tertiaire, elles s'appliquent toutes aux

protéines. À notre connaissance, aucune équipe n'a développé une méthode de quantification de la relation entre la structure tertiaire et l'activité appliquée à l'ARN.

Chapitre 4 Modeling RNA tertiary structure motifs by graph-grammars

4.1 Résumé

Nous avons développé les grammaires de graphes pour encoder le graphe d'interactions d'un motif d'ARN, Sarcin-Ricin, qui est présent dans la boucle E du ribosome et qui est sensible aux toxines α-sarcin et ricin. Les grammaires de graphes appliquées au Sarcin-Ricin, identifient les séquences qui se replieraient dans ce motif.

Nous avons confirmé la pertinence biologique des séquences du Sarcin-Ricin par une comparaison avec celles trouvées dans un alignement d'un site connu de plus de 800 séquences ribosomales d'ARN de bactéries.

L'article « Modeling RNA tertiary structure motifs by graph-grammars » dont les auteurs sont K. St-Onge, P. Thibault, S. Hamel, et F. Major, a été publié dans le journal « Nucleic Acids Research » en 2007 [41].

4.2 Partage du travail

Dans cet article, j'ai développé l'outil informatique *MC-Seq*, une grammaire de graphes et je l'ai appliquée au motif d'ARN du Sarcin-Ricin (présent dans l'ARN ribosomal des bactéries) sous la supervision de F. Major et S. Hamel. P. Thibault a modélisé la structure de la figure 6C.

Modeling RNA tertiary structure motifs by graph-grammars

Karine St-Onge[1,2], Philippe Thibault[1,2], Sylvie Hamel[1] and François Major[1,2†]

[1]Department of Computer Science and Operations Research
Université de Montréal
PO Box 6128, Downtown station
Montreal, Quebec H3C 3J7
CANADA

[2]Institute for Research in Immunology and Cancer
Université de Montréal
PO Box 6128, Downtown station
Montreal, Quebec H3C 3J7
CANADA

[†]To whom correspondence should be addressed.
François Major
Tel: 514 343 6752
Fax: 514 343 5839

Nucleic Acids Research (2007)
(Publié)

Abstract

We encoded the tertiary structural features of the sarcin-ricin motif using a graph-grammar, and we employed it to derive a set of valid sarcin-ricin sequences. We confirmed the biological relevance of the derived sequences by using an alignment of the bacterial sequences of 23S ribosomal RNA subunits at their sarcin-ricin sites. The graph-grammar allowed us to make biologically coherent alignment hypotheses that were assessed by tertiary structure prediction and three-dimensional modeling, which point to plausible evolutionary events that might have occurred at the sarcin-ricin sites in the ribosome. The use of indivisible interaction cycles made possible the utilization of graph-grammars to represent RNA motifs. Here, we also showed the cycles are fundamental RNA building blocks. In spite the fact that we removed all instances of the sarcin-ricin motif from the learning database, the five cycles that constitute it were found in other structural contexts and so to produce an equivalent graph-grammar that derives the same set of sarcin-ricin sequences. Generating RNA graph-grammars and deriving their sequence space are both tractable processes.

Introduction

Recently, the resolution of RNA X-ray crystal structures revealed, in the context of their biologically active hosts, several RNA motifs that were previously studied experimentally as individual fragments. Many of these motifs were predicted from comparative sequence analysis [1], indicating the existence of a relation between their sequence and structure. The recent structures confirmed that RNA motifs fold in stable conformations, in the context of their hosts, and are found involved in important intra- and inter-molecular stabilization interactions, as well as in catalytic domains [3][4]. Consequently, it is now largely acknowledged that RNA motifs are crucial elements of RNA tertiary structure and function [6][8][5][7].

During the last decade, RNA motifs have been computationally represented, among others, by stochastic context free grammars (SCFG) [9], covariance models [10][11][12], secondary structure profiles [15][14], and constraint networks [17]. Most of these

computational models are inferred from sequence alignments. They allow us to parse, or fit, RNA sequences into their plausible secondary structure elements and to seek for new instances in genomic data. In addition to parse RNA sequences, some computational models, such as SCFGs, directly generate the set of sequences that are compatible to any given motif, which is a necessary step towards *in silico* selection and the engineering of RNA sequences with predetermined structure and function.

Indeed, current computational representations are sensitive and subject to their input sequence alignments. Unfortunately, inferring structure and alignment goes beyond secondary structure, covariation analysis, and sequence similarity. It involves a difficult and iterative process of pattern matching and modeling. The putative patterns in each alignment must be validated in terms of their structure, which base pair substitution rules, for instance, are constrained by the local structural context that include subtle factors including base stacking [3][42]. As a consequence, it is currently difficult to consider structural information that is not explicitly encoded in an alignment. For instance, one would need to use isostericity matrices to include additional base pairs that are not seen in a current alignment [3], and sequence homology scores (cf. *Blast*) to match approximate single-stranded regions [14].

As RNA crystallographic data accumulate [43], we can now conceive a direct inference of RNA motif information that is pertinent and necessary to improve genomic searches and sequence alignment. Here, we show how graph-grammars [44][45] were used to encode RNA motifs in the context of their tertiary structure and to parse and generate compatible RNA sequences. We exemplify the use of graph-grammars by generating a set of sequences that accommodate the sarcin-ricin motif [46][47]. We then applied the graph-grammar to make hypotheses about the sequence alignment of the sarcin-ricin regions of the 23S ribosomal RNA (rRNA) subunit of Bacteria. Finally, we used graph-grammars to show how the sequence-structure relation of a specific motif (here the hexaloop) can be established. Deriving a graph-grammar from small RNA patterns, or motifs, and predicting their compatible sequences are both tractable processes.

Materials & Methods

RNA graph and tertiary structure. The tertiary structure of an RNA can be represented computationally by a graph, $G = \{V, E\}$, where V is the set of nucleotide vertices, or nodes, and E is the set of interaction edges. In comparison to secondary structure, which describes the sequence (backbone interaction) and the canonical Watson-Crick base pairs of the RNA, the tertiary structure includes all nucleotide interactions: the backbone, the canonical and non-canonical base pairs (base-base H-bonds), the base-backbone and base-sugar H-bonds, and the base stacking. Note that more than one interaction per edge can exist, and so the edge types include all possible interaction combinations (cf. backbone/stack, backbone/pair, and so on).

Nomenclature. Consider the three-dimensional (3-D) structure of the sarcin-ricin motif shown in **Figure 24A**. In order to represent this structure in an RNA graph, we need to specify the nucleotide nodes and the type of their interacting edges. Consequently, we need an unambiguous nomenclature to distinguish all types. **Figure 24B** shows the tertiary structure and graph of the sarcin-ricin motif.

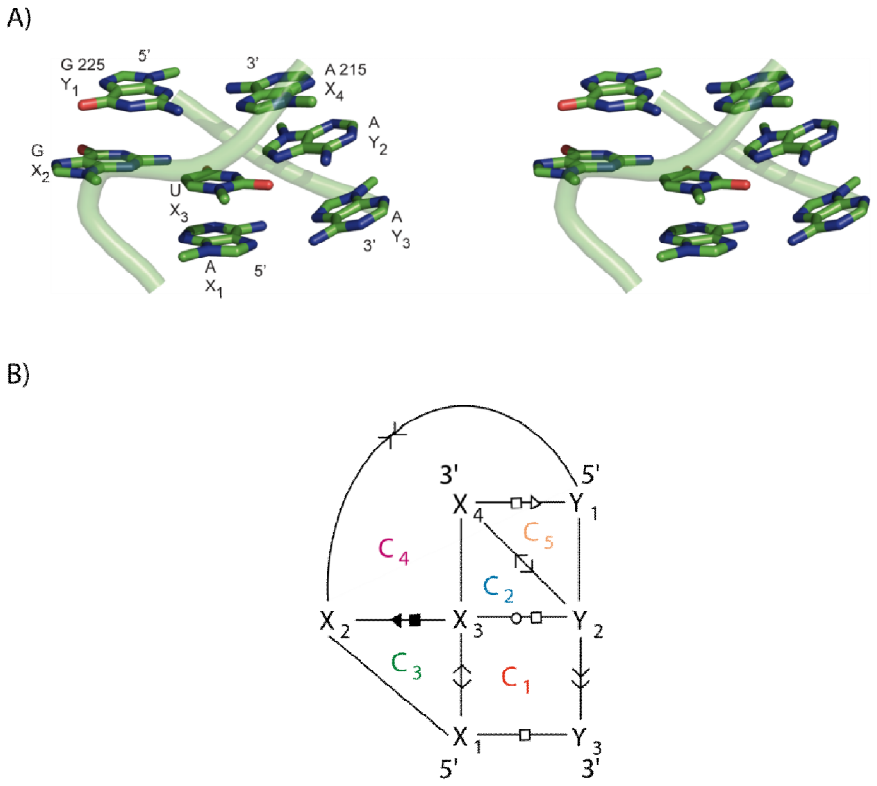


Figure 24. The sarcin-ricin motif. A) Stereoview of the 3-D structure. The nucleotides are labeled by the X_i (5'-strand) and Y_i (3'-strand). The backbone is shown using a light green cylinder. Nitrogen atoms are in blue; oxygen in red; and carbon in green. The hydrogen atoms are not shown. B) Tertiary structure and cycles. The minimal cycle basis of the sarcin-ricin motif is made of five minimum cycles: C_1 to C_5. The symbols used to indicate base stacking and base pairing are described in Materials & Methods (see Nomenclature).

Base stacking. Arrows are used to indicate the orientation of a base, independently of the backbone direction. The tips of the arrows indicate the normal of the base pyrimidine plane, as defined in a classical A-RNA type double-helix, where the normal vectors are oriented towards the 3'-strand endpoint [26]. Two orientations in two bases result in four base stacking types: upward (>>), downward (<<), outward (<>), and inward (><). Two arrows pointing in the same direction (upward and downward) corresponds to the stacking type in canonical A-RNA double-helices. We choose upward or downward depending on which

base we name first (cf. A>>B means B is stacked upward of A, or A is stacked downward of B). The two other types are less frequent in RNAs, respectively inward (A><B; A or B is stacked inward of, respectively B or A) and outward (A<>B; A or B is stacked outward of, respectively B or A). Note that all base stacking types are present in the sarcin-ricin motif shown in **Figure 24**.

Base pairing. The Leontis and Westhof nomenclature is used to describe the base pairs [24]. We indicate the face where the interacting chemical groups are found in each partner. We assign different names and symbols to each of the three faces of a base: the Watson-Crick face, ● (*cis*); ❍ (*trans*), the Hoogsteen face, ■ (*cis*); □ (*trans*), and the Sugar face, ◀ (*cis*); ◁ (*trans*) [24]. The *cis*/*trans* notation is for the relative orientation of the backbone according to the median of the plane formed by the two bases. When the two bases interact by H-bonds on the same face, only one symbol is used. For instance, $X\square\square Y$ is written $X\square Y$. We can also discuss the relative orientation of the bases in a base pair by using the arrows described above for base stacking. Similarly, a base pair can be parallel, if the two normal vectors point in the same direction, or antiparallel if not.

Seed motif. The sarcin-ricin motif shown in **Figure 24** was extracted from the 23S rRNA subunit of *Haloarcula Marismortui* of the Protein Data Bank [43] (PDB code 1JJ2). It is located at position A‘0’212-G‘0’213-U‘0’214-A‘0’215 / G‘0’225-A’0’226-A‘0’227. This instance was used as a seed motif to build the graph-grammar of the sarcin-ricin motif.

The graph of the sarcin-ricin motif is composed of seven nucleotides, forming two strands, 212-215 and 225-227, five backbone, four base pairs and four base stacking (7 nodes and 11 edges). In **Figure 24**, we generalized the graph by renaming its nodes with $X_1 - X_4$ and $Y_1 - Y_3$. The numbering of the Xs and Ys implies the backbone interactions. X_1 and X_3 stack outward, X_1<>X_3. $X_1$and Y_3 form a parallel *trans* Hoogsteen/Hoogsteen base pair, $X_1\square Y_3$. X_2 and X_3 form a parallel *cis* Sugar/Hoogsteen base pair, X_2◀■X_3. X_2 and Y_1 stack inward, X_2><Y_1. X_3 and Y_2 form an antiparallel *trans* Watson-Crick/Hoogsteen base pair,

X_3❍□Y_2. X_4 and Y_1 form an antiparallel trans Hoogsteen/Sugar base pair, X_4□▷Y_1. X_4 and Y_2 stack outward, X_4<>Y_2. Finally, Y_2 and Y_3 stack upward, Y_2 >>Y_3.

Shortest Cycle Basis. In **Figure 24B**, the indivisible cycles, or cycles for short, of the sarcin-ricin motif are identified by C_1 to C_5. RNA cycles are small RNA fragments defined by, as their name suggests, cycles of nucleotide interactions, or edges in the RNA graph [2]. The cycles of the sarcin-ricin graph, shown in **Figure 24B**, form its shortest cycle basis, as called in graph theory [27].

Graph-grammars. The first step to build the graph-grammar of an RNA motif is to determine its shortest cycle basis. The *MC-Cycle* computer program, developed in our laboratory, computes the shortest cycle basis of an RNA graph by implementing Horton's algorithm (unpublished results), which was developed for general graphs [27].

Formally, a graph-grammar, $\mathrm{H} = \{\mathrm{N}, \Sigma, \mathrm{P}\}$, is constituted of a set of non-terminal symbols, N, a set of terminal symbols, Σ, and a set of production rules, P. In the context of the sarcin-ricin motif, $N = \{C_1, C_2, \cdots, C_5\}$ is the set of cycles, $\Sigma = \{S_1, S_2, \cdots, S_5\}$ is the set of cycle sequences for each cycle, and P is a consistent assignment of Σ to S (see Derivation below).

Terminal symbols. To obtain the sequences of each cycle of an RNA motif, we search for their individual instances (not necessarily in the motif context) in the high-resolution (3 Å or better) 3-D X-ray crystallographic structures (RNA-3A) of the Protein Data Bank [43]. These instances are found by using a tool available in our laboratory, *MC-Search*, which takes as input a cycle and a database of 3-D structures (unpublished data). We consider the sequences of all instances matched in RNA-3A.

Derivation. We derive a set of consistent sequences for the motif by assigning the cycle sequences (terminals) to the cycles (non-terminals). This process is called "derivation" in formal grammars. We assign sequences to cycles by matching the nucleotides in the

common positions. Consider, once again, the five cycles of the sarcin-ricin motif (see **Figure 24**). For instance, C_1 and C_2 share X_3 and Y_2, and C_1 and C_4 share X_3. We define a two-dimensional table: cycle nucleotides (columns) × cycles (rows) (see **Figure 25**). A unique identifier labels each nucleotide. For each cycle and for each cycle sequence, we systematically replace the identifiers by their corresponding nucleotides. If the introduction of the cycle sequence does not introduce two different nucleotides in one position, then it is accepted and we proceed to the next cycle. If, in the contrary, the cycle sequence creates a conflict with at least one of the previously assigned position, then it is rejected and we try the next sequence for this cycle. If all the cycle sequences have been tried without success, then we "backtrack" to the next sequence of the previous cycle (see **Figure 26**). A sequence is compatible to the motif if a cycle sequence has been found compatible for each cycle.

C_1	X_1	X_3	Y_2	Y_3
C_2	X_3	X_4	Y_2	
C_3	X_1	X_2	X_3	
C_4	X_2	X_3	X_4	Y_1
C_5	Y_1	Y_2	X_4	

Figure 25. Derivation table. The table is made of one cycle per row and their corresponding nucleotides in the columns. The colors match the colors in **Figure 24**.

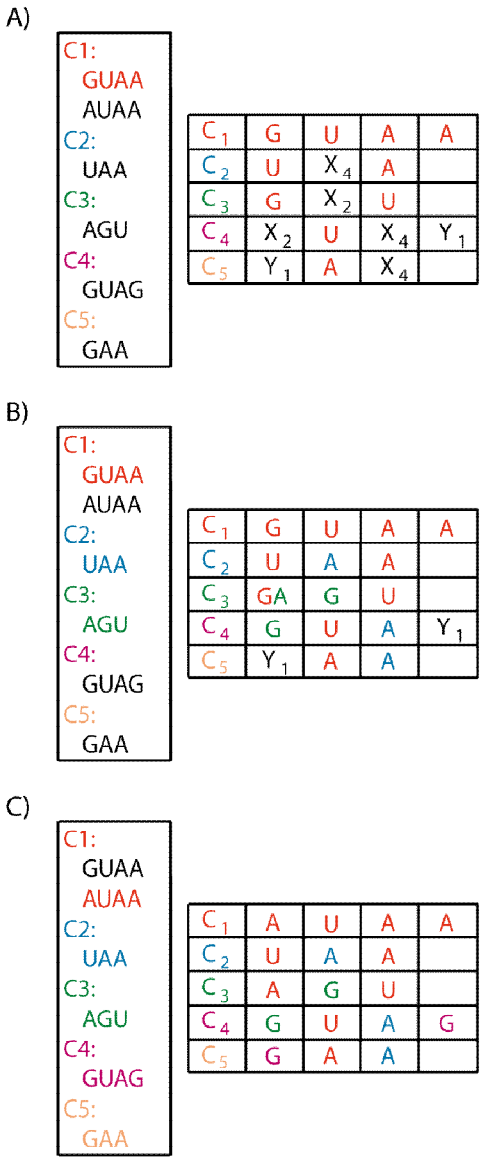


Figure 26. Graph-grammar derivation (on a reduced set of sequences). A) Insertion of the first C_1 sequence: GUAA (in red). B) Insertion of the first C_3 sequence, AGU (in green). This insertion is not possible because the first nucleotide of C_3, A, does not match with the first nucleotide of C_1, G, which was previously inserted in the table. Since no other sequence is available for C_3, the algorithm backtracks to the previous cycle, C_1, and selects its next sequence, AUAA. C) Last step. The insertion of the C_5 sequence, GAA, completes the sequence of the entire motif and represents a valid derivation of the graph-grammar: AGUA / GAA. The order of the nucleotides corresponds to the order given by the labels in **Figure 25**.

Insertions. RNA cycles are subject to insertions [2]. For instance, a nucleotide can be inserted between two others that are connected by a backbone/stack edge. The inserted nucleotide can bulge out, preserving the original stacking interaction and, thus, the cycle. To measure the impact of the backbone in the formation of the cycles and in the derivation results of the graph-grammar, we extracted the cycle instances with and without the presence of the backbone. As the number of RNA cycles is finite, we built *a priori* a database of sequences and cycles to speed up the building of graph-grammars.

Alignment. Westhof provided us with an alignment of the bacterial sequences of the 23S rRNA subunit (personal communication). The alignment corresponds to the one used to develop the concept of isostericity matrices [3]. We located in the alignment the positions that match the sequences derived by the graph-grammar, independently of the gaps (cf. ACGU matches A-CG--U). This defines a map indexed by the n-tuples of the positions that match in the motif strands (cf. we have two strands in the sarcin-ricin motif, so the map is made of 2-tuples) (see **Figure 27**). Starting at the position of the match in the seed motif crystal structure, we searched for matches in all sequences in a way to minimize the Manhattan distance to the seed motif match:

$$distance(\,(46-30),(46-10)\,) \;=\; |\,46\;-\;46\,|\;+\;|\,30\;-\;10\,|\;=\;20.$$

A)

```
         1        10        20        30        40        50        60        70        80
X15364   -----UGGGG-AG-GGGGA-A--CCCGCCG-AACUGAAACAUCUUA-GU-AGGCGGAGGAAGA---GAAAGC-AAAUUGC 01
X02729   --------GG-AU-UGGUA-A--CGCGGGG-GAUUGAAGCAUCUUA-GU-ACCCGCAGGAAAA---GAAAUC-AACU-GA 02
X72495   ------CGCA-AU-GGGGA-A--CGCCGAG-AACUGAAACAUCUCA-GU-AUCGGCAGGAAAA---GAAAAC-GUAAUGU 03
```

B)

```
(46-18): {1,3}
(46-30): {1,3}
(46-36): {1,2,3}
(46-59): {1,2,3}
(46-67): {1,2,3}
```

Figure 27. A parse-map in sequence data. A) First step. The graph-grammar identifies all sites corresponding to the two strands in the sarcin-ricin motif (shown in bold and underlined). B) The map data structure. The 2-tuples are mapped to the sequences that contain them. In the example, the 2-tuple (46-18) is found in sequences #01 and #03, (46-30) in #01 and #03, and so on.

Tertiary structure prediction. We evaluated the alignments that were not derived by the graph-grammar by tertiary structure prediction. We submitted the sequence to *MC-Fold*, a tertiary structure prediction program, currently under development in our laboratory (unpublished results). Among the optimal and suboptimal solutions proposed by *MC-Fold*, we compared the alignment with the prediction that minimizes the edition of the alignment.

Root-Mean-Square Deviations. We used the RNA fragment distance metric developed earlier in our laboratory [22] to compute the root-mean-square deviations (RMSD) between pairs 3-D structures.

Results and Discussion

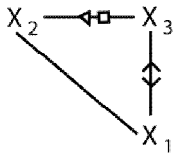
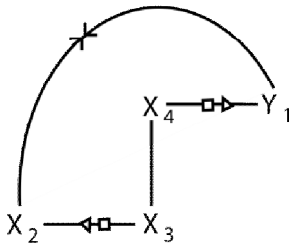
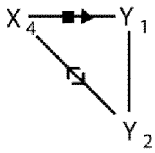
Sarcin/Ricin cycle sequences. The 3-D structure and shortest cycle basis of the sarcin-ricin motif are shown in **Figure 24**. The instances of each of the five individual cycles of the motif were searched in RNA-3A. In **Table 2**, we report for each cycle of the sarcin-ricin motif the number of instances, the number of sequences, and the highest RMSD between any instance and the seed motif. **Table 2** also shows the sequences of the base pairs shared

by two cycles and the RNA graphs of the most distant instances (if different from the seed motif).

We note no variation in the number of C_1 instances, 319, and sequences, 7, in presence or absence of the backbone. This suggests that the structural context of the non Watson-Crick tandem ❍□/□ limits the sequence variability beyond the specific and local geometry of the base pair. Among the 256 = 16 × 16 possible theoretical sequences, 120 = 10 (❍□) × 12 (□) sequences would be supported by isostericity matrices [3], whereas only 7 were observed in RNA-3A. Outside the context of the X_1□Y_3 base pair, such as in C_2, the X_3❍□Y_2 base pair accommodates more sequences, up to 15 in the absence of the backbone. Only 10 sequences for ❍□ base pairs are supported by isostericity matrices [3]. For C_2, there are 64 = 16 × 4 theoretical sequences, of which 34 were observed in RNA-3A in absence of the backbone, and only 5 in presence of the backbone.

In the case of C_2, we note a significant increase in the number of sequences in the absence of the backbone (34 compared to 5). The presence of the backbone constrains the sequence space, an observation that can be made in all cycles but C_1. The backbone interactions in C_1 are combined to base stacking, whereas in all other cycles we find lone backbone interactions. This indicates that base stacking affects the sequence space, but at a lower rate than the backbone.

Table 2. Sarcin-ricin cycle sequences. For each cycle, the number of instances found in RNA-3A, the number of different sequences, the RMSD between the most distant instance and the seed motif, the sequences of the base pairs shared by two cycles, and the RNA graph of the most distant instances are given. The RMSD are in Å.

Cycle	Backbone	Absent	Present	Sequences	RNA graph
C_1	#Instances	319	319	A❍□A	
	#Sequences	7	7	A❍□G	
	RMSD	2.7	2.7	G❍□G	
				U❍□A	
C_2	#Instances	1980	640	All but A❍□U	
	#Sequences	34	5		
	RMSD	7.1	1.7		
C_3	#Instances	2453	294	A◀■A	
	#Sequences	20	2	A◀■C	
	RMSD	6.9	1.8	C◀■A	
				C◀■C	
				G◀■A	
				G◀■G	
				A◀■U	
				U◀■C	
				U◀■G	
C_4	#Instances	755	327	A□▷N A◀■A	
	#Sequences	16	3	C□▷A A◀■C	
	RMSD	5.3	3.1	G□▷G C◀■A	
				U□▷G G◀■A	
				G◀■G	
				G◀■U	
				U◀■A	
C_5	#Instances	2453	1619	A□▷A	
	#Sequences	20	8	A□▷C	
	RMSD	6.7	3.4	A□▷G	
				C□▷A	
				C□▷C	
				C□▷U	
				G□▷G	
				G□▷U	
				U□▷G	

The high RMSD of the most distant C_2 instance is introduced by a flipping of the base in X_3. We observed that base flipping occurs in the absence of the backbone in all but the C_1 cycle, whereas base flipping occurs only in C_3 and C_4 in presence of the backbone. Consequently, the backbone restricts, but does not avoid, base flipping.

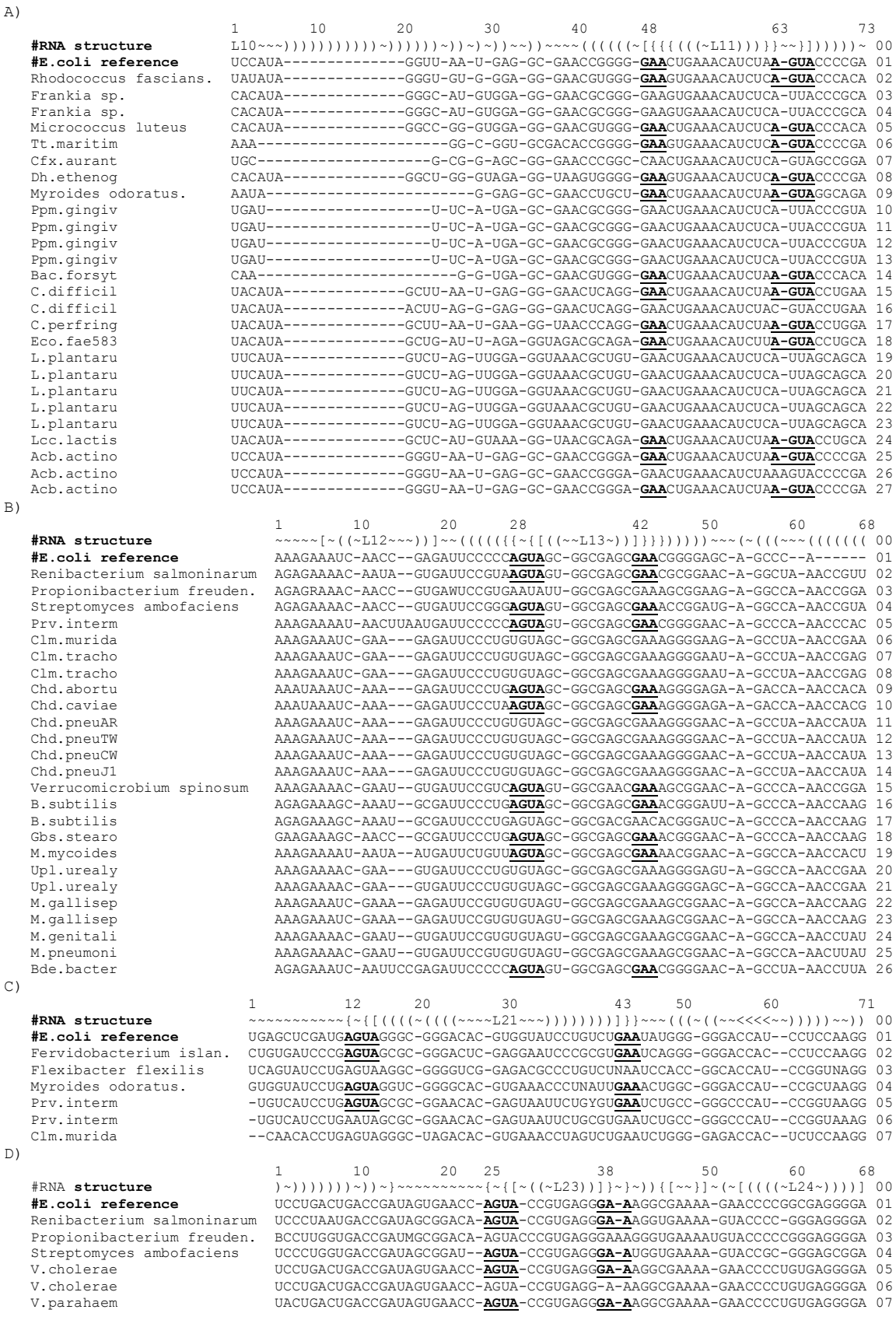
The sequence space of C_3 is highly constrained by the backbone and includes a rare base pair between two adjacent nucleotides in the sequence, here the X_2◀■X_3 base pair. This base pairing type can accommodate up to 14 sequences according to the isostericity matrices [3], but the specific context of C_3 allows for only one. The two possibilities for X_1 (A and G) brings to 2 the number of sequences.

Sarcin-ricin sequences. The application of the production rules of the sarcin-ricin graph-grammar result from the combination of its cycles, in any order, and their assignment of the cycle sequences. **Table 3** shows the sequences derivated by the graph-grammar. Four sarcin-ricin sequences are found in presence of the backbone: AGUA-AAA, AGUA-GAA, GGUA-AAA, and GGUA-GAA (in bold in **Table 3**), and 22 in absence of the backbone (see **Table 3**). The *lex parsimoniae* principle would favor the most simplified set of sequences, here the four sequences. Note, however, that sarcin-ricin sites made of three strands, with inserted nucleotides between X_1 and X_2, are found in RNA-3A (cf. G‘0’1071-G’0’1292-U’0’1293-A’0’1294 / G’0’911-A’0’912-A’0’913 in PDB entry 1JJ2 and other 23S, and A‘A’2302-G’A’953-U’A’954-A’A’955 / A’A’1012-A’A’1013-A’A’1014 in PDB entry 1K8A and other 50S). If we restrict the RMSD of each cycle to a maximum of 3 Å, we obtain the same four sequences (data not shown). If we remove the instances of the sarcin-ricin motifs from RNA-3A, the four sequences are still derived, showing that the cycles that make the sarcin-ricin motif appear elsewhere in RNA-3A and outside the context of the sarcin-ricin motif.

Table 3. Sarcin-ricin sequences. For present (bold) and absent (regular) backbone, the numbers and the sequences derived by the graph-grammar are listed.

Backbone	#Sequences	Sequences					
Absent	22	AAAA-AGA	AAAA-GGA	AAAU-GGA	AGAA-AGA	AGAA-GGA	AGAU-GGA
		AGGA-GGA	AGGA-GGC	AGGC-AGA	AGGC-AGC	**AGUA-AAA**	**AGUA-GAA**
		CAAA-AGG	CAAA-GGG	CAAU-GGG	GAAA-AAA	GAAA-GAA	GGAA-AAA
		GGAA-GAA	GGAG-GAA	**GGUA-AAA**	**GGUA-GAA**		
Present	4	**AGUA-AAA**	**AGUA-GAA**	**GGUA-AAA**	**GGUA-GAA**		

Sarcin/Ricin Alignment. **Figure 28** shows the sarcin-ricin sites in the alignement of the bacterial 23S rRNA subunit, as confirmed by the graph-grammar sequences (shown in bold and underlined in **Figure 28**). To preserve the global visibility of the alignment, for each underived sequence, the above and beyond sequences were chosen and put in a smaller alignment. For instance, **Figure 28A** shows the alignment of 27 sarcin-ricin sites in L11, which includes the 13 unsupported sites (among the 806 sequences). The supported sites are located at position (63-48), where the first nucleotide of the 4-nucleotide strand is at position 63 and the first nucleotide of the 3-nucleotide stand is at position 48.


A)
B)
C)
D)
#RNA structure
#E.coli reference

E)

```
                    1   5    10    16      17           30        40        50        60  64    70
#RNA structure      )(([[{~{[~((~~~~~ [...] }}~)))~)~~~))~~)))))))~~)}~)~))}}}}}}}}}}]}}}}}]}}]))~ 00
#E.coli reference   CAUAAGUA-ACGAUAA [...] AAUGGC-GUAAUGA-UGGCCAG-GCUGU-CUCCACCCGAGACUCAGUGAAAUUG 01
C.perfring          CAUCAGUA-GCGAGA- [...] AAUGGC-GUAAUGA-CUUGGGC-ACUGU-CUCAACUGUAAAUCCGGCGAAGUUG 02
C.perfring          CAUCAGUA-GCGAGA- [...] AAUGGCAGUAAUGA-CUUGGGC-ACUGUUCUCAACUGUAAAUCCGGC--AGUUG 03
C.perfring          CAUCAG-A-GCGAGA- [...] AAUGGC-GUAAUGA-CUUGGGC-ACUGU-CUCAACUGUAAAUCCGGCGAAGUUG 04
C.perfring          CAUCAGUA-GCGAGA- [...] AAUGGC-GUAAUGA-CUUGGGC-ACUGU-CUCAACUGUAAAUCCGGCGAAGUUG 05
Stp.aur832          UGUGAGUA-GCGAAA- [...] AAAGGC-GUAACGA-UUUGGGC-ACUGU-CUCAACGAGAGACUCGGUGAAAUCA 06
Stp.aur832          UGUGAGUC-GCGAAA- [...] AAAGGC-GUAACGA-UUUGGGC-ACUGU-CUCAACGAGAGACUCGGUGAAAUCA 07
Stp.epid62          UGUGAGUA-GCGAAA- [...] AAAGGC-GUAACGA-UUUGGGC-ACUGU-CUCAACGAGAGACUCGGUGAAAUCA 08
Lcc.lactis          UAUGAGUA-GCGCAA- [...] AAAGGC-GUAAUGA-UUUGGGC-ACUGU-CUCAACGAGAGACUCGGUGAAAUUU 09
L.gasseri           UAUGAGUA-GCGAAA- [...] N-NNNNNNN-NNNNNNN-NNNNN-NNNNNNNNNNNNNNNNNNNNNNNNNNNN 10
L.johnsoni          UAUGAGUA-GCGAAA- [...] AAAGGU-GUAAUGA-UUUGGGC-ACUGU-CUCAACGAGAGACUCGGUGAAAUUA 11
Atb.ferrox          CAUGAGUA-GCGAUAA [...] AAUGGC-GUAAUGA-UGGCCAC-ACUGU-CUCCACCCGAGACUCAGCGAAGUUG 12
Cox.burnet          CAUAAGUA-ACGAUAA [...] ACGA-UAGCCAC-GCUGU-CUCCACCCAAGACUCAGUGAAAUUGAAAUCGCUGU 13
Fnc.tulare          CAUGAGUA-ACGUAA- [...] AAUGGC-GUAACGA-UGGCCAC-ACUGU-CUCCACCAGAGACUCAGUGAAAUUG 14
```

Figure 28. Alignment of the bacterial 23S rRNA sequences. The alignment contains 806 sequences. The alignment was broken in 5 alignments: (A) to (E). Each smaller alignment was made of the sequences that were not derived by the graph-grammar and their above and below sequences. The parentheses represent canonical base pairs. The braces and brackets represent non-canonical base pairs. The tild characters, '~', are used for the unpaired nucleotides. Each sarcin-ricin site is assigned a loop number (cf. L11, L13, etc). The last site is span by too many nucelotides to be assigned a loop. Each sequence is given a unique number (on the right side). The source species of the sequences are indicated on the left. The sequence of *E. coli* is the structural reference, indicated by '#' character. The nucleotides shown in bold and underlined correspond to the sarcin-ricin sites that are derived by the graph-grammar. A) Site 'B'204-'B'205-'B'206-'B'207 / 'B'189-'B'190-'B'191 in PDB entry 2AWB, corresponding to position (63-48) in the alignment. B) 'B'241-'B'242-'B'243-'B'244 / 'B'254-'B'255-'B'256; (28-42). C) 'B'371-'B'372-'B'373-'B'374 / 'B'400-'B'401-'B'402; (12-43). D) 'B'457-'B'458-'B'459-'B'460 / 'B'469-'B'470-'B'471; (25-38). E) 'B'1265-'B'1266-'B'1267-'B'1268 / 'B'2012-'B'2013-'B'2014; (5-64).

For each unsupported site, we considered three possible hypotheses: i) a different tertiary structure (see Materials & Methods); ii) a sequencing error; and iii) a different alignment. All sarcin-ricin sites in the alignment are supported by isostericity matrices, as proposed by Westhof and coworkers, but for one exception in the last sarcin-ricin site of *Cox.burnet* (discussed below).

In **Figure 28A**, a different tertiary structure is suggested by *MC-Fold* (see Materials & Methods) for all 13 underived sequences. It aligns the sarcin-ricin site at position (62-49). The predicted tertiary structure and its cycles are shown in **Figure 29A**. An interesting feature of this structure is the presence of canonical Watson-Crick base pairs. We wanted to measure the distance of such a structure with the seed motif, and thus built a 3-D model using the computer program *MC-Sym* [21]. A superimposition of the model with the seed motif is shown in **Figure 29B**. The small RMSD (2.4 Å) between the model and the seed motif suggests a possible interpolation, a series of punctual mutations that could have transformed the motif during evolution. The characteristic inward X_2><Y_1 stacking of the seed sarcin-ricin motif is reproduced, but the nucleotides G and U do not allow for the formation of the typical X_2◀■X_3 base pair. However, one can see (**Figure 29B**) that the two nucleotides are positioned face-to-face and ready to form H-bonds, given the right mutations would align the appropriate donor-acceptor groups. We also built a 3-D model of the Weshtof's alignment hypothesis, which is shown in **Figure 29C**. As one can see, the 3-D model fits better the seed motif (RMSD of 0.9 Å), as the sequence is closer to that of the seed motif. In this case, the typical sarcin-ricin X_2◀■X_3 base pair is played by the U◀■U.

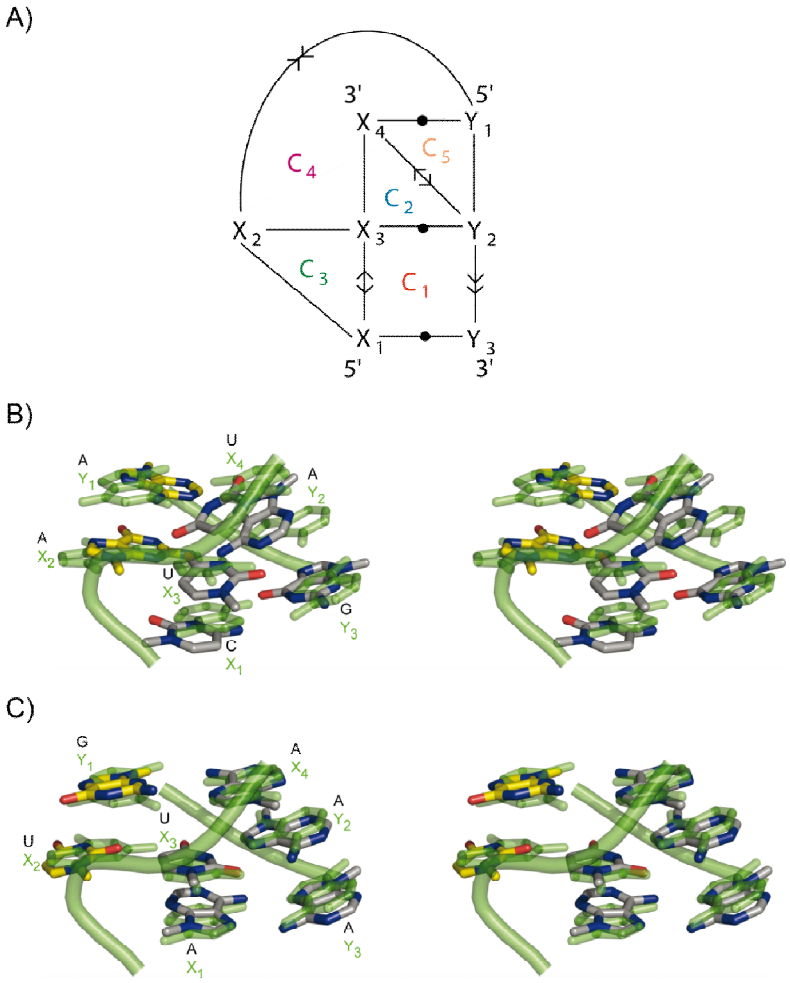


Figure 29. Unusual sarcin-ricin structure. A) Tertiary structure and cycles. The shortest cycle basis of the unusual sarcin-ricin structure shows five cycles, C_1 to C_5, characterized by canonical Watson-Crick base pairs. B) Stereoview of the *MC-Fold* model (2.4 Å RMSD). The model (colored) is superimposed on the seed motif (green). The nucleotides are labeled by the X_i (5'-strand) and Y_i (3'-strand). The backbone of the seed motif is shown using a light green cylinder. The backbone of the model is not shown. The nitrogen atoms in the model are in blue; oxygen in red; and carbon in gray. The carbon atoms shown in yellow emphasize the unconventional inward stacking between X_2 and Y_1, a characteristic feature of the sarcin-ricin motif. X_2 and X_3 do not pair. The hydrogen atoms are not shown. C) Stereoview of the alignment model (0.9 Å RMSD). The model (colored) is superimposed of on the seed motif (green). The color and numbering nomenclature is the same as in (B). X_2 and X_3, U and U, base pair as in the seed motif.

We noticed possible sequencing errors. In the sequence of *Cfx.aurant*, at position 48, a G would make more sense than a C. Even though the C is supported by isostericity matrices, a G is found in all other 805 sequences. In the *C.difficil* #16 sequence, the C at position 63 could have been the result of an insertion or of a sequencing error, rather than playing the role of X_1, which can be played by the preceding A if we assume an insertion. Again here, an A is found in all other 805 sequences. Finally, the graph-grammar signaled the inserted A at position 64 in the sequence of *Acb.actino* #26. Interestingly, the hypothesis of the insertion is equally sound as shifting the gap in all other sequences.

Figure 28B shows the alignment of the sarcin-ricin site in L13 at position (28-42). Among the 806 sequences, only 15 sequences are not derived by the graph-grammar. All but one sequence is also supported by tertiary structure prediction (*MC-Fold* data not shown). In the sequence of *B.subtilis* #17, *MC-Fold* predicts a sarcin-ricin motif at position (28-41), also compatible with isostericity matrices. This hypothesis creates a new gap, at position 40 for instance. Another possibility is a sequencing error at position 44, where the C is more likely to be an A. In this case, all approaches would support the alignment. Another possible sequencing error is at position 29 in the sequence of *Propionibacterium freuden*., where a G would fit better with all other 805 sequences. Finally, in the 13 unsupported sites with a U at position 29, a gap, to use the G at position 27, could be inserted. The U in this case would be seen as an insertion.

Figure 28C shows the sarcin-ricin site in L21 at position (12-43), where only two sequences are not derived by the graph-grammar. However, if the N in position 43 of the *Flexibacter flexilis* sequence is a G, then the graph-grammar can parse it and *MC-Fold* supports it as well. *MC-Fold* also supports the alignment of the *Prv.interm* #06 sequence. To be supported by the graph-grammar, the A at position 13 in this sequence of must be a G, as in all other 805 sequences.

Figure 28D shows the sarcin-ricin site in L23 at position (25-38), where only two sequences are not derived by the graph-grammar. *MC-Fold* supports a sarcin-ricin at the same position in the sequence of *Propionibacterium freuden*., but keeps the gap at position

40. In the sequence of *V.cholerae* #06, *MC-Fold* positions the sarcin-ricin at (25-37), but shifts the gap to position 37. The two above hypotheses are valid for the graph-grammar.

Figure 28E shows the last sarcin-ricin site at position (5-64), where 5 sequences are not derived by the graph-grammar. The four first underived sequences would be valid for the graph-grammar if sequencing errors are hypothesized: -- to GA in position 64 in *C.perfring* #03; - to U in position 7 in *C.perfring* #04; C to A in position 8 in Stp.aur832 #07; and NNN by GAA in position 64 in *L.gasseri*. For the fifth sequence, *Cox.burnet*, we noted that shifting to the right the sequence by 10 positions, starting at position 36, moves the GAA at position 54 to position 64, and then makes it valid for all approaches. Recall that isostericity matrices do not support the original alignment.

Conclusions

We were able to encode the tertiary structural features of the sarcin-ricin motif in a graph-grammar, and to employ it to derive valid sarcin-ricin sequences and to help align properly the sarcin-ricin sites in the bacterial sequences of 23S rRNA subunits. Producing and using the graph-grammar were both tractable, and the results obtained were sound and useful. The cycles allow us to build such a computational representation of RNA motifs.

We showed here that the cycles are independent RNA building blocks, as they are found in different contexts. We removed all occurrences of sarcin-ricin motifs from RNA-3A and were still deriving the valid sarcin-ricin sequences. Occurrences of the five sarcin-ricin cycles of its shortest cycle basis were found as individual elements in other motifs.

Deriving the sequences of an RNA motif can be thought of as a generalization of the isostericity base pair concept, but to a larger tertiary structure context that include all types of nucleotide interactions. For instance, we found that the valid sequences for the X_3❍□Y_2 base pair are not the same in two different contexts (C_1 and C_2). Other factors such as the backbone geometry have a role in the sequence space of an RNA motif, since removing it

increases the number of sequences that preserve all other interactions of the motif, and, thus, possibly its function.

Tertiary structure prediction and 3-D modeling, when combined to graph-grammars, are even more powerful tools to assess multiple hypotheses, and to make new ones. Tertiary structure predictions can be transformed in precise 3-D models, which, fed to a graph-grammar, can be used to derive new valid sequences.

In the current status of available structural and genomic data, the alignment protocol we propose can generate many valid hypotheses, which still require human intervention in the alignment process. It is not clear at this time that a graph-grammar search engine to identify RNA motifs in genomic data is necessary, as the use of currently available models combined to better alignments may improve greatly the rates of false positives and negatives. Or perhaps, the interplay of the nucleotide interactions that compose specific motifs reduces the sequence-structure signal to a point where identifying RNA families in genomic data is and will stay ticklish.

The definition of an RNA motif is in general vague, not always as precise as for the sarcin-ricin motif. For instance, many RNA families have stems that vary in length, and base pairing and stacking types that change from species to species. The effective graph-grammars of such motifs can be produced automatically, but a different graph-grammar is needed for each version of the motif. We can combine manually several graph-grammars to accommodate many motif versions. However, it would be more effective to consolidate many motif versions in a single graph-grammar automatically.

Acknowledgments

We would like to thank Martin Larose and Romain Rivière for helping us with some implementation details, Philippe Thibault for building the 3-D models with *MC-Sym*, and Marc Parisien for helpful discussions. This work was supported by grants from the Canadian Institutes of Health Research (CIHR) (MT-14604) and from the Natural Sciences

and Engineering Research Council of Canada (NSERC) (170165-01) to FM; NSERC (262965-06) to SH. SH holds a NSERC University Faculty Award. KS is supported by a scholarship from the Fonds Québécois de la Recherche sur la Nature et les Technologies. FM is a member of the Centre Robert-Cedergren.

Chapitre 5 RNA sequence design using a three-dimensional quantitative structure-activity relationships approach

5.1 Résumé

Nous avons développé une méthode pour déterminer les groupes chimiques impliqués pour une fonction donnée et prédire si une structure tertiaire peut remplir cette fonction. Nous avons appliqué notre méthode sur un ensemble de 12 séquences d'une instance d'un motif Sarcin-Ricin d'un ribosome bactérien, pour identifier les groupes chimiques. Nous avons appliqué notre méthode sur un nouvel ensemble de 8 séquences, que nous avons testées expérimentalement afin de montrer que nos prédictions sont précises.

L'article « RNA sequence design using a three-dimensional quantitative structure-activity relationships approach » dont les auteurs sont K. St-Onge, V. Lisi, S. Hamel, et F. Major, sera soumis pour publication au journal « Nature Structural and Molecular Biology ».

5.2 Partage du travail

Dans cet article, j'ai développé l'outil informatique *MC-QSAR*, un modèle pour quantifier la relation entre la structure et l'activité de l'ARN et je l'ai appliqué à un ensemble de séquences du Sarcin-Ricin sous la supervision de F. Major et S. Hamel. V. Lisi a effectué les expériences en laboratoire.

RNA sequence design using a three-dimensional quantitative structure-activity relationships approach

Karine St-Onge, Véronique Lisi, Sylvie Hamel and François Major[†]

Institute for Research in Immunology and Cancer,
Department of Computer Science and Operations Research
Université de Montréal
PO Box 6128, Downtown station
Montreal, Quebec H3C 3J7
CANADA

[†]To whom correspondence should be addressed.
François Major
Tel: 514 343 6752
Fax: 514 343 5839

Nature Structural and Molecular Biology
(En préparation)

Abstract

Ribonucleic acids are first-choice molecules to intervene in cellular programs since they both carry and control genetic information. Tens of RNA therapeutics are currently under clinical trials, and RNA-synthetic components that react to prefixed conditions and release therapeutics RNAs have been conceived and prototyped. However, engineering RNAs with predetermined structure and function requires profound understanding of RNA structure-activity relationships. Here, we introduce a principal component analysis approach, *MC-QSAR*, to discriminate RNA variants that preserve a given function. *MC-QSAR* builds the 3D structural profile of active sequences from a set of known active and inactive sequences. After the supervised learning step, the 3D structure of new sequences is matched to the profile: those fitting the profile are predicted to be active, while those that do not are predicted to be inactive. To exemplify *MC-QSAR*, we built a training set of twelve 23S ribosomal sarcin-ricin loop sequences, four known to be viable and eight to be lethal. Information about the key nucleotides of this loop was obtained by selecting the best parameters using leave-one-out cross validation. These nucleotides are either involved directly with the loop interactions with elongation factors or to maintain the necessary structural features in place for thereof. Besides, *MC-QSAR* was successful in predicting the outcomes of 23 out of 24 sarcin-ricin loops in transgenic bacteria.

Introduction

The unmatched flexibility of ribonucleic acids (RNAs) makes them molecules of choice for therapeutics [48]. RNA sequences fold in three-dimensional (3D) structures that confer precise biological function. The surface of RNA 3D structures offers a variety of chemical groups including several hydrogen donors and acceptors, the reactive 2'OH, and an electronegative phosphodiester chain. In tandem with proteins to carry and control genetic information, RNAs and proteins form numerous complexes.

RNA structures are dynamics [49]. The many possible conformations available to an RNA are determined by its sequence of nucleotides. These conformations determine function.

Therefore, any change in the sequence can alter its available structures and function. However, predicting to which extent any given mutation in the RNA sequence has on its function is difficult. To address this problem, we are taking a quantitative structure-activity relationships (QSAR) approach.

QSAR approaches are classically used to optimize the chemistry of ligands bound to receptors in the context of rational drug design. QSAR methods have been used to optimize protein primary [50] and topological structures [51][52], as well as topological descriptors of chemical structures [53]. QSAR methods have been applied to RNAs as well.

They have been used successfully to probe anticancer activity [34], to predict the local binding affinity constants between a specific nucleotide and an antibiotic [37] or to quickly identify and predict miRNAs [38]. The predictor variables used to model these RNA QSAR approaches go from 2D descriptors like chemical topology (planar representation of atoms and their chemical bonds) [34], to vectors of numerical values representing nucleotide properties (experimental molar absorption coefficients, single excitation energies, oscillation strength values, etc.) [37], or finally to thermodynamic molecular descriptors such as free energies, entropies and melting temperatures [38]. All those models were calibrated using linear discriminant analysis techniques and validated using either cross-validation techniques [34], leave-one-out jack-knife experiments [37] or ROC curve analysis [38].

Here, we introduce a QSAR method, *MC-QSAR*, to determine if an RNA sequence exhibits the same function as that measured experimentally on a training set of sequences. The 3D structures accessible to the training set sequences are predicted using high-resolution experimental structures, if any, and RNA 3D structure prediction. In comparison to traditional QSAR approaches, the goal here is not to optimize any profile, but rather to distinguish between among new sequences those that would be active (exhibit the function of the training sequences) and inactive (do not exhibit the function).

The training set can be composed of active and inactive sequences. Then, we use principal component analysis [30] (PCA) to build the electrostatic profile of active structures based on the commonly exposed charges of the 3D structures of the active sequences that are absent in the inactive sequences. The predicted structures of a new sequence for which we want to determine functionality are then mapped to this electrostatic profile. If the structure of the new sequence fits the profile, then it is determined to be active or inactive otherwise.

To illustrate and validate *MC-QSAR*, we use the sarcin-ricin loop (SRL), located in one of the longest conserved RNA sequences in large ribosomal subunits [54] (see **Figure 30A**). The SRL contains the classical GNRA tetraloop [55]. The name SRL is attributed to the fact that two protein toxins, α-sarcin and ricin, bind to it and inactivate protein synthesis by blocking the ribosome's interactions with elongation factors (EFs) [46][56]. The α-sarcin catalyzes the hydrolysis of the phosphodiester linkage between the R and A of the GNRA tetraloop [57], while the ricin favours the depurination of the N in the GNRA [58]. Since the SRL is highly conserved across ribosomes, it makes it an interesting case to study its structure-activity relationships.

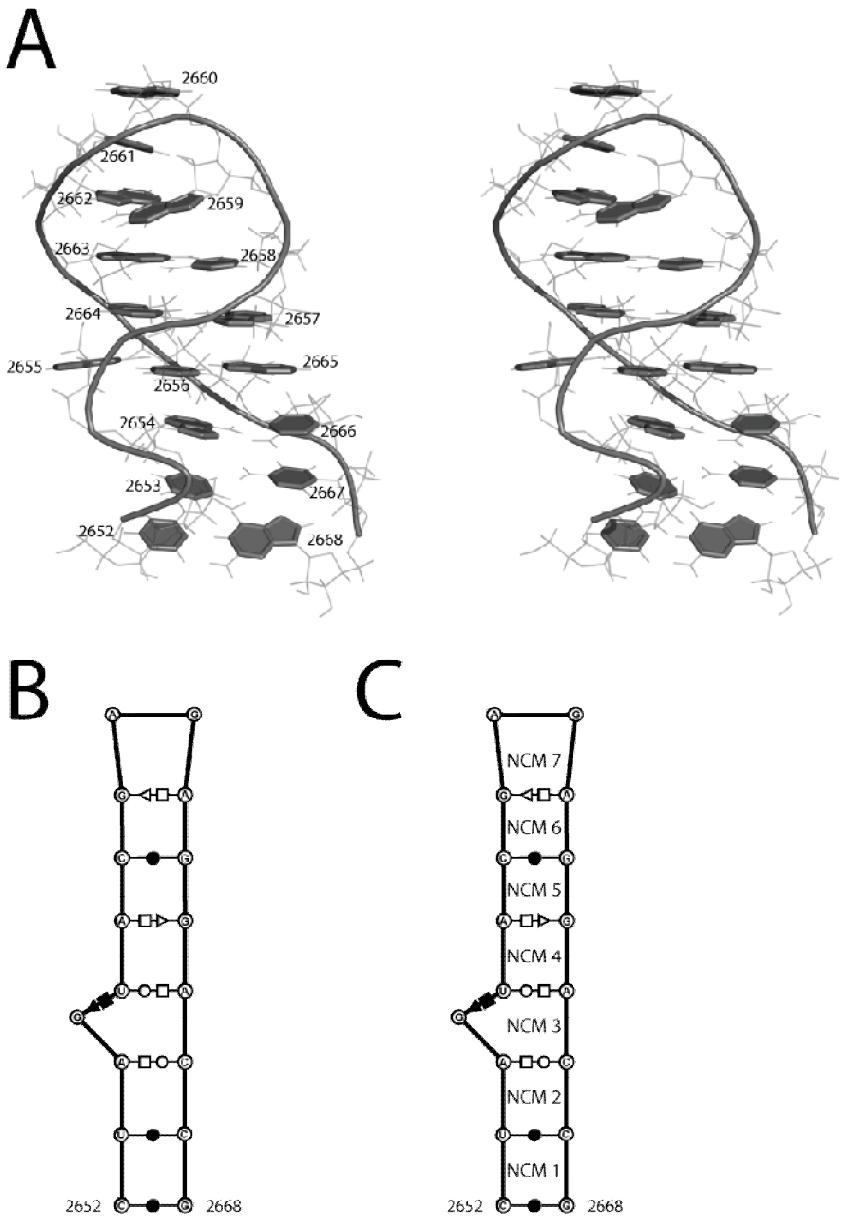


Figure 30. SRL. A) Stereo view of the SRL. Nucleotides are illustrated by planar forms and phosphodiester links by cylinders. B) Tertiary structure of the SRL in the 23S rRNA of *E. coli* (B2652-B2668) using *MC-Annotate* [22]. Canonical base pairs are represented with a black circle according to Leontis-Westhof notation [24], sugar edge is represented with a triangle and hoogsteen edge with a square. Filled symbol indicates that the base pair is in cis orientation and blank symbol in trans. Dark line represents phosphodiester link. C) NCMs are identified into the tertiary structure of the SRL. Same symbols are used as that in B).

To enhance the validation of *MC-QSAR*, we use two more examples: domain II tested for erythromycin resistance and P loop tested for cell growth. The domain II mutations that cause resistance to erythromycin are located in a hairpin structure between nucleotides 1198 and 1247 (see **Figure 31B**). This is close to a short open reading frame in the 23S rRNA that encodes a pentapeptide whose expression in vivo renders cells resistant to erythromycin. Therefore, a possible mechanism of resistance caused by domain II mutations may be related to an increased expression of the pentapeptide [59]. The P loop mutations that cause lethal mutants are located between nucleotides 2249 and 2254 (see **Figure 32B**). Evidence is present for the participation of the P loop of 23 S rRNA in establishing the tertiary structure of the peptidyl transferase center. Nucleotide substitutions were introduced into the P loop, which participates in peptide bond formation through direct interaction with the CCA end of P site-bound tRNA [60].

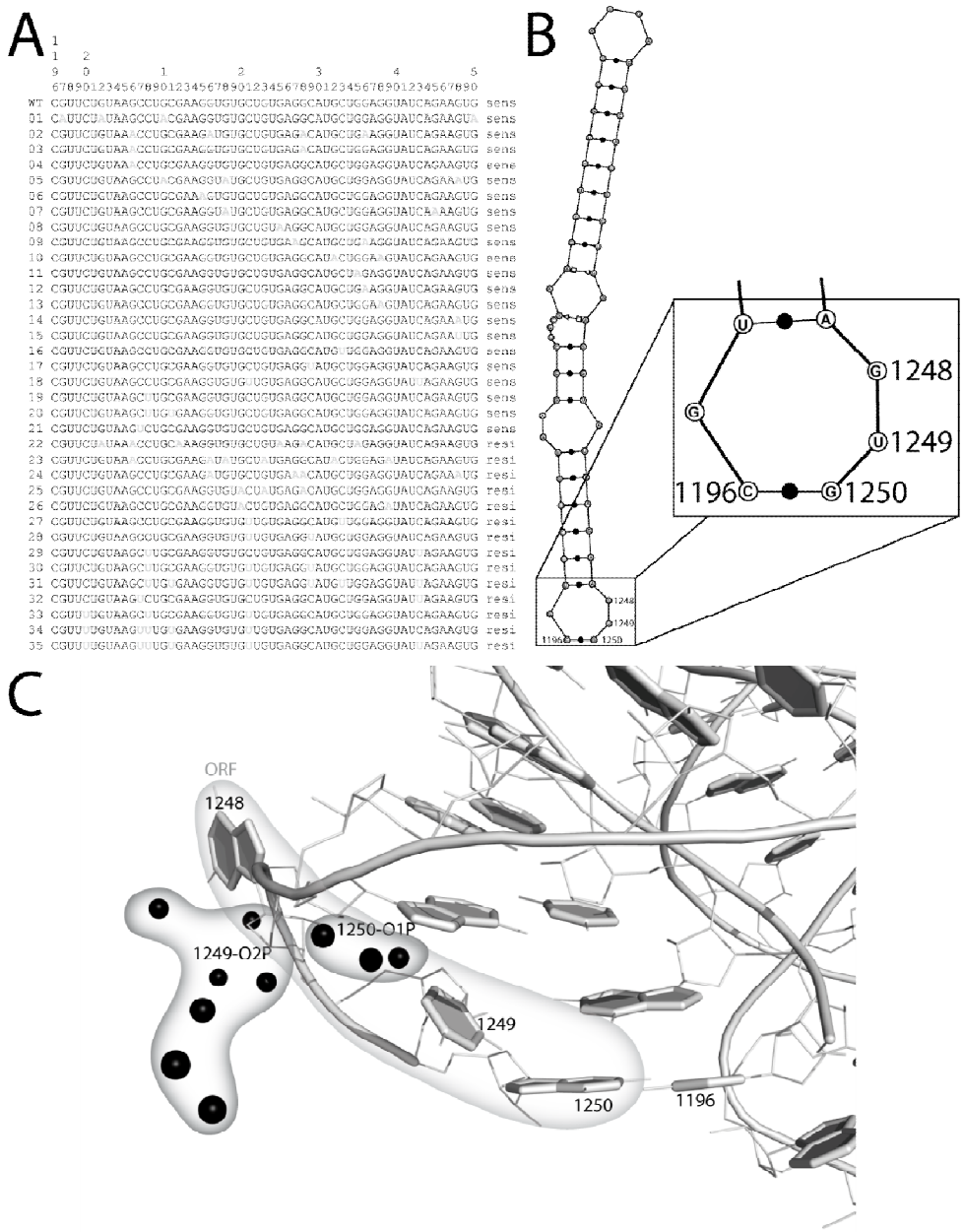


Figure 31. Domain II tested for erythromycin resistance. A) The thirty-six sequences from nucleotides 1196 to 1250 of the 23S *E. coli* that were tested for erythromycin resistance [59][61][62][63], mutations are indicated by bold/gray nucleotides. The numbering and wild-type (WT) sequence comes from the 23S *E. coli*. Notation sens/resi indicates that the sequence is sensitive or resistant to erythromycin. B) The tertiary structure using *MC-Annotate* [22]. A close-up of the most significant NCM from the PCA analysis is shown. The numbering is the same as in A). Canonical base pairs are

represented with a black circle according to Leontis-Westhof notation [24], sugar edges are represented with a triangle and hoogsteen edges with a square. Filled symbol indicates that the base pair is in cis orientation and blank symbol in trans. Dark line represents phosphodiester link. C) The 3D structure where the domain II is represented. Nucleotides are illustrated by planar forms and phosphodiester links by cylinders. The black areas represent significant atoms (indicated by spheres, where a bigger sphere is more significant) for discriminate sequences according to erythromycin resistance. The gray area indicates the short open reading frame (ORF) between nucleotides 1248 and 1250.

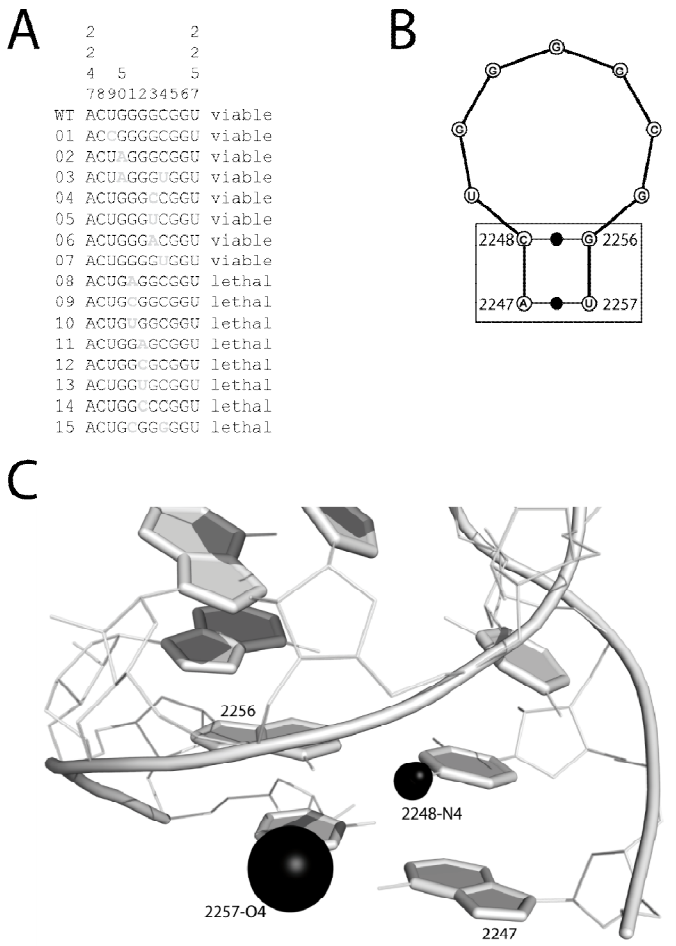


Figure 32. P loop tested for cell growth. A) The sixteen sequences from nucleotides 2247 to 2257 of the 23S *E. coli* that were tested for cell growth [24][60][64][65][66][67][68][69] [70], mutations are indicated by bold/gray nucleotides. The numbering and wild-type (WT) sequence comes from the 23S *E. coli*. Notation viable/lethal indicates that the sequence is viable or lethal (cell growth). B) The tertiary structure using *MC-Annotate* [22]. The most significant NCM from the PCA analysis is shown with a box. The numbering is the same as in A). Canonical base pairs are represented with black circle according to Leontis-Westhof notation [24], filled symbol indicates that the base pair is in cis orientation. Dark line represents phosphodiester link. C) The 3D structure where the hairpin is represented. Nucleotides are illustrated by planar forms and phosphodiester links by cylinders. Significant atoms are represented with spheres, where a bigger sphere is more significant for discriminating sequences according to cell growth.

Results

SRL predictions

Data set. **Table 4** shows the *MC-QSAR* activity predictions of the training set sequences. The four viable sequences and seven of the eight lethal sequences were correctly predicted. So we accurately predicted activity of 94% of the sequences from the training set. Moreover, we predict the activity of thirty new sequences (see **Table 5**): eight sequences that were in the alignment of the bacterial 23S rRNA subunit; eleven variants of the 2658-2663 base pair; and eleven randomly generated sequences that conserve dinucleotides composition. From the eight sequences that were in the alignment, we predict five sequences as viable and three as lethal. Then from the eleven variants of the 2658-2663 base pair, we predict three sequences as viable and eight as lethal. Finally from the eleven randomly generated sequences, we predict three sequences as viable and five as lethal. Note that no model is produced for three of the random sequences so the prediction is not possible.

Table 4. Predictions of the training set. Sequences are identified with the number (or WT) to their left in the first column. WT identifier is for the wild-type sequence, from the 23S rRNA of *E. coli* (B2652-B2668). The type of model for each sequence is indicated in the second column; MFE for the minimum free energy model and RMSD for the closest model in RMSD of the seed structure. The activities (viable/lethal for growth of cells [71][72][73][74]) and the activity predictions (viable/lethal) from LOOCV are shown for each sequence in third and fourth column.

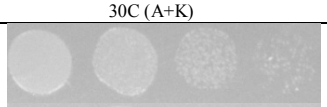
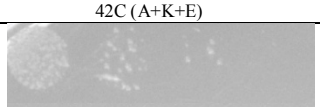
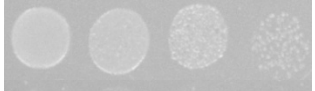
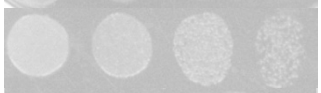
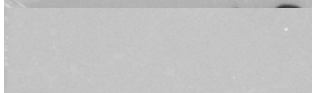
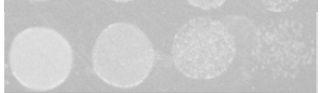
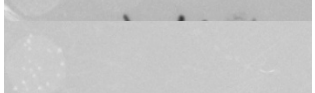
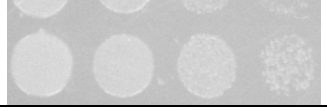
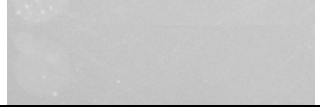
IDs	Model	Activity	Prediction
WT	MFE	viable	viable
WT	RMSD	viable	viable
01	MFE	lethal	lethal
02	MFE	viable	viable
02	RMSD	viable	viable
03	MFE	lethal	lethal
04	MFE	lethal	lethal
05	MFE	lethal	lethal
06	MFE	viable	viable
06	RMSD	viable	viable
07	MFE	lethal	lethal
08	MFE	lethal	viable
09	MFE	lethal	lethal
10	MFE	lethal	lethal
11	MFE	viable	viable
11	RMSD	viable	viable

Table 5. The data set. Sequences are identified with the number (or WT) to their left in the first column. WT identifier is for the wild-type sequence, from the 23S rRNA of *E. coli* (B2652-B2668). Mutations in sequences are shown in gray in the second column. Crosses are used to identify sequences that are in the alignment of bacterial 23S rRNA sequences in the third column. The words "viable" or "lethal" are used to identify sequences that are tested experimentally (growth cells) in 4 papers [71] [72] [73][74] in the fourth to seventh column. The activity predictions are indicated in the eighth column. Note that no model is produced for three of the random sequences so the prediction is not possible.

IDs	Sequences	Alignment	Chan *et al.*, JMB, 2000	Macbeth and Wool, JMB, 1999	Chan et al., JMB, 2006	Chan and Wool, JMB, 2008	Activity predictions
WT	CUAGUACGAGAGGACCG	×	viable	viable	viable	viable	viable
01	CUAGUAGGAGACGACCG		lethal			lethal	lethal
02	CUAGUAUGAGAAGACCG	×	viable				viable
03	CUAGUAAGAGAUGACCG		lethal				lethal
04	CUAGUACGAGACGACCG		lethal			lethal	lethal
05	CUACUACGAGAGGACCG			lethal		lethal	lethal
06	CUAAUACGAGAGGACCG	×		viable		viable	viable
07	CUAUUACGAGAGGACCG			lethal		lethal	lethal
08	CUAGUACGAGCGGACCG				lethal		viable
09	CUAGUACGAGGGGACCG				lethal		lethal
10	CUAGUACGAGUGGACCG				lethal		lethal
11	CUAGUACGUGAGGACCG					viable	viable
12	UUAGUACGAGAGGACCG	×					viable
13	CUAGUACGAGAGGACCA	×					viable
14	AUAGUACGAGAGGACCU	×					lethal
15	UUAGUACGCAAGGACCG	×					viable
16	CUUGUACGAGAGGACCG	×					viable
17	UUUGUACGAGAGGACCA	×					lethal
18	UUAGUACGAGAGGAUUU	×					lethal
19	CUAGUACGAGAGGCCCG	×					viable
20	CUAGUAAGAGAAGACCG						lethal
21	CUAGUAAGAGACGACCG						lethal
22	CUAGUAAGAGAGGACCG						lethal
23	CUAGUACGAGAAGACCG						lethal
24	CUAGUACGAGAUGACCG						lethal
25	CUAGUAGGAGAAGACCG						lethal
26	CUAGUAGGAGAGGACCG						lethal
27	CUAGUAGGAGAUGACCG						lethal
28	CUAGUAUGAGACGACCG						viable
29	CUAGUAUGAGAGGACCG						viable
30	CUAGUAUGAGAUGACCG						viable
31	UCAGUAGACCGGAGACG						viable
32	CUAGGACGAGAGUCACG						-
33	AGAGUCGACUAGGGACC						viable
34	CGAGUCACUAGGGAGAC						-
35	GGAGUAGACCACUCGGA						lethal
36	AGAGUACCUCGGAGACG						lethal
37	GGAGUACCGAGGACUCA						viable
38	CGAGUAGACGGGACUCA						-
39	GGUACUCGAGAGGACCA						lethal
40	UCGGGACGAGAGUACCA						lethal
41	AGGACUCGAGAGGUACC						lethal

Prediction set. From the thirty sequences from the data set, we selected eight sequences to be experimentally tested (growth of cells): three sequences that were in the alignment of the bacterial 23S rRNA subunit; three variants of the 2658-2663 base pair; and two randomly generated sequences that conserve dinucleotides composition. The activity predictions are shown in **Table 6**, where all of the eight tested sequences were correctly predicted.

Table 6. New sequences prediction. Sequences are identified with numbers (or WT) in the first column. WT identifier is for the wild-type sequence, from the 23S rRNA of *E. coli* (B2652-B2668). Mutations in sequences are shown in gray in the second column. The third and fourth columns show the effect on the growth of *E. coli* cells of mutations in a 23 S rRNA gene in a plasmid-encoded rrnB operon. (30C A+K and 42C A+K+E). The experimental activities and the activity predictions from *MC-QSAR* (viable/lethal) are shown for each sequence in the fifth and sixth column.

IDs	Sequences	Experimental		Activities	Predictions
		30C (A+K)	42C (A+K+E)		
WT	CUAGUACGAGAGGACCG			viable	viable
12	UUAGUACGAGAGGACCG			viable	viable
13	CUAGUACGAGAGGACCA			viable	viable
19	CUAGUACGAGAGGCCCG			viable	viable
26	CUAGUAGGAGAGGACCG			lethal	lethal
28	CUAGUAUGAGACGACCG			viable	viable
30	CUAGUAUGAGAUGACCG			viable	viable
36	AGAGUACCUCGGAGACG			lethal	lethal
40	UCGGGACGAGAGUACCA			lethal	lethal

PCA analysis. The most influent nucleotide cyclic motif (NCM) in SRL according to the LOOCV analysis is the NCM5 (see **Figure 30C**) composed of nucleotides 2657-2658, 2663-2664. From that NCM, we observe that the O6 atom from nucleotide 2664, N3 (2663), N4 (2658) and O2P (2657 and 2658) (see **Figure 33**) play a role in cell growth. However, the most significant atom is O6 from nucleotide 2664, means that a mutation that modifies the electrostatic area near this atom has more chance to change the activity of the sequence.

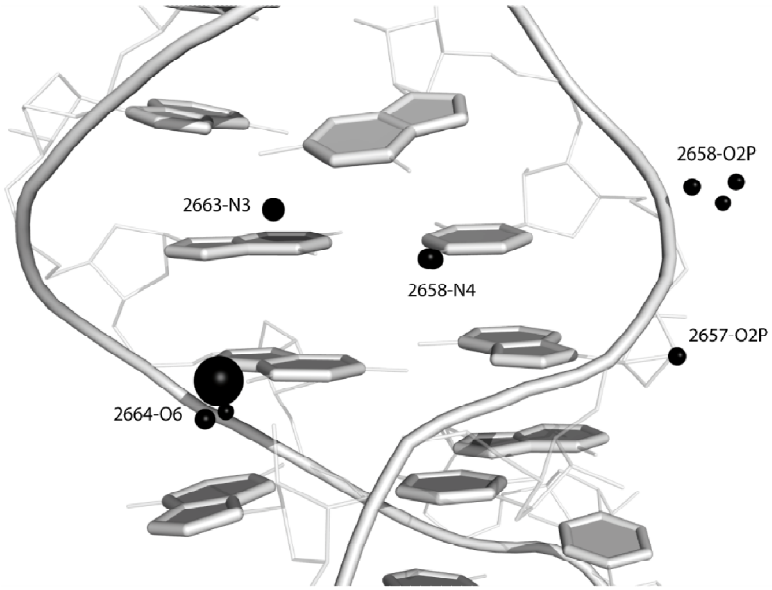


Figure 33. PCA analysis. The 3D representation of the SRL centered on the NCM 5 composed by nucleotides 2657, 2658, 2663 and 2664. Nucleotides are illustrated by planar forms and phosphodiester links by cylinders. The atoms that are the most important are represents with black spheres.

<u>*Other example predictions*</u>

Domain II tested for erythromycin resistance. The LOOCV analysis reveals that the NCM containing nucleotides 1196-1198 and 1248-1250 is the most significant for discriminate sequences according to erythromycin resistance. That analysis shows that the best number of clusters to use is twenty-seven. Finally, the LOOCV allows us to correctly predict thirty-

three of the forty-four sequences from the training set. Then, the PCA analysis indicates two significant atoms (O2P from nucleotide 1249 and O1P from 1250) that play a role in erythromycin resistance.

P loop tested for cell growth. The LOOCV analysis reveals that the NCM containing nucleotides 2247, 2248, 2256 and 2257 is the most significant for discriminating sequences according to cell growth. This analysis reveals that the best number of clusters is thirteen. In addition, the LOOCV enables us to accurately predict twelve of the sixteen sequences from the training set. Next, the PCA analysis indicates two significant atoms (N4 from nucleotide 2248 and O4 from 2257) that play a role in cell growth.

Discussion

SRL predictions

Using the *MC-Sym* software [21], we can now predict the 3D structure of RNA sequences more easily than ever before. Accurate predictions are essential to study the electrostatic features of RNAs. Using the PCA analysis, we can now study simultaneously the effect of a large number of features and capture those that explain a phenomenon (here cell growth of RNA sequences). The combined advances in RNA structure predictions and PCA analysis allow us to develop an RNA 3D QSAR method, *MC-QSAR*.

MC-QSAR allows us to identify the electrostatic profile of active structures based on the commonly exposed charges of the 3D structures of the active sequences that are absent in the inactive sequences, but also to distinguish between among new sequences those that would be active and inactive. We model new SRL sequences (see **Table 7**) and we determine functionality by mapping them to the electrostatic profile. We correctly predict twenty-three of the twenty-four experimentally-tested sequences, therefore 96% of predictions are correct.

Table 7. The activity predictions. Sequences are identified with the number (or WT) to their left in the first column. WT identifier is for the wild-type sequence, from the 23S rRNA of *E. coli* (B2652-B2668). Mutations in sequences are shown in gray in the second column. Crosses are used to identify sequences that are in the alignment of bacterial 23S rRNA sequences in the third column. The origin of NCM [20] is mentioned, in the fourth to tenth column, for each NCM and each sequence, where "blank" indicates that we use all occurrences of the NCM in the PDB (Protein Data Bank); "make" indicates that the NCM does not appear in the PDB and we build it using base pairing substitution into a backbone template from the PDB; "mut." indicates that the NCM does not appear in the PDB and it is not possible to construct it as described by the "make" and therefore we mutate the nucleobase identity to the specified sequence into the WT NCM; bold identifiers indicate that the 2655-2656 nucleotides are not paired; italic identifiers indicate that the mutation is made into other NCMs than the WT one; and underline identifiers indicate that base pair types are not the same as the WT one. The number of models for each sequence is shown, the minimum free energy [75] and the minimal distance from the seed structure in the eleventh to thirteenth column. The model used for each sequence is indicated in the fourteenth column; MFE for the minimum free energy model and RMSD for the closest model in RMSD of the seed structure. The activity (viable/lethal for growth of cells from [73][74][72][71] and our experiments) and the activity prediction (viable/lethal) are shown for each sequence in the fifteenth to sixteenth column.

IDs	Sequences	Align	NCM 1	NCM 2	NCM 3	NCM 4	NCM 5	NCM 6	NCM 7	Nb. models	Min. free energy	Distance from seed	Models	Activities	Predictions
WT	CUAGUACGAGAGGACCG	×								82	-154.244	0.292780	MFE	viable	viable
WT	CUAGUACGAGAGGACCG	×								82	-154.244	0.292780	RMSD	viable	viable
01	CUAGUAGGAGACGACCG									123	-153.915	0.545254	MFE	lethal	lethal
02	CUAGUAUGAGAAGACCG	×								38	-141.877	0.887899	MFE	viable	viable
02	CUAGUAUGAGAAGACCG	×								38	-141.877	0.887899	RMSD	viable	viable
03	CUAGUAAGAGAUGACCG									297	-148.853	0.399450	MFE	lethal	lethal
04	CUAGUACGAGACGACCG									4755	-137.172	1.043702	MFE	lethal	lethal
05	CUACUACGAGAGGACCG				**make**					2716	-137.959	0.853049	MFE	lethal	lethal
06	CUAAUACGAGAGGACCG	×			**make**					687	-147.103	0.720840	MFE	viable	viable
06	CUAAUACGAGAGGACCG	×			**make**					687	-147.103	0.720840	RMSD	viable	viable
07	CUAUUACGAGAGGACCG				**make**					1888	-134.788	0.839605	MFE	lethal	lethal
08	CUAGUACGAGCGGACCG								mut.	5462	-122.86	0.744745	MFE	lethal	viable
09	CUAGUACGAGGGGACCG								make	1230	-145.995	0.611596	MFE	lethal	lethal
10	CUAGUACGAGUGGACCG								make	328	-139.397	0.725250	MFE	lethal	lethal
11	CUAGUACGUGAGGACCG									36	-148.779	0.660903	MFE	viable	viable
11	CUAGUACGUGAGGACCG									36	-148.779	0.660903	RMSD	viable	viable
12	UUAGUACGAGAGGACCG	×								3296	-138.229	0.705416	MFE	viable	viable
13	CUAGUACGAGAGGACCA	×								3404	-139.714	1.039440	MFE	viable	viable
14	AUAGUACGAGAGGACCU	×								3316	-135.676	0.667155	MFE	-	lethal
15	UUAGUACGCAAGGACCG	×								6807	-142.586	0.756959	MFE	-	viable
16	CUUGUACGAGAGGACCG	×			**mut.**					328	1350.44	4.530229	MFE	-	viable
17	UUUGUACGAGAGGACCA	×		make	**mut.**					904	1618.42	4.692289	MFE	-	lethal
18	UUAGUACGAGAGGAUUU	×			**make**					10000	-137.25	1.134765	MFE	-	lethal
19	CUAGUACGAGAGGCCCG	×			**make**	make				10000	790.175	1.181008	MFE	viable	viable

20	CUAGUAAGAGAAGACCG				mut.	mut.		424	-93.1315	0.921480	MFE	-	lethal
21	CUAGUAAGAGACGACCG							54	-129.534	0.924141	MFE	-	lethal
22	CUAGUAAGAGAGGACCG				mut.	make		440	-91.5752	0.683969	MFE	-	lethal
23	CUAGUACGAGAAGACCG							3567	-138.342	0.799503	MFE	-	lethal
24	CUAGUACGAGAUGACCG							21	-106.055	1.017566	MFE	-	lethal
25	CUAGUAGGAGAAGACCG				make	mut.		46	-104.791	0.880023	MFE	-	lethal
26	CUAGUAGGAGAGGACCG							76	-123.753	0.887077	MFE	lethal	lethal
27	CUAGUAGGAGAUGACCG							168	-143.322	0.682374	MFE	-	lethal
28	CUAGUAUGAGACGACCG							39	-126.435	1.316499	MFE	viable	viable
29	CUAGUAUGAGAGGACCG							33	-147.811	0.470994	MFE	-	viable
30	CUAGUAUGAGAUGACCG							1349	-141.013	1.123861	MFE	viable	viable
31	UCAGUAGACCGGAGACG		**make**	make	*mut.*	*mut.*	*mut.*	586	2086.3	5.368651	MFE	-	viable
32	CUAGGACGAGAGUCACG		**make**	make				0	-	-	-	-	-
33	AGAGUCGACUAGGGACC		**make**	mut.	*mut.*	*mut.*	*mut.*	1806	800.374	4.652008	MFE	-	viable
34	CGAGUCACUAGGGAGAC		**make**	mut.	*mut.*	*mut.*	make	0	-	-	-	-	-
35	GGAGUAGACCACUCGGA		**make**	make			make	10302	-52.8482	3.597762	MFE	-	lethal
36	AGAGUACCUCGGAGACG		**make**	make		*mut.*	*mut.*	67321	1.005e+07	4.192982	MFE	lethal	lethal
37	GGAGUACCGAGGACUCA		**make**	make		*mut.*	*mut.*	1476	-27.0752	5.000734	MFE	-	viable
38	CGAGUAGACGGGACUCA		**make**	make	*mut.*	mut.	*mut.*	0	-	-	-	-	-
39	GGUACUCGAGAGGACCA	make	**mut.**	mut.				74650	-98.6177	1.928139	MFE	-	lethal
40	UCGGGACGAGAGUACCA	make	**make**	make				53679	-107.701	3.467435	MFE	lethal	lethal
41	AGGACUCGAGAGGUACC		**make**	mut.				71198	1.000e+08	6.110375	MFE	-	lethal

Beyond the activity predictions, we analyze the PCA analysis and we observe that the O6 atom from nucleotide 2664 plays a major role in cell growth (see **Figure 33**). According to that result, Spackova and Sponer identified an important ion-binding site, interconnecting the N7 atom from nucleotide 2263 and O6 from 2664 [76]. Moreover, the 2664 is within the site of ribosome-inactivating proteins [77].

Our PCA analysis also reveals that O2P atom from nucleotide 2657 among others play a role in cell growth (see **Figure 33**). Consistent with this result, Uchiumi and co-workers recognized that the 2657 nucleotide is strongly protected by the binding of L3 and L6 proteins with the RNA [78]. Moreover Spackova and Sponer found a hydration site in this area between N2 atom from nucleotide 2664 and O2P from 2657. In the crystal structure these atoms are in close contact, but during the equilibration period the distance between N2 (2664) and O2P (2657) increases and a hydration site is formed [76].

In another vein, for the clusters that participate the most in the PCA analysis (data not shown), we observe that two of them are positioned near to the N7 atom of the 2663 nucleotide and the C5 atom of the 2658 nucleotide as identified by Correll and co-workers. These functional groups, which are identical in the wild-type and the viable mutation but are different in the lethal mutation, are part of the putative EF binding surface [79].

Other example predictions

Domain II tested for erythromycin resistance. The PCA analysis indicates two significant atoms (O2P from nucleotide 1249 and O1P from 1250) that play a role in erythromycin resistance (see **Figure 31**). Agree with this result, Nteo mentions that translation of a pentapeptide (E-peptide) in cis, encoded in the rRNA has been reported to mediate erythromycin resistance in *Escherichia coli*. This E-peptide is encoded in a short open reading frame (ORF) between nucleotides 1248 and 1265 at the junctions of domain II and III. Mutations that affect translation initiation signals of the E-peptide mini gene (Shine-Dalgarno region and initiator codon GUG) abolish erythromycin resistance, suggesting that the size of the peptide and its amino acid sequence are essential for its functions [80].

P loop tested for cell growth. The PCA analysis indicates two significant atoms (N4 from nucleotide 2248 and O4 from 2257) that play a role in cell growth (see **Figure 32**). In correspondence with this result, Moazed and Noller among others identified that nucleotides 2256 and 2257 are located within the P-loop of domain V of the 23 S rRNA and are protected by the 3' terminus of the P-site-bound peptidyl tRNA [81][82][83][84].

Methods

SRL seed structure

The SRL tertiary structure of the *E. coli* 23S rRNA (PDB 2AWB B2652-B2668) is a typical SRL example, which we define as the seed structure (see **Figure 30A**). This structure is used as the template for a bacterial alignment of sequences in eight hundred and six species [Personal communication] (see **Table 8**). The sequences from the alignment that we use in this work have tested mutations [71][72][73][74] between nucleotides 2655 to 2663. From this region of mutations, we choose to include three additional base pairs to this sequence to complete the seed structure: C2652●G2668, U2653●C2667, and A2654□○C2666. With these additional base pairs, the structure will be more stable for the 3D structure prediction.

Table 8. The alignment sequences. Sequences are identified with the number (or WT) in the first column. WT identifier is for the wild-type sequence, from the 23S rRNA of *E. coli* (B2652-B2668). The alignment of bacterial 23S rRNA sequences from residue 2652 to 2668 represents the SRL in the second column. Mutations in sequences are shown in gray in the second column. The frequency of each sequence among the eight hundred and six sequences of the alignment is indicated in the third column.

IDs	Sequences	Frequency
WT	CUAGUACGAGAGGACCG	457
12	UUAGUACGAGAGGACCG	261
13	CUAGUACGAGAGGACCA	72
14	AUAGUACGAGAGGACCU	6
02	CUAGUAUGAGAAGACCG	2
15	UUAGUACGCAAGGACCG	1
16	CUUGUACGAGAGGACCG	1
17	UUUGUACGAGAGGACCA	1
18	UUAGUACGAGAGGAUUU	1
05	CUACUACGAGAGGACCG	1
19	CUAGUACGAGAGGCCCG	1
-	CUAGUACGANAGGACCG	1
-	CUAKUACGAGAGGACCG	1

The seed structure includes a base triple: G2655◄■U2656○□A2665 (see **Figure 30**) and is made of seven NCMs, numbered 1 to 7 in **Figure 30B**. NCMs were shown to be building blocks of RNA structures [2][20], and thus we use it to predict the 3D structure of SRL sequences.

Data set

The data set is made of forty-two sequences (see **Table 5** and **Table 7**). Twelve were taken from the literature (training set): the seed sequence (WT) and eleven mutants (01, 02, ..., 11) that were tested in bacteria cell growth [71][72][73][74]. Eight sequences (12, 13, ..., 19) were obtained from the bacterial alignment of 23S rRNAs. Eleven variants (20, 21, ..., 30) of the 2658-2663 base pair [73]. Finally, eleven random sequences (31, 32, ..., 41) that conserve the dinucleotide composition from the seed sequence.

Prediction set

For the prediction set, we choose sequences 12, 13 and 19 from the alignment: sequences 12 and 13 have many occurrences (two hundred and seventy-one and seventy-two

occurrences) while sequence 19 has only one occurrence. We also choose sequences 26, 28 and 30; they have mutations localized in the same base pair as sequences 01, 02, 03 and 04 [73]. We include sequences 36 and 40 that are random sequences that conserve the dinucleotide composition from the seed sequence.

QSAR method

The **Figure 34** illustrates how we determine the essential features of the biological function, here the viability and the growth of cells, from a set of sequences. To do this, we build a set of 3D models using *MC-Sym* for each sequence from the training set (see *Structure prediction* section). From this set of models, we divide each model into a set of NCMs (see *Split into NCMs* section). We clusterize atoms from each NCM to obtain a set of atom's clusters (see *Atoms clustering* section). Then we analyze the electrostatic features of each cluster considering accessible area and partial charge of atoms within those clusters (see *Features computation* section) and we encode this information in vectors. We use the Principal Component Analysis (PCA) [30] to convert the electrostatic feature's vectors of each model into uncorrelated variables called principal components (see *Principal Component Analysis* section). Those principal components are used to build the electrostatic profile of the training set. Finally, to determine the activity of a new sequence (see *Activity prediction* section), we analyze the electrostatic profile of this new sequence relative to that from training set sequences.

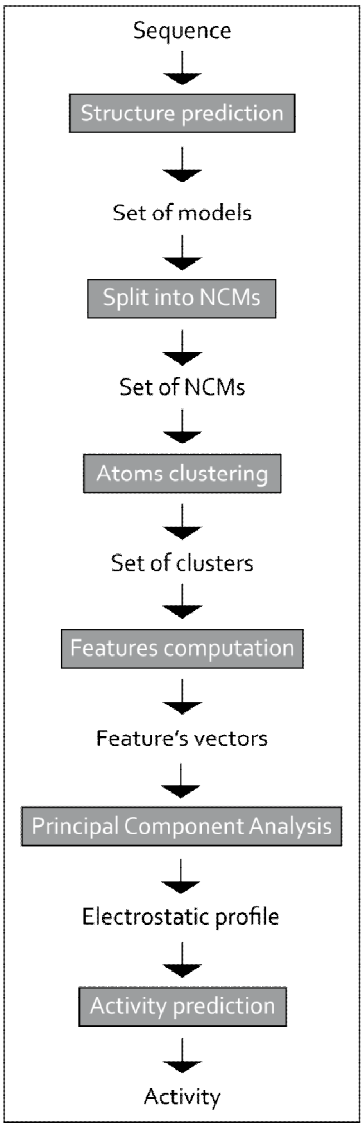


Figure 34. QSAR method. The QSAR method determines the essential features of the biological function from a set of sequences. To do this, we build a set of 3D models using a structure prediction program, *MC-Sym* [21] for each sequence. From this set of models, we split each model into a set of NCMs. We clusterize atoms from each NCM to obtain a set of the atom's clusters. Then we analyze the electrostatic features of each cluster and we encode this information in vectors. We use the PCA [30] to convert the electrostatic features vectors of each model into principal components and we build the electrostatic profile of the training set. Finally, to determine the activity of a new sequence, we analyze the electrostatic profile of this new sequence relative to that from training set sequences.

Structure prediction. For each sequence from the data set, we use *MC-Sym* to generate a set of 3D models that conserve base pair types from the seed structure (see **Table 7** and Supplementary Material for *MC-Sym* scripts). We refine each 3D models using energy minimization (TINKER limited memory L-BFGS [75]) and keep the minima. For the viable sequences from the training set (WT, 02, 06, and 11), we also keep the closest model in RMSD to the seed structure (see **Figure 35**), which gives us a set of eight viable and eight lethal models from the training set.

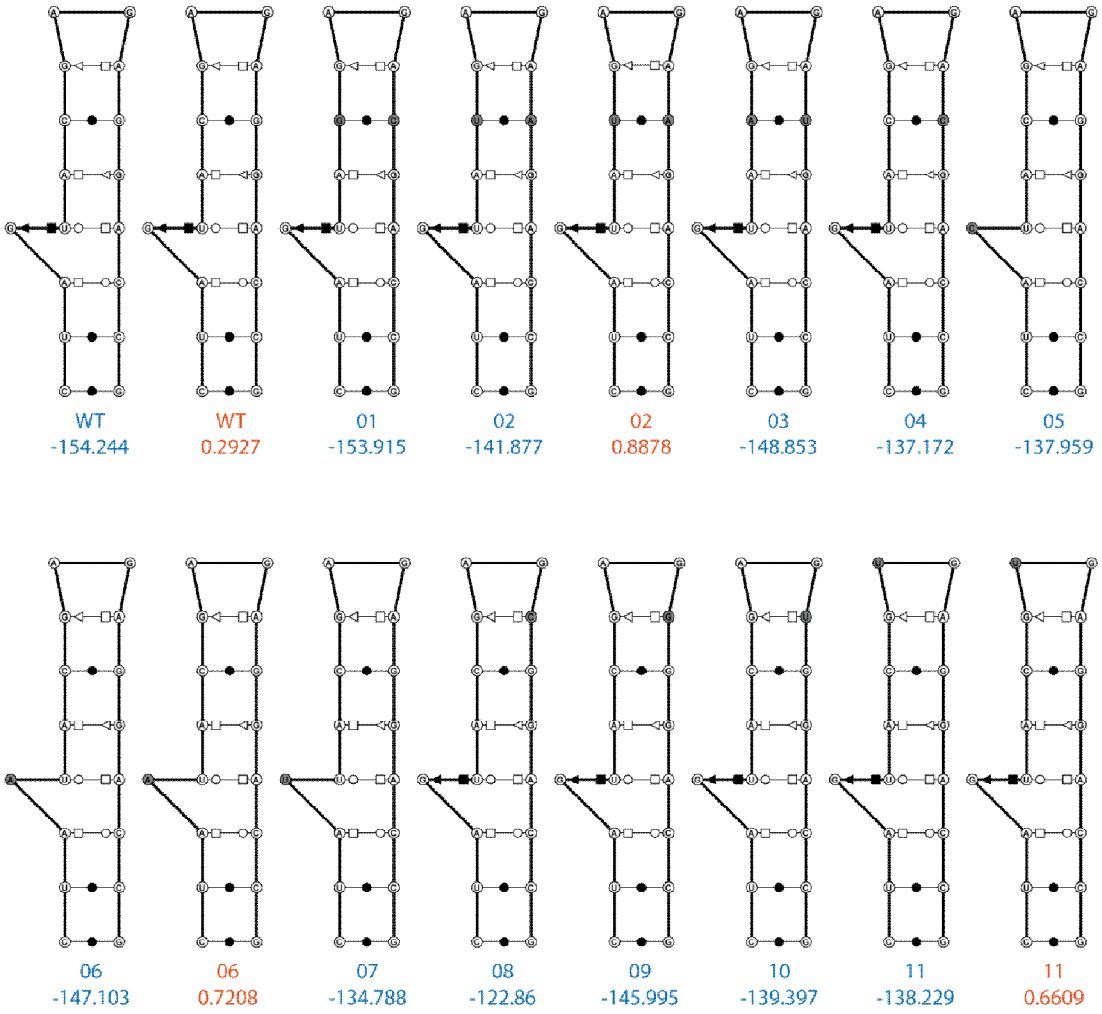


Figure 35. Training set models. The tertiary structure of the models used for the training set. Models are labelled, under tertiary structure, using IDs of sequences (WT, 01, …, 11). Under labels, the free energy in kcal/mole (blue) and the RMSD in Angstrom (orange) of each model are shown. The models are annotated using *MC-Annotate* [22]. Canonical base pairs are represented with a black circle according to Leontis-Westhof notation [24], sugar edge is represented with a triangle and hoogsteen edge with a square. Filled symbol indicates that the base pair is in cis orientation and blank symbol in trans. Dark line represents phosphodiester link. Nucleotide's gray background indicates mutations from WT sequence.

Split into NCMs. We align models together using *MC-RMSD* [22] on all atoms of the structures. We divide the whole structure analysis into NCM analyses as the divide and conquer strategy. Models are split in NCMs (see **Figure 30C**) where the first NCM (called NCM 1) is composed by nucleotides 2652-2653,2667-2668; the second NCM (called NCM 2) is composed by nucleotides 2653-2654,2666-2667; etc. At this step, we obtain a set of seven NCMs, where each NCM is composed of an alignment from the data set models.

Atoms clustering. For each NCM, we use the k-means algorithm [85] on atoms into the align models to cluster them (see **Figure 36A**). The idea of the k-means algorithm is to assign each atom into k clusters, where an atom is assigned to the cluster that minimizes the distance with the cluster centroid (center of mass, average of all atoms). In this study, we use k=13 because it is the k that maximizes the LOOCV analysis (see *Leave-One-Out analysi*s section). At this step, atoms (from data set's models) of each NCM are combined into a set of thirteen clusters.

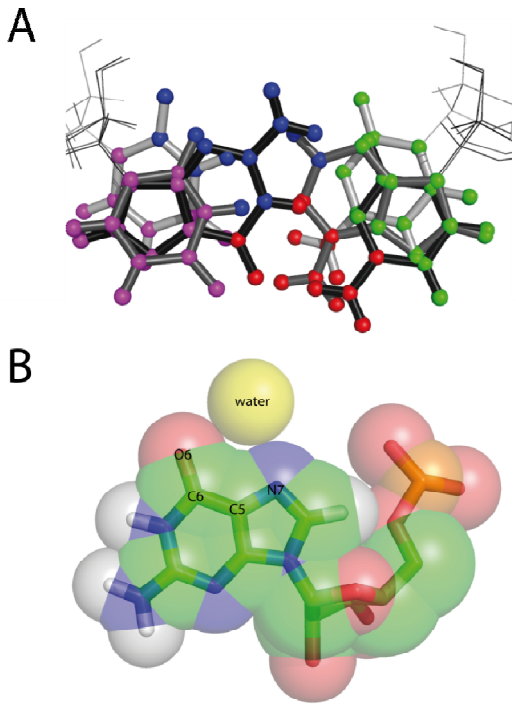


Figure 36 Examples. A) An example of an alignment of three base pairs (GC in black, UA in dark gray and UC in pale gray). Atoms used for the clustering are shown with spheres. The clustering atoms in 4 clusters are represented by each color (purple, blue, red and green). B) The accessible surface is calculated by a probe sphere (here a water molecule in yellow) as it rolls over the RNA (here a guanine (G) nucleotide represented by van der waals radius). In this example, the water molecule has access to the O6 and N7 atoms, but not to the C5 and C6 atoms. Nitrogen atoms are in blue; oxygen in red; carbon in green; phosphate in orange and hydrogen in white.

Features computation. We build a feature's vector for each model from the data set which contains the atom's electrostatic contributions of the thirteen clusters within a model. To do this for the cluster i within the model m for example, we sum the partial charge area (see **Eq. (3)**) of each atom within the cluster i for the model m, see **Eq. (2)**.

$$\boldsymbol{PartialCharge_{i,m}} = \sum_{j=1}^{\boldsymbol{NbAtoms_{i,m}}} \boldsymbol{PartialChargeArea_j} \qquad \textbf{(2)}$$

where $PartialCharge_{i,m}$ is the partial charge of the cluster i within the model m, $NbAtoms_{i,m}$ is the number of atoms that are members of the cluster i within the model m, $PartialChargeArea_j$ is the partial charge area of atom j (see **Eq. (3)**).

To calculate the partial charge area of an atom j (see **Eq. (3)**), we multiply its accessible area (see example in **Figure 36B**) obtained with the *pymol* program [86] (for a probe of water where its radius is 1.4A) with its partial charge (atomic partial charge parameters from Amber [87], see **Table 9**).

$$\boldsymbol{PartialChargeArea_j = PartialCharge_j \times Area_j} \qquad \textbf{(3)}$$

where $PartialChargeArea_j$ is the partial charge area of atom j, $PartialCharge_j$ is the partial charge of atom j, $Area_j$ is the accessible area of atom j.

At this step, we obtain a feature's vector of length thirteen (from the LOOCV analysis) for each model from the data set.

Table 9. Partial charges. Partial charge [87] for each atom (in row) in each type of nucleotide (in column).

Atoms/Nucleotides	A	C	G	U
C2	0.5875	0.7538	0.7657	0.4687
C4	0.3053	0.8185	0.1222	0.5952
C5	0.0515	0.5215	0.1744	-0.3635
C6	0.7009	0.0053	0.4770	-0.1126
C8	0.2006		0.1374	
N1	-0.7615	-0.0484	-0.4787	0.0418
N2			-0.9672	
N3	-0.6997	-0.7584	-0.6323	-0.3549
N4		-0.9530		
N6	-0.9019			
N7	-0.6073		-0.5709	
N9	-0.0251		0.0492	
O2		-0.6252		-0.5477
O4				-0.5761
O6			-0.5597	
H1			0.3424	
H2	0.04			
H3				0.3164
H5		0.1928		0.1811
H6		0.1928		0.2188
H8	0.1553		0.1640	
1H2			0.4364	

2H2			0.4364	
1H4		0.4234		
2H4		0.4234		
1H6	0.4115			
2H6	0.4115			
C1’	0.0394	0.0066	0.0191	0.0674
C2’	0.0670	0.0670	0.0670	0.0670
C3’	0.2022	0.2022	0.2022	0.2022
C4’	0.1065	0.1065	0.1065	0.1065
C5’	0.0558	0.0558	0.0558	0.0558
O2’	-0.6139	-0.6139	-0.6139	-0.6139
O3’	-0.5246	-0.5246	-0.5246	-0.5246
O4’	-0.3548	-0.3548	-0.3548	-0.3548
O5’	-0.4989	-0.4989	-0.4989	-0.4989
H1’	0.2007	0.2029	0.2006	0.1824
H2’	0.0972	0.0972	0.0972	0.0972
H3’	0.0615	0.0615	0.0615	0.0615
H4’	0.1174	0.1174	0.1174	0.1174
1H5’	0.0679	0.0679	0.0679	0.0679
2H5’	0.0679	0.0679	0.0679	0.0679
HO2’	0.4186	0.4186	0.4186	0.4186
P	1.1662	1.1662	1.1662	1.1662
O1P	-0.7760	-0.7760	-0.7760	-0.7760
O2P	-0.7760	-0.7760	-0.7760	-0.7760

Principal Component Analysis. For performing the machine learning step, we use feature’s vectors from the models part of the training set. We build the electrostatic profile of active structures based on the commonly exposed charges of the active sequences that are absent in the inactive sequences. To do this, we use PCA [30] to convert the feature’s vectors of each model into uncorrelated variables called principal components (see **Figure 37**). The first principal component is the one that represents the data more accurately (the highest variance), the second principal component is the second best representative of the data (the second highest variance), and so on. Models are now vectors of values called scores. In order to get an electrostatic profile as accurate as possible, we compute as many components as electrostatic features (here, resulting in thirteen principal components).

At this step, we obtain the electrostatic profile of models from the training set by performing the PCA on feature’s vectors thereof.

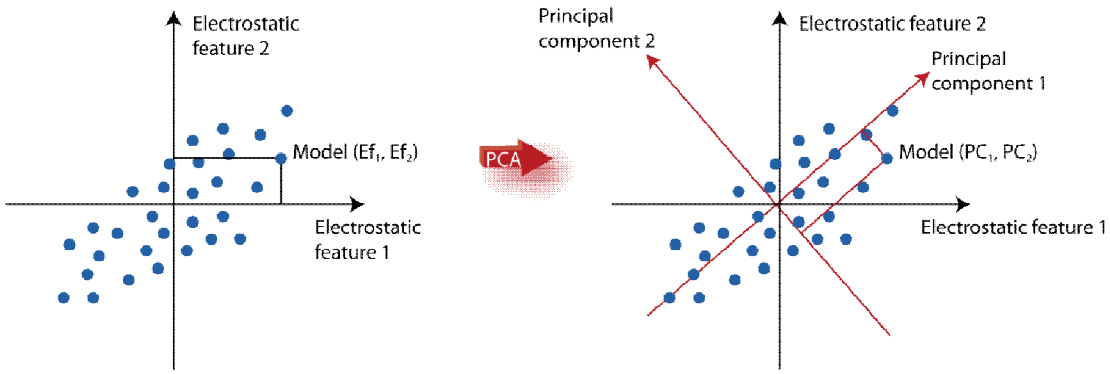


Figure 37. Principal component analysis. An example of PCA with two dimensions. Left) Models (blue dots) are represented in a graph where coordinates of each dot are coupled (Electrostatic feature 1, Electrostatic feature 2). Right) Same models as left. The first principal component represents the axis with the most variance. The second component is the orthogonal axis with the most variance to the first one. A model is the couple (Principal component 1, Principal component 2).

Activity prediction. To determine the activity of a new sequence, we analyze the electrostatic profile of this new sequence relative to that from training set sequences. To do this, we compute the distance, using weighted Euclidean distance as shown in **Eq. (4)**, between the electrostatic profile of the model from the new sequence and that of models from the training set.

$$\mathbf{Distance}_{i,j} = \sum_{k=1}^{NbComp} V_k\left(Score_{i,k} - Score_{j,k}\right)^2 \qquad (4)$$

where $Distance_{i,j}$ is the weighted Euclidean distance between the model i and the model j, $NbComp$ is the number of components in the PCA analysis (here, $NbComp$=13), V_k is the variance of component k, and $Score_{i,k}$ is the PCA score of the model i for the component k. We determine the activity of a new sequence as the same as the activity of the closest model from the training set of this model's new sequence, using the nearest neighbor algorithm.

Leave-One-Out analysis

To select the best set of parameters (NCM analysis and number of clusters used in the k-means algorithm) for our method *MC-QSAR*, we perform the Leave-One-Out Cross-Validation (LOOCV) technique [88] on electrostatic profiles from the training set. For each NCM analysis (see *Split into NCMs* section), we execute different LOOCV using one to fifteen clusters (see *Atoms clustering* section) for each NCM analysis. Among the one hundred and five LOOCV (result of the combination of seven NCMs and fifteen cluster sizes), we keep one that minimizes the prediction errors (see *Activity prediction* section). Here, one analysis produces one prediction error (see **Table 4**), the NCM analysis for NCM 5 (see **Figure 30C**) with thirteen clusters.

Other example predictions

Domain II tested for erythromycin resistance. We applied *MC-QSAR* on thirty-six sequences from nucleotides 1196 to 1250 of the 23S *E. coli* that were tested for erythromycin resistance [59][61][62][63]. Then, we applied LOOCV to identify the best NCM and the number of clusters to use.

P loop tested for cell growth. We applied *MC-QSAR* on sixteen sequences from nucleotides 2247 to 2257 of the 23S *E. coli* that were tested for cell growth [24][60][64][65][66][67][68][69][70]. Then, we apply LOOCV to identify the best NCM and the number of clusters to use.

Experimental methods.

Bacterial stains, plasmids and mutagenesis. *E. Coli* DH1 strain containing the thermolabile lcI repressor (referred to as DH1/cI) and the pLK45 plasmid were generously provided by the Wool lab. The pLK45 plasmid contains the *rrn*B operon under the control of the lP_L promoter, an ampicillin selection marker and a 23S mutation conferring erythromycin resistance (A2058G). An increase in temperature to 42°C induces expression of the pLK45 plasmid encoded rRNA. Mutageneses of the 23S ribosome were performed using the

QuickChange II XL mutagenesis kit from Agilents technologies according to the manufacturer's instructions.

Growth assay. DHI/cI cells with wild type or mutant 23S rRNA were grown in LB containing 50μg/mL ampicillin and 30μg/mL kanamycin at 30°C to an absorbance of 0.6 at 650nm, diluted (10^{-1} to 10^{-5}) and applied in 7μl drops on agar plates containing either 50μg/mL ampicillin and 30μg/mL kanamycin or 50μg/mL ampicillin, 50μg/mL kanamycin and 50μg/mL erythromycin. The plates were incubated at 30°C (amicillin and kanamycin) or 42°C (ampicillin and kanamycin – ampicillin, kanamycin and erythromycin) for 16-20 hours. Experiments were performed twice in duplicates and representative results are shown.

Acknowledgments

We would like to thank Paul Dallaire and Marc-Frédérick Blanchette for constructive discussions. We thank Eric Westhof for providing the alignment. This work was supported by grants from the Canadian Institutes of Health Research (CIHR) (MT-14604) to F.M., and from the Natural Sciences and Engineering Research Council of Canada (NSERC) (170165-01) to F.M. and (262965-2011) to S.H. S.H. holds a NSERC University Faculty Award. K.S. is supported by a scholarship from the Fonds Québécois de la Recherche sur la Nature et les Technologies.

Supplementary Material

NCM scripts

MC-Search script format

NCM 01

```
sequence (RNA A1 Seq.#1)
sequence (RNA A16 Seq.#2)
relation (
A1 A17 { W/W }
A2 A16 { pairing }
)
```

NCM 02

```
sequence (RNA A2 Seq.#1)
sequence (RNA A15 Seq.#2)
relation (
A2 A16 { pairing }
A3 A15 { H/W }
)
```

NCM 03 #1

```
sequence (RNA A3 AAU)
sequence (RNA A14 AC)
relation (
A3 A15 { H/W }
A4 A5  { S/H }
A5 A14 { W/H }
)
```

NCM 03 #2

```
sequence (RNA A3 AAU)
sequence (RNA A14 AC)
relation (
A3 A15 { H/W }
A5 A14 { W/H }
)
```

NCM 04

```
sequence (RNA A5 Seq.#1)
```

```
sequence (RNA A13 Seq.#2)
relation (
A5 A14 { W/H }
A6 A13 { H/S }
)
```

NCM 05 #1

```
sequence (RNA A6 Seq.#1)
sequence (RNA A12 Seq.#2)
relation (
A6 A13 { H/S }
A7 A12 { W/W }
)
```

NCM 05 #2

```
sequence (RNA A6 Seq.#1)
sequence (RNA A12 Seq.#2)
relation (
A6 A13 { pairing }
A7 A12 { pairing }
)
```

NCM 06 #1

```
sequence (RNA A7 Seq.#1)
sequence (RNA A11 Seq.#2)
relation (
A7 A12 { W/W }
A8 A11 { S/H }
)
```

NCM 06 #2

```
sequence (RNA A7 Seq.#1)
sequence (RNA A11 Seq.#2)
relation (
A7 A12 { pairing }
A8 A11 { pairing }
)
```

NCM 07 #1

```
sequence (RNA A8 Sequence)
relation (
A8 A11 { S/H }
)
```

NCM 07 #2

```
sequence (RNA A8 Sequence)
relation (
A8 A11 { pairing }
)
```

Make NCM script format

2_2

```
makeCycle.exe OUTmakeCycle.pdb 0.3 2_2 Sequence 2_2-NNNN.pdb Pb.#1.pdb Pb.#2.pdb >
Output.txt
```

3_2

```
makeCycle.exe OUTmakeCycle.pdb 0.3 3_2 Sequence 3_2-NNNNN.pdb Pb.#1.pdb Pb.#2.pdb >
Output.txt
```

4

```
makeCycle.exe OUTmakeCycle.pdb 0.3 4 Sequence 4-NNNN.pdb Pb.#1.pdb > Output.txt
```

Mutate NCM script format

NCM 03 #1

```
for file in MC-Sym/NCM-DB/NCM03AGUAC/*.pdb.gz MC-Sym/NCM-DB/NCM03ACUAC/*.pdb.gz MC-Sym/NCM-
DB/NCM03AAUAC/*.pdb.gz MC-Sym/NCM-DB/NCM03AUUAC/*.pdb.gz MC-Sym/NCM-DB/NCM03UGUAC/*.pdb.gz
MC-Seq/MC-Sym/NCM-DB/NCM03AGUAU/*.pdb.gz MC-Sym/NCM-DB/NCM03AGUCC/*.pdb.gz ; do molsep $file
; done
for file in *.pdb ; do echo $file ; makeSS.exe MC-Sym/NCM-DB/NCM03Sequence/${file} Sequence
${file} ; done
mcsearch MC-SearchScript MC-Sym/NCM-DB/NCM03Sequence/*model*.pdb
rm MC-Sym/NCM-DB/NCM03Sequence/*model*.pdb
```

NCM 03 #2

```
for file in MC-Sym/NCM-DB/NCM03AGUAC/*.pdb.gz ; do molsep $file ; done
for file in *.pdb ; do echo $file ; makeSS.exe MC-Sym/NCM-DB/NCM03Sequence/${file} Sequence
${file} ; done
```

NCM 04

```
for file in MC-Sym/NCM-DB/NCM04UAGA/*.pdb.gz ; do molsep $file ; done
for file in *.pdb ; do echo $file ; makeSS.exe MC-Sym/NCM-DB/NCM04Sequence/${file} Sequence
${file} ; done
```

NCM 05 #1

```
for file in MC-Sym/NCM-DB/NCM05AACG/*.pdb.gz MC-Sym/NCM-DB/NCM05AAUG/*.pdb.gz MC-Sym/NCM-
DB/NCM05ACAG/*.pdb.gz MC-Sym/NCM-DB/NCM05ACCG/*.pdb.gz MC-Sym/NCM-DB/NCM05ACGG/*.pdb.gz MC-
Sym/NCM-DB/NCM05ACUG/*.pdb.gz MC-Sym/NCM-DB/NCM05AGCG/*.pdb.gz MC-Sym/NCM-
```

```
DB/NCM05AGGG/*.pdb.gz MC-Sym/NCM-DB/NCM05AGUG/*.pdb.gz MC-Sym/NCM-DB/NCM05AUAG/*.pdb.gz MC-
Sym/NCM-DB/NCM05AUCG/*.pdb.gz MC-Sym/NCM-DB/NCM05AUGG/*.pdb.gz MC-Sym/NCM-
DB/NCM05AUUG/*.pdb.gz ; do molsep $file ; done
for file in *.pdb ; do echo $file ; makeSS.exe MC-Sym/NCM-DB/NCM05Sequence/${file} Sequence
${file} ; done
mcsearch MC-SearchScript MC-Sym/NCM-DB/NCM05Sequence/*model*.pdb
rm MC-Sym/NCM-DB/NCM05Sequence/*model*.pdb
```

NCM 05 #2

```
for file in MC-Sym/NCM-DB/NCM05ACGG/*.pdb.gz ; do molsep $file ; done
for file in *.pdb ; do echo $file ; makeSS.exe MC-Sym/NCM-DB/NCM05Sequence/${file} Sequence
${file} ; done
```

NCM 06 #1

```
for file in MC-Sym/NCM-DB/NCM06AGAC/*.pdb.gz MC-Sym/NCM-DB/NCM06AGAU/*.pdb.gz MC-Sym/NCM-
DB/NCM06CGAA/*.pdb.gz MC-Sym/NCM-DB/NCM06CGAC/*.pdb.gz MC-Sym/NCM-DB/NCM06CGAG/*.pdb.gz MC-
Sym/NCM-DB/NCM06CGAU/*.pdb.gz MC-Sym/NCM-DB/NCM06CGCG/*.pdb.gz MC-Sym/NCM-
DB/NCM06CGGG/*.pdb.gz MC-Sym/NCM-DB/NCM06CGUG/*.pdb.gz MC-Sym/NCM-DB/NCM06GGAC/*.pdb.gz MC-
Sym/NCM-DB/NCM06GGAG/*.pdb.gz MC-Sym/NCM-DB/NCM06GGAU/*.pdb.gz MC-Sym/NCM-
DB/NCM06UAAG/*.pdb.gz MC-Sym/NCM-DB/NCM06UGAA/*.pdb.gz MC-Sym/NCM-DB/NCM06UGAC/*.pdb.gz MC-
Sym/NCM-DB/NCM06UGAG/*.pdb.gz /MC-Sym/NCM-DB/NCM06UGAU/*.pdb.gz ; do molsep $file ; done
for file in *.pdb ; do echo $file ; makeSS.exe MC-Sym/NCM-DB/NCM06Sequence/${file} Sequence
${file} ; done
mcsearch MC-SearchScript MC-Sym/NCM-DB/NCM06Sequence/*model*.pdb
rm MC-Sym/NCM-DB/NCM06Sequence/*model*.pdb
```

NCM 06 #2

```
cd MC-Sym/NCM-DB/NCM06Sequence
for file in *.pdb.gz ; do molsep $file ; done
for file in *.pdb ; do echo $file ; makeSS.exe MC-Sym/NCM-DB/NCM06Sequence/${file} Sequence
${file} ; done
```

NCM 07 #1

```
for file in MC-Sym/NCM-DB/NCM07GAGA/*.pdb.gz MC-Sym/NCM-DB/NCM07GAGC/*.pdb.gz MC-Sym/NCM-
DB/NCM07GAGG/*.pdb.gz MC-Sym/NCM-DB/NCM07GAGU/*.pdb.gz MC-Sym/NCM-DB/NCM07GCAA/*.pdb.gz MC-
Sym/NCM-DB/NCM07GUGA/*.pdb.gz ; do molsep $file ; done
for file in *.pdb ; do echo $file ; makeSS.exe MC-Sym/NCM-DB/NCM07Sequence/${file} Sequence
${file} ; done
mcsearch MC-SearchScript MC-Sym/NCM-DB/NCM07Sequence/*model*.pdb
rm MC-Sym/NCM-DB/NCM07Sequence/*model*.pdb
```

NCM 07 #2

```
for file in MC-Sym/NCM-DB/NCM07GAGA/*.pdb.gz ; do molsep $file ; done
```

```
for file in *.pdb ; do echo $file ; makeSS.exe MC-Sym/NCM-DB/NCM07Sequence/${file} Sequence
${file} ; done
```

Used scripts table

NCM	Sequence	Scripts			Seq.#1	Seq.#2	Pb.#1	Pb.#2
		MC-Search	Make NCM	Mutate NCM				
01	AGCC	X			AG	CC		
01	AGCG	X			AG	CG		
01	AUCU	X			AU	CU		
01	CGAC	X			CG	AC		
01	CGCA	X			CG	CA		
01	CUCA	X			CU	CA		
01	CUCG	X			CU	CG		
01	GGCA	X			GG	CA		
01	GGGA	X			GG	GA		
01	UCCA	X			UC	CA		
01	UCCG	X			UC	CG		
01	UUCA	X			UU	CA		
01	UUCG	X			UU	CG		
01	UUUU	X			UU	UU		
02	CAAC	X			CA	AC		
02	CGCC	X	2_2		CG	CC	CC	GC
02	GAAC	X			GA	AC		
02	GAGA	X			GA	GA		
02	GAGG	X			GA	GG		
02	GAUC	X			GA	UC		
02	GGAC	X			GG	AC		
02	GUCC	X	2_2		GU	CC	GC	UC
02	UAAC	X			UA	AC		
02	UACC	X			UA	CC		
02	UAUU	X			UA	UU		
02	UUCC	X	2_2		UU	CC	UC	UC
03	AAUAC	#2	3_2		AAU	AC	AC	UA
03	ACUAC	#2	3_2		ACU	AC	AC	UA
03	AGGCA	#2	3_2		AGG	CA	AA	GC
03	AGUAC	#1			AGU	AC		
03	AGUAG	#2	3_2		AGU	AG	AG	UA
03	AGUAU	#2	3_2		AGU	AU	AU	UA
03	AGUCC	#2	3_2		AGU	CC	AC	UC
03	AGUCG	#2	3_2		AGU	CG	AG	UC
03	AGUCU	#2	3_2		AGU	CU	AU	UC
03	AGUGA	#2	3_2		AGU	GA	AA	UG
03	AUUAC	#2	3_2		AUU	AC	AC	UA
03	GACUA	#2		#1	GAC	UA		

03	GGGAC	#2	3_2		GGG	AC	GC	GA
03	UACAC	#2		#2	UAC	AC		
03	UGUAC	#2		#2	UGU	AC		
04	CUGA	X		X	CU	GA		
04	GUGU	X		X	GU	GU		
04	GAUA	X	2_2		GA	UA	GA	AU
04	GAUC	X	2_2		GA	UC	GC	AU
04	UAAC	X	2_2		UA	AC	UC	AA
04	UAAG	X	2_2		UA	AG	UG	AA
04	UAGA	X			UA	GA		
04	UAGC	X	2_2		UA	GC	UC	AG
04	UAUC	X	2_2		UA	UC	UC	AU
04	UCGA	X		X	UC	GA		
04	UCGG	X		X	UC	GG		
05	AAAG	#1		#2	AA	AG		
05	AACG	#1			AA	CG		
05	AAGG	#1		#2	AA	GG		
05	AAUG	#1			AA	UG		
05	ACAG	#1			AC	AG		
05	ACCG	#1			AC	CG		
05	ACGA	#1			AC	GA		
05	ACGG	#1			AC	GG		
05	ACGU	#1			AC	GU		
05	ACUG	#1			AC	UG		
05	AGAG	#1	2_2		AG	AG	AG	GA
05	AGCG	#1			AG	CG		
05	AGCU	#1			AG	CU		
05	AGGA	#1		#2	AG	GA		
05	AGGG	#1			AG	GG		
05	AGUG	#1			AG	UG		
05	AUAG	#1			AU	AG		
05	AUCG	#1			AU	CG		
05	AUGG	#1			AU	GG		
05	AUUG	#1			AU	UG		
05	CAGG	#1		#1	CA	GG		
05	CGGG	#2		#1	CG	GG		
05	UCGG	#1			UC	GG		
06	ACGG	#2		#1	AC	GG		
06	AGAA	#1		#2	AG	AA		
06	AGAC	#1			AG	AC		
06	AGAG	#1	2_2		AG	AG	AG	GA
06	AGAU	#1			AG	AU		
06	CCGG	#2		#1	CC	GG		
06	CGAA	#1			CG	AA		
06	CGAC	#1			CG	AC		
06	CGAG	#1			CG	AG		

06	CGAU	#1			CG	AU	
06	CGCG	#1			CG	CG	
06	CGGG	#1			CG	GG	
06	CGUG	#1			CG	UG	
06	GAAC	#1			GA	AC	
06	GAAG	#1		#1	GA	AG	
06	GAGG	#2		#1	GA	GG	
06	GGAA	#1		#2	GG	AA	
06	GGAC	#1			GG	AC	
06	GGAG	#1			GG	AG	
06	GGAU	#1			GG	AU	
06	UAAG	#1			UA	AG	
06	UGAA	#1			UG	AA	
06	UGAC	#1			UG	AC	
06	UGAG	#1			UG	AG	
06	UGAU	#1			UG	AU	
07	ACCA	#1	4		AC	CA	AA
07	ACCG	#2		#1	AC	CG	
07	ACGG	#2		#1	AC	GG	
07	ACUA	#1		#1	AC	UA	
07	CGAG	#2		#1	CG	AG	
07	CUAG	#1		#1	CU	AG	
07	CUCG	#2		#1	CU	CG	
07	GAGA	#1			GA	GA	
07	GAGC	#1		#2	GA	GC	
07	GAGG	#1	4		GA	GG	GG
07	GAGU	#1	4		GA	GU	GU
07	GCAA	#1			GC	AA	
07	GUGA	#1			GU	GA	

MC-Sym scripts

Scripts are generated *via* http://www.major.iric.ca/MC-Pipeline/ with default parameters except these

Sequence	Fragment RMSD	Merge RMSD	Clash threshold	Backtrack limit	Bond threshold	Model diversity	Model limit
WT	0.1	0.05	1.5	50%/50%	3.0	0.1	Not used
01	0.1	0.06	1.5	50%/50%	3.0	0.1	Not used
02	0.1	0.07	1.5	50%/50%	3.0	0.1	Not used
03	0.1	0.06	1.5	50%/50%	3.0	0.1	Not used
04	0.1	0.955	1.5	50%/50%	3.0	0.5	Not used
05	0.1	0.353	1.5	50%/50%	3.0	1.0	Not used
06	0.1	0.3	1.5	50%/50%	3.0	1.0	Not used
07	0.1	0.352	1.5	50%/50%	3.0	1.0	Not used
08	0.1	0.15	1.5	50%/50%	3.0	0.1	Not used
09	0.1	0.23	1.5	50%/50%	3.0	0.5	Not used

10	0.1	0.52	1.5	50%/50%	3.0	1.0	Not used
11	0.1	0.05	1.5	50%/50%	3.0	0.1	Not used
12	0.1	0.72	1.5	50%/50%	3.0	1.0	Not used
13	0.1	1.06	1.5	50%/50%	3.0	1.0	Not used
14	0.1	0.704	1.5	50%/50%	3.0	1.0	Not used
15	0.1	0.72	1.5	50%/50%	3.0	1.0	Not used
16	1.0	1.155	0.01	50%/50%	4.5	1.5	Not used
17	1.0	1.2	0.01	100%/100%	4.5	1.0	Not used
18	0.0001	5000.0	Not used	100%/100%	Not used	0.001	Not used
19	1.0	1.0	0.01	100%/100%	4.3	1.0	Not used
20	0.1	0.2	1.5	50%/50%	3.0	0.1	Not used
21	0.1	0.82	1.5	50%/50%	3.0	1.0	Not used
22	0.1	0.42	1.5	50%/50%	3.0	0.5	Not used
23	0.1	0.83	1.5	50%/50%	3.0	0.5	Not used
24	1.0	5.0	0.01	100%/100%	4.5	0.1	Not used
25	0.1	0.34	1.5	50%/50%	3.0	0.5	Not used
26	0.1	0.6	1.5	50%/50%	3.0	1.0	Not used
27	0.1	0.25	1.5	50%/50%	3.0	0.5	Not used
28	0.1	0.77	1.5	50%/50%	3.0	1.0	Not used
29	0.1	0.2	1.5	50%/50%	3.0	0.5	Not used
30	0.1	0.22	1.5	50%/50%	3.0	0.1	Not used
31	0.1	2.0	0.01	50%/50%	4.5	0.1	Not used
32	0.1	100.0	Not used	100%/100%	Not used	0.1	Not used
33	0.1	4.046	0.01	50%/50%	4.5	0.1	Not used
34	0.1	100.0	Not used	100%/100%	Not used	0.1	Not used
35	0.1	2.551	0.0001	100%/100%	10.0	0.1	Not used
36	0.1	4.053	0.01	50%/50%	4.0	0.1	Not used
37	0.1	4.0	1.0	50%/50%	3.0	0.1	Not used
38	0.1	100.0	Not used	100%/100%	Not used	0.1	Not used
39	0.1	1.394	0.5	50%/50%	3.0	0.1	Not used
40	0.1	0.411	1.5	50%/50%	3.5	0.1	Not used
41	0.1	2.8	0.5	50%/50%	3.0	0.1	Not used
01-00	default	0.3	1.5	25%/33%	2.0	1.0	10000
01-01	default	0.3	1.5	25%/33%	2.0	1.0	10000
01-02	default	0.3	1.5	25%/33%	2.0	1.0	10000
01-03	default	0.2	1.5	25%/33%	2.0	1.0	10000
01-04	default	0.3	1.5	25%/33%	2.0	1.0	10000
01-05	default	0.3	1.5	25%/33%	2.0	1.0	10000
01-06	default	0.32	1.5	25%/33%	2.0	1.0	10000
01-08	default	0.45	1.5	25%/33%	2.0	1.0	10000
01-09	default	2.0	1.5	25%/33%	2.0	1.0	10000
01-10	default	0.46	1.5	25%/33%	2.0	1.0	10000
26-00	default	0.5	1.0	25%/33%	2.0	1.0	10000
26-01	default	0.5	1.0	25%/33%	2.0	1.0	10000
26-02	default	0.5	1.0	25%/33%	2.0	1.0	10000
26-03	default	0.48	1.0	25%/33%	2.0	1.0	10000

26-11	default	0.55	1.0	25%/33%	2.0	1.0	10000
26-15	default	0.55	1.0	25%/33%	2.0	1.0	10000
26-19	default	0.82	1.0	25%/33%	2.0	1.0	10000
26-20	default	0.8735	1.0	25%/33%	2.0	1.0	10000
26-24	default	1.3	1.0	25%/33%	2.0	1.0	10000
26-25	default	0.78	1.0	25%/33%	2.0	1.0	10000
30-00	default	0.28	1.0	25%/33%	2.0	1.0	10000
30-01	default	0.318	1.0	25%/33%	2.0	1.0	10000
30-02	default	0.317	1.0	25%/33%	2.0	1.0	10000
30-03	default	0.24	1.0	25%/33%	2.0	1.0	10000
30-04	default	0.652	1.0	25%/33%	2.0	1.0	10000
30-05	default	0.45	1.0	25%/33%	2.0	1.0	10000
30-06	default	0.6	1.0	25%/33%	2.0	1.0	10000
30-07	default	1.15475	1.0	25%/33%	2.0	1.0	10000
30-09	default	0.35	1.0	25%/33%	2.0	1.0	10000
30-11	default	0.415	1.0	25%/33%	2.0	1.0	10000
36-00	default	0.55	1.0	25%/33%	2.0	1.0	10000
36-01	default	0.42	1.0	25%/33%	2.0	1.0	10000
36-02	default	0.43	1.0	25%/33%	2.0	1.0	10000
36-03	default	0.67	1.0	25%/33%	2.0	1.0	10000
36-05	default	0.33	1.0	25%/33%	2.0	1.0	10000
36-06	default	0.3	1.5	25%/33%	2.0	1.0	10000
36-07	default	0.88	1.0	25%/33%	2.0	1.0	10000
36-08	default	0.23	1.5	25%/33%	2.0	1.0	10000
36-09	default	0.82	1.0	25%/33%	2.0	1.0	10000
36-10	default	1.0	1.0	5%/10%	2.0	1.0	10000
40-00	default	0.46	1.5	25%/33%	2.0	1.0	10000
40-01	default	0.37	1.5	25%/33%	2.0	1.0	10000
40-02	default	0.35	1.5	25%/33%	2.0	1.0	10000
40-08	default	2.7	1.5	25%/33%	2.0	1.0	10000
40-09	default	0.48	1.5	25%/33%	2.0	1.0	10000
40-10	default	0.35	1.5	25%/33%	2.0	1.0	10000
40-11	default	0.39	1.5	25%/33%	2.0	1.0	10000
40-13	default	0.46	1.5	25%/33%	2.0	1.0	10000
40-14	default	0.23	1.5	25%/33%	2.0	1.0	10000
40-21	default	0.44	1.5	25%/33%	2.0	1.0	10000

Discussion et Conclusion

Le premier objectif de cette thèse était d'utiliser les grammaires de graphes pour développer un outil informatique permettant de prédire l'appartenance d'une séquence à une famille d'ARN. Une des applications est d'identifier les séquences qui en font partie et ainsi déterminer un ensemble de séquences qui se replient dans une même structure. Toutefois, les grammaires de graphes ne tiennent pas compte des interactions des groupes chimiques spécifiques avec des éléments extra-moléculaires, comme d'autres macromolécules ou ligands. Pour tenir compte de ces interactions, le deuxième objectif de cette thèse était de développer un modèle qui tient compte de la position des groupes chimiques à la surface des structures tertiaires pour prédire l'activité d'une séquence. L'hypothèse étant qu'un ensemble de groupes chimiques à des positions conservées dans des séquences actives déplacés dans des séquences inactives sont probablement impliqués dans des interactions qui sont nécessaires à leur fonction.

La réalisation de ces objectifs à débuté avec le développement d'une grammaire de graphes afin de modéliser la structure tertiaire d'une famille d'ARN. Nous avons d'abord modélisé la boucle E du ribosome qui contient le motif Sarcin-Ricin et identifié un ensemble de quatre séquences qui s'y replient. La grammaire est constituée des cycles d'interactions comme objet de premier ordre. Nous avons démontré que ces cycles sont des éléments indépendants de construction d'ARN. Nous avons retiré les instances du motif Sarcin-Ricin de la base de données de structures (« Jackknife » [89]) et nous avons dérivé le même ensemble de séquences. Nous avons confirmé la pertinence biologique de cet ensemble de séquences par une comparaison des séquences avec un alignement de plus de 800 séquences ribosomiques bactériennes (Communication personnelle, Westhof).

Dans un deuxième temps et toujours avec l'exemple de la boucle E, nous avons ensuite cherché les groupes chimiques de cette boucle qui pourraient être impliqués dans

des interactions avec des facteurs d'élongation. Une fois les groupes identifiés, nous avons prédit par modélisation tridimensionnelle les séquences qui positionnent correctement ces groupes dans leurs structures tertiaires. Nous avons confirmé le pouvoir prédictif de ce modèle à l'aide d'un ensemble de 24 variants dont la viabilité à été vérifiée expérimentalement dans notre laboratoire. De ces variants, 23 d'entre eux ont été correctement prédits par notre modèle.

À l'aide de notre grammaire de graphe, nous avons étudié les séquences et les structures tertiaires de chaque cycle composant la structure du Sarcin-Ricin. Cette étude a révélé que l'espace des séquences dépend grandement des interactions entre tous les nucléotides reliés par des interactions de types paire de bases et empilement dans la structure tertiaire, c'est-à-dire pas uniquement entre deux paires de bases adjacentes. Cela suggère l'importance du contexte pour la relation entre la séquence et la structure. On peut en conclure qu'une grammaire de graphes contextuelle qui se base sur les motifs cycliques de nucléotides plutôt que la structure secondaire classique modélise mieux la relation séquence-structure des ARN.

Pour ce qui est d'identifier des séquences fonctionnelles d'un motif, les méthodes existantes utilisent une représentation trop simplifiée de la structure d'ARN, ce qui limite leur pouvoir prédictif. La force de notre modèle est en grande partie due à la précision de la modélisation tridimensionnelle des structures développée dans notre laboratoire.

Dans le cadre de cette thèse, nous avons développé une grammaire de graphes basée sur les cycles d'interactions. Il est cependant possible de remanier la grammaire de graphes afin d'utiliser d'autres types de sous-graphes cycliques d'interactions, comme les NCM par exemple. Les types de sous-graphes cycliques utilisés découperont le graphe d'interactions de manière différente. Le découpage du motif aura un impact sur l'ensemble des séquences générées. Par exemple, un découpage en NCM peut produire plus de séquences qu'un

découpage en cycles parce que les cycles sont plus contraignants que les NCM. En ce sens, les cycles considèrent l'interaction d'empilement alors que les NCM non.

À plus grande échelle, il est également possible d'utiliser une grammaire de graphes basée sur des motifs prédéfinis, comme les définitions qu'on trouve dans la littérature, par exemple. Les séquences générées par cette grammaire représenteraient un agencement de motifs, ce qui peut être utile pour étudier une grande molécule d'ARN, un ARNt par exemple. De plus, cette grammaire pourrait être récursive, par exemple générer des ensembles de séquences pour les motifs, qui à leur tour serviraient à générer de plus longues séquences.

Nous pourrions aussi utiliser différemment notre modèle qui identifie des séquences fonctionnelles d'un motif en se basant sur la position des groupes chimiques. Contrairement à identifier les propriétés qui dissocient un ensemble de séquences actives d'un ensemble de séquences inactives, nous pourrions nous en servir pour identifier les propriétés identiques à un ensemble de séquences donné. En effet, à l'inverse de sélectionner les propriétés les plus pondérées, c'est-à-dire celles qui discriminent un ensemble de séquences, nous pourrions choisir les propriétés les moins pondérées, c'est-à-dire celles qui sont les plus communes à un ensemble de séquences. Cette application de notre modèle pourrait être employée sur les familles de Rfam afin de mieux les comprendre et les caractériser.

Finalement, nous avons atteint les objectifs de cette thèse, c'est-a-dire que le développement des grammaires de graphes nous a permis d'identifier des séquences qui font partie d'une famille d'ARN et nous sommes en mesure d'identifier des séquences fonctionnelles d'un motif. Cet aboutissement a été possible grâce à la particularité du motif Sarcin-Ricin, toutefois les grammaires de graphes autant que notre modèle d'identification des groupes chimiques ont leurs limites. Étant donné que les grammaires de graphes sont construites à partir des cycles d'interactions observés dans les structures existantes, alors générer des séquences pour un motif constitué de nouveaux cycles d'interactions (c'est-à-dire non observés auparavant) devient impossible ou laborieux. Ceci veut dire qu'il faut

prédéterminer différemment un ensemble de séquences représentant chaque nouveau cycle d'interactions. Pour ce qui est de notre modèle qui identifie des séquences fonctionnelles d'un motif en se basant sur la position des groupes chimiques, sa limite est la modélisation tridimensionnelle. En effet, le pouvoir prédictif du modèle est directement proportionnel à la précision de la modélisation. Prédire des séquences actives pour une grande molécule d'ARN ou pour un motif dont la structure est inconnue s'avère donc plus difficile.

Bibliographie

[1] R. R. Gutell, J. C. Lee, and J. J. Cannone, "The accuracy of ribosomal RNA comparative structure models," *Curr. Opin. Struct. Biol.*, vol. 12, pp. 301-310, 2002.

[2] S. Lemieux and F. Major, "Automated extraction and classification of RNA tertiary structure cyclic motifs," *Nucleic Acids Res.*, vol. 34, pp. 2340-2346, 2006.

[3] A. Lescoute, N. B. Leontis, C. Massire, and E. Westhof, "Recurrent structural RNA motifs, isostericity matrices and sequence alignments," *Nucleic Acids Res.*, vol. 33, pp. 2395-2409, 2005.

[4] J. H. Cate, et al., "RNA tertiary structure mediation by adenosine platforms," *Science*, vol. 273, pp. 1696-1699, 1996.

[5] A. Chworos, et al., "Building programmable jigsaw puzzles with RNA," *Science*, vol. 306, pp. 2068-2072, 2004.

[6] S. R. Holbrook, "RNA structure: the long and the short of it," *Curr. Opin. Struct. Biol.*, vol. 15, pp. 302-308, 2005.

[7] N. B. Leontis, A. Lescoute, and E. Westhof, "The building blocks and motifs of RNA architecture.," *Curr. Opin. Struct. Biol.*, vol. 16, pp. 279-287, 2006.

[8] S. Pasquali, H. H. Gan, and T. Schlick, "Modular RNA architecture revealed by computational analysis of existing pseudoknots and ribosomal RNAs," *Nucleic Acids Res.*, vol. 33, pp. 1384-1398, 2005.

[9] Y. Sakakibara, et al., "Stochastic contextfree grammars for tRNA modeling," *Nucleic Acids Res.*, vol. 22, pp. 5112-5120, 1994.

[10] S. R. Eddy and R. Durbin, "RNA sequence analysis using covariance models," *Nucleic Acids Res.*, vol. 22, pp. 2079-2088, 1994.

[11] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy, "Rfam: an RNA family database," *Nucleic Acids Res.*, vol. 31, pp. 439-441, 2003.

[12] Z. Weinberg and W. L. Ruzzo, "Exploiting conserved structure for faster annotation of

non-coding RNAs without loss of accuracy," *Bioinformatics*, vol. 20, pp. i334-i341, 2004.

[13] P. P. Gardner, et al., "Rfam: updates to the RNA families database," *Nucleic Acids Res.*, vol. 37, no. Issuesuppl 1, pp. D136-D140, 2009.

[14] D. Gautheret and A. Lambert, "Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles," *J. Mol. Biol.*, vol. 313, pp. 1003-1011, 2001.

[15] A. Lambert, et al., "The ERPIN server: an interface to profile-based RNA motif identification," *Nucleic Acids Res.*, vol. 32, pp. W160-W165, 2004.

[16] D. Gautheret, F. Major, and R. Cedergren, "Pattern searching/alignment with RNA primary and secondary structures: an effective descriptor for tRNA," *Comput. Appl. Biosci.*, vol. 6, pp. 325-331, 1990.

[17] P. Thebault, S. Givry, T. Schiex, and C. Gaspin, "Searching RNA motifs and their intermolecular contacts with constraint networks," *Bioinformatics*, vol. 22, pp. 354-361, 2006.

[18] J. Sponer, J. Leszczyński, and P. Hobza, "Nature of Nucleic Acid−Base Stacking: Nonempirical ab Initio and Empirical Potential Characterization of 10 Stacked Base Dimers. Comparison of Stacked and H-Bonded Base Pairs," *J. Phys. Chem.*, vol. 100, no. 13, p. 5590–5596, 1996.

[19] H. A. Heus, S. S. Wijmenga, H. Hoppe, and C. W. Hilbers, "The detailed structure of tandem G.A mismatched base-pair motifs in RNA duplexes is context dependent," *J. Mol. Biol.*, vol. 271, no. 1, pp. 147-158, 1997.

[20] M. Parisien and F. Major, "The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data," *Nature*, vol. 452, pp. 51-55, 2008.

[21] F. Major, et al., "The combination of symbolic and numerical computation for three-dimensional modeling of RNA," *Science*, vol. 253, pp. 1255-1260, 1991.

[22] P. Gendron, S. Lemieux, and F. Major, "Quantitative analysis of nucleic acid three-dimensional structures," *J. Mol. Biol.*, vol. 308, pp. 919-936, 2001.

[23] S. Lemieux and F. Major, "RNA canonical and non-canonical base pairing types: a

recognition method and complete repertoire," *Nucleic Acids Res.*, vol. 30, pp. 4250-4263, 2002.

[24] N. B. Leontis and E. Westhof, "Geometric nomenclature and classification of RNA base pairs," *RNA*, vol. 7, pp. 499-512, 2001.

[25] N. B. Leontis, J. Stombaugh, and E. Westhof, "The non-Watson-Crick base pairs and their associated isostericity matrices," *Nucleic Acids Res.*, vol. 16, pp. 3497-3531, 2002.

[26] F. Major and P. Thibault, "RNA Tertiary Structure Prediction," in *Bioinformatics: From Genomes to Therapies*. Weinheim, Germany: Wiley-VCH, 2007, pp. 491-536.

[27] J. D. Horton, "A polynomial-time algorithm to find the shortest cycle basis of a graph," *SIAM J. Comp.*, vol. 16, pp. 358-366, 1987.

[28] L. Boltzmann, "The Second Law of Thermodynamics," in *Theoretical physics and Philosophical Problems: Selected Writings*, L. Boltzmann, Ed. Dordrecht, Netherlands: D. Reidel, 1974, pp. 14-32.

[29] U. Montanari, "Networks of constraints: Fundamental properties and applications to picture processing," *Information Sciences*, vol. 7, pp. 95-132, 1974.

[30] K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philos. Mag.*, vol. 6, no. 2, p. 559–572, 1901.

[31] G. A. Fichant and C. Burks, "Identifying potential tRNA genes in genomic DNA sequences," *J. Mol. Biol.*, vol. 220, pp. 659-671, 1991.

[32] M. Höchsmann, B. Voss, and R. Giegerich, "Pure Multiple RNA Secondary Structure Alignments: A Progressive Profile Approach. IEEE/ACM Trans," *Comput. Biol. Bioinformatics*, vol. 1, pp. 53-62, 2004.

[33] P. Thebault, *Formalisme CSP et localisation de motifs structurés dans les textes génomiques*. Toulouse: Thèse de doctorat, École Doctorale Biologie Santé Biothecnologies, 2004.

[34] A. Morales Helguera, J. E. Rodriguez-Borges, X. Garcia-Mera, F. Fernandez, and M. N. Cordeiro, "Probing the Anticancer Activity of Nucleoside Analogues: A QSAR Model Approach Using an Internally Consistent Training Set," *J. Med. Chem.*, vol. 50,

pp. 1537-1545, 2007.

[35] R. Todeschini, V. Consonni, and M. Pavan, *Dragon Software, version 2.1*. 2002.

[36] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*. Wiley-VCH, 2000.

[37] Y. Marrero Ponce, J. A. Castillo Garit, and D. Nodarse, "Linear indices of the 'macromolecular graph's nucleotides adjacency matrix' as a promising approach for bioinformatics studies. Part 1: Prediction of paromomycin's affinity constant with HIV-1 Psi-RNA packaging region," *Bioorg. Med. Chem.*, vol. 13, pp. 3397-3404, 2005.

[38] H. Gonzalez-Diaz, S. Vilar, L. Santana, G. Podda, and E. Uriarte, "On the applicability of QSAR for recongnition of miRNA bioorganic structures at early stages of organism and cell development: Embryo and stem cells," *Bioorg. Med. Chem.*, vol. 15, pp. 2544-2550, 2007.

[39] I. J. Tinoco, et al., "Improved estimation of secondary structure in ribonucleic acids," *Nat. New Biol.*, vol. 246, pp. 40-41, 1973.

[40] S. M. Freier, et al., "Improved free-energy parameters for predictions of RNA duplex stability," *Proc. Natl. Acad. Sci. USA*, vol. 83, pp. 9373-9377, 1986.

[41] K. St-Onge, P. Thibault, S. Hamel, and F. Major, "Modeling RNA tertiary structure motifs by graph-grammars," *Nucleic Acids Res.*, vol. 35, no. 5, pp. 1726-1736, 2007.

[42] N. B. Leontis, J. Stombaugh, and E. Westhof, "Motif prediction in ribosomal RNAs Lessons and prospects for automated motif prediction in homologous RNA molecules," *Biochimie*, vol. 84, p. 961, 2002.

[43] H. M. Berman, et al., "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, pp. 235-242, 2000.

[44] M. Nagl, "Set theoretic approaches to graph grammars," in *Graph-Grammars and their Application to Computer Science*. Springer: Ehrig, H., Nagl, M., Rozenberg, G. and Rosenfeld, A. (eds),, 1987, pp. 41-54.

[45] C. V. Jones, "An integrated modeling environment based on attributed graphs and graph-grammars," *Dec. Support Syst.*, vol. 10, pp. 255-275, 1993.

[46] A. A. Szewczak and P. B. Moore, "The Sarcin/Ricin Loop, a Modular RNA," *J. Mol. Biol.*, vol. 247, p. 81–98, 1995.

[47] A. A. Szewczak, P. B. Moore, Y. L. Chang, and I. G. Wool, "The conformation of the sarcin/ricin loop from 28S ribosomal RNA," *Proc. Natl Acad. Sci. U.S.A.*, vol. 90, pp. 9581-9585, 1993.

[48] J. C. Burnett and J. J. Rossi, "RNA-Based Therapeutics: Current Progress and Future Prospects," *Chem. Biol.*, vol. 19, pp. 60-71, 2012.

[49] M. H. Bailor, X. Sun, and H. M. Al-Hashimi, "Topology links RNA secondary structure with global conformation, dynamics, and adaptation," *Science*, vol. 327, pp. 202-206, 2010.

[50] J. Caballero, L. Fernández, J. I. Abreu, and M. Fernández, "Amino Acid Sequence Autocorrelation vectors and ensembles of Bayesian-Regularized Genetic Neural Networks for prediction of conformational stability of human lysozyme mutants," *J. Chem. Inf. Model.*, vol. 46, no. 3, pp. 1255-1268, 2006.

[51] M. A. Cabrera, I. Gonzalez, C. Fernandez, C. Navarro, and M. Bermejo, "A topological substructural approach for the prediction of P-glycoprotein substrates," *J. Pharm. Sci.*, vol. 95, pp. 589-606, 2006.

[52] M. Perez Gonzalez, H. Gonzalez Diaz, R. Molina Ruiz, M. A. Cabrera, and R. Ramos de Armas, "TOPS-MODE based QSARs derived from heterogeneous series of compounds. Applications to the design of new herbicides," *J. Chem. Inf. Comput. Sci.*, vol. 43, pp. 1192-1199, 2003.

[53] Z. Xiao, et al., "Antitumor Agents. 213. Modeling of Epipodophyllotoxin Derivatives Using Variable Selection k Nearest Neighbor QSAR Method," *J. Med. Chem.*, vol. 45, no. 11, p. 2294–2309, 2002.

[54] J. J. Cannone, et al., "The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs," *BMC Bioinf.*, vol. 3, p. 2, 2002.

[55] F. M. Jucker and A. Pardi, "GNRA tetraloops make a U-turn," *RNA*, vol. 1, pp. 219-222, 1995.

[56] D. Moazed, J. M. Robertson, and H. F. Noller, "Interaction of elongation factors EF-G and EF-Tu with a conserved loop in 23 S RNA," *Nature*, vol. 334, p. 362–364, 1988.

[57] Y. Endo and I. G. Wool, "The site of action of alpha sarcin on eukaryotic ribosomes," *J. Biol. Chem.*, vol. 257, p. 9054–9060, 1982.

[58] Y. Endo, M. Mitsui, M. Motizuki, and K. Tsurugi, "The mechanism of action of ricin and related toxic lectins on eukaryotic ribosomes. The site and the characteristics of the modification in 28 S ribosomal RNA caused by the toxins," *J. Biol. Chem.*, vol. 262, p. 5908–5912, 1987.

[59] M. Dam, S. Douthwaite, T. Tenson, and A. S. Mankin, "Mutations in domain II of 23 S rRNA facilitate translation of a 23 S rRNA-encoded pentapeptide conferring erythromycin resistance," *J. Mol. Biol.*, vol. 259, pp. 1-6, 1996.

[60] S. T. Gregory and A. E. Dahlberg, "Mutations in the conserved P loop perturb the conformation of two structural elements in the peptidyl transferase center of 23 S ribosomal RNA," *J. Mol. Biol.*, vol. 285, no. 4, pp. 1475-1483, 1999.

[61] S. Douthwaite, T. Powers, J. Y. Lee, and H. F. Noller, "Defining the structural requirements for a helix in 23 S ribosomal RNA that confers erythromycin resistance," *J. Mol. Biol.*, vol. 209, pp. 655-665, 1989.

[62] C. Aagaard and S. Douthwaite, "Requirement for a conserved, tertiary interaction in the core of 23S ribosomal RNA," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 91, pp. 2989-2993, 1994.

[63] I. Leviev, S. Levieva, and R. A. Garrett, "Role for the highly conserved region of domain IV of 23S-like rRNA in subunit-subunit interactions at the peptidyl transferase centre," *Nucleic Acids Res.*, vol. 23, pp. 1512-1517, 1995.

[64] M. Bocchetta, L. Xiong, and A. S. Mankin, "23S rRNA positions essential for tRNA binding in ribosomal functional sites," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 95, no. 7, pp. 3525-3530, 1998.

[65] R. Green, R. R. Samaha, and H. F. Noller, "Mutations at nucleotides G2251 and U2585 of 23 S rRNA perturb the peptidyl transferase center of the ribosome," *J. Mol. Biol.*, vol. 266, no. 1, pp. 40-50, 1997.

[66] S. T. Gregory, K. R. Lieberman, and A. E. Dahlberg, "Mutations in the peptidyl transferase region of E. coli 23S rRNA affecting translational accuracy," *Nucleic Acids Res.*, vol. 22, no. 3, pp. 279-284, 1994.

[67] K. R. Lieberman and A. E. Dahlberg, "The importance of conserved nucleotides of 23 S ribosomal RNA and transfer RNA in ribosome catalyzed peptide bond formation," *J. Biol. Chem.*, vol. 269, no. 23, pp. 16163-16169, 1994.

[68] M. O'Connor, et al., "Genetic probes of ribosomal RNA function," *Biochem. Cell Biol.*, vol. 73, pp. 859-868, 1995.

[69] B. T. Porse, H. P. Thi-Ngoc, and R. A. Garrett, "The Donor Substrate Site within the Peptidyl Transferase Loop of 23 S rRNA and its Putative Interactions with the CCA-end of N-blocked Aminoacyl-tRNAPhe," *J. Mol. Biol.*, vol. 264, no. 3, pp. 472-483, 1996.

[70] C. M. Spahn, J. Remme, M. A. Schäfer, and K. H. Nierhaus, "Mutational analysis of two highly conserved UGG sequences of 23 S rRNA from Escherichia coli," *J. Biol. Chem.*, vol. 271, no. 51, pp. 32849-32856, 1996.

[71] Y. L. Chan and I. G. Wool, "The integrity of the sarcin/ricin domain of 23 S ribosomal RNA is not required for elongation factor-independent peptide synthesis," *J. Mol. Biol.*, vol. 378, no. 1, pp. 12-19, 2008.

[72] Y. L. Chan, J. Dresios, and I. G. Wool, "A pathway for the transmission of allosteric signals in the ribosome through a network of RNA tertiary interactions," *J. Mol. Biol.*, vol. 355, no. 5, pp. 1014-1025, 2006.

[73] Y. L. Chan, A. S. Sitikov, and I. G. Wool, "The phenotype of mutations of the base-pair C2658.G2663 that closes the tetraloop in the sarcin/ricin domain of Escherichia coli 23 S ribosomal RNA," *J. Mol. Biol.*, vol. 298, no. 5, pp. 795-805, 2000.

[74] M. R. Macbeth and I. G. Wool, "The phenotype of mutations of G2655 in the sarcin/ricin domain of 23 S ribosomal RNA," *J. Mol. Biol.*, vol. 285, no. 3, pp. 965-75, 1999.

[75] J. W. Ponder, *TINKER - software tools for molecular design, version 4.2*. 2004.

[76] N. Spackova and J. Sponer, "Molecular dynamics simulations of sarcin–ricin," *Nucleic*

Acids Res., vol. 34, no. 2, pp. 697-708, 2006.

[77] A. Yassin and A. S. Mankin, "Potential new antibiotic sites in the ribosome revealed by deleterious mutations in RNA of the large ribosomal subunit," *J. Biol. Chem.*, vol. 282, no. 33, pp. 24329-24342, 2007.

[78] T. Uchiumi, N. Sato, A. Wada, and A. Hachimori, "Interaction of the Sarcin/Ricin Domain of 23 S Ribosomal RNA with Proteins L3 and L6*," *J. Biol. Chem.*, vol. 274, no. 2, pp. 681-686, 1999.

[79] C. C. Correll, J. Beneken, M. J. Plantinga, M. Lubbers, and Y. L. Chan, "The common and the distinctive features of the bulged-G motif based on a 1.04 Å resolution RNA structure," *Nucleic Acids Res.*, vol. 31, no. 23, pp. 6806-6818, 2003.

[80] M. D. Nteo, *Investigation of Novel Erythromycin Resistance Mechanisms Arising from Heterologous Expression of Gram Positive DNA in Escherichia Coli*. Johannesburg, South Africa: University of Johannesburg, 2006.

[81] D. Moazed and H. F. Noller, "Sites of interaction of the CCA end of peptidyl-tRNA with 23S rRNA," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 88, no. 9, pp. 3725-3728, 1991.

[82] D. Moazed and H. F. Noller, "Interaction of tRNA with 23S rRNA in the ribosomal A, P, and E sites," *Cell*, vol. 57, no. 4, pp. 585-597, 1989.

[83] G. Steiner, E. Kuechler, and A. Barta, "Photo-affinity labelling at the peptidyl transferase centre reveals two different positions for the A- and P-sites in domain V of 23S rRNA," *EMBO J.*, vol. 7, no. 12, pp. 3949-3955, 1988.

[84] R. Green, C. Switzer, and H. F. Noller, "Ribosome-catalyzed peptide-bond formation with an A-site substrate covalently linked to 23S ribosomal RNA," *Science*, vol. 280, no. 5361, pp. 286-289, 1998.

[85] H. Steinhaus, "Sur la division des corps matériels en parties," *Bull. Acad. Polon. Sci.*, vol. 4, no. 3, pp. 801-804, 1956.

[86] W. L. DeLano, "The PyMOL Molecular Graphics System," DeLano Scientific, 2002.

[87] J. Wang, P. Cieplak, and P. A. Kollman, "How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules?," *J. Comput. Chem.*, vol. 21, pp. 1049-1074, 2000.

[88] M. Plutowski, S. Sakata, and H. White, "Cross-validation estimates IMSE," in *Advances in Neural Information Processing Systems 6*. San Mateo, CA: Morgan Kaufman, 1994, pp. 391-398.

[89] M. Quenouille, "Notes on bias in estimation," *Biometrika*, no. 43, pp. 353-360, 1956.

[90] E. A. Maxwell, *Methods of Plane ProjectiveGeometry Based on the Use of General Homogeneous Coordinates*. Cambridge, England: Cambridge University Press, 1946.

[91] J. Nocedal, "Updating Quasi-Newton Matrices with Limited Storage," *Mathematics of Computation*, vol. 35, pp. 773-782, 1980.

[92] R. Battiti and F. Masulli, "BFGS Optimization for Faster and Automated," in *International Neural Network Conference*, Paris, France, 1990, p. 757–760.

[93] B. Lee and F. M. Richards, "The interpretation of protein structures: estimation of static accessibility," *J. Mol. Biol.*, vol. 55, no. 3, pp. 379-400, 1971.

[94] A. Shrake and J. A. Rupley, "Environment and exposure to solvent of protein atoms. Lysozyme and insulin," *J. Mol. Biol.*, vol. 79, no. 2, pp. 351-371, 1973.

[95] G. M. Voronoï, "Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Deuxième mémoire : recherches sur les parallélloèdres primitifs.," *J. Reine Angew. Math.*, no. 134, p. 198–287, 1908.

Annexe

MC-RMSD

Nous utilisons une distance métrique de fragments tertiaires d'ARN pour calculer la moyenne des racines carrées des déviations (RMSDs) « Root-Mean-Square Deviations » entre deux paires de structures tertiaires. Dans notre laboratoire, un outil à été développé, *MC-RMSD* [22], qui prend en entrée deux fragments tertiaires d'ARN et calcule la distance en Angstrom (Å) entre ceux-ci. Un Angstrom est une unité de longueur valant 0,1 nanomètre, soit 10^{-10} mètre

Brièvement, pour deux bases par exemple, *MC-RMSD* établi le référentiel local de celles-ci et calcule la distance en utilisant des matrices de transformations homogènes [90]. Le référentiel local est déterminé comme suit :

- $\vec{u}$: Vecteur unitaire de l'atome N1 vers l'atome C2 pour les pyrimidines (N9 et C4 pour les purines).
- $\vec{v}$: Vecteur unitaire de l'atome N1 vers l'atome C6 pour les pyrimidines (N9 et C8 pour les purines).
- $\vec{y}$: Vecteur unitaire orienté par $u + v$.
- $\vec{z}$: Vecteur unitaire orienté par le produit vectoriel $u \times v$.
- $\vec{x}$: Vecteur unitaire, suivant la règle de la main droite, est donné par $y \times z$.

La position relative des référentiels locaux peuvent s'exprimer en utilisant des matrices de transformations homogènes. Ces matrices (4×4) encodent les opérations géométriques nécessaires afin de déplacer un référentiel local sur l'autre.

Minimisation

Dans le but de raffiner les structures tertiaires d'ARN que nous modélisons, nous utilisons le logiciel TINKER [75]. Il utilise l'algorithme d'optimisation numérique (L-

BFGS) [91], une version « limited memory » de la méthode de Broyden–Fletcher–Goldfarb–Shanno (BFGS) [92], qui prend en entrée une structure tertiaire et ajuste les coordonnées des atomes dans le but de minimiser l'énergie. La **Figure 38** illustre un exemple de minimisation de structure.

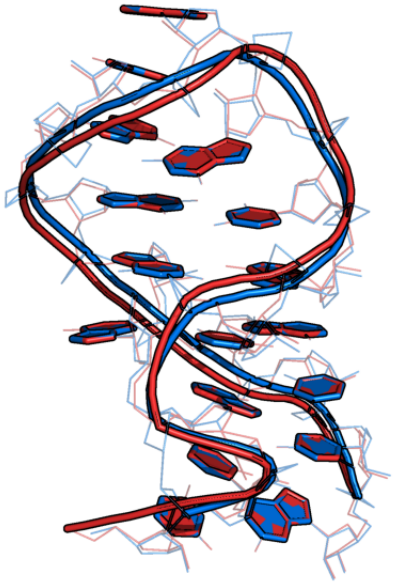


Figure 38. Minimisation. La structure en rouge est un modèle généré par *MC-Sym* et la structure bleue le résultat de la minimisation appliquée sur la structure rouge. Les liaisons phosphodiesters sont illustrées par un cylindre. Les nucléotides sont illustrés par des formes planaires.

K-means

Nous utilisons une des techniques de classification non supervisée « clustering » les plus utilisées, K-means [85], pour regrouper les atomes des structures tertiaires d'ARN. Étant donné un entier K, K-means partitionne les éléments en K groupes ne se chevauchant pas. Ce résultat est obtenu en positionnant K centroïdes dans les régions de l'espace les plus peuplées. Chaque observation est alors affectée au centroïde le plus proche (règle dite de la

Distance Minimale). Chaque classe contient donc les éléments qui sont plus proches d'un certain centroïde que de tout autre centroïde (voir **Figure 39**).

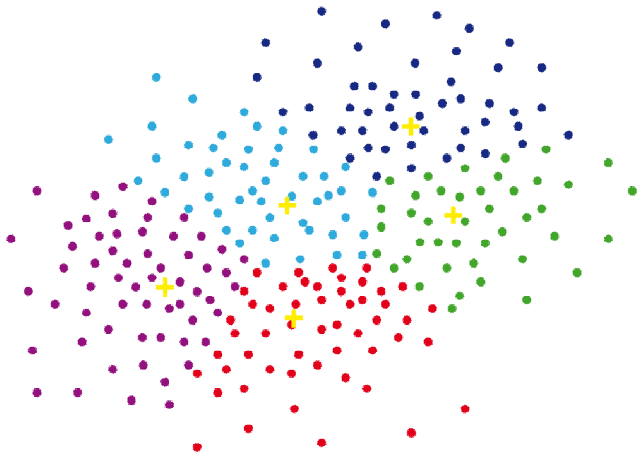


Figure 39. K-Means. Regroupement dans l'espace bidimensionnel de points dans 5 groupes (magenta, cyan, bleu, rouge et vert). Les croix jaunes représentent les centroïdes de chaque groupe.

Surface accessible d'un atome

La surface accessible d'un atome est l'aire de la surface de la biomolécule qui est accessible par un solvant, généralement une molécule d'eau. Cette surface est habituellement calculée en Angstrom carré et a été décrite en premier par Lee et Richards en 1971 [93]. La surface est typiquement calculée par l'algorithme « rolling ball » développé par Shrake et Rupley en 1973 [94]. Cet algorithme utilise une sphère (solvant) d'un rayon particulier pour sonder la surface de la molécule (voir **Figure 40**).

La surface d'une molécule, représentée par des boules de van der Waals, est constituée par la frontière de l'ensemble de ces boules définies par le rayon de van der Waals de chaque atome de la molécule.

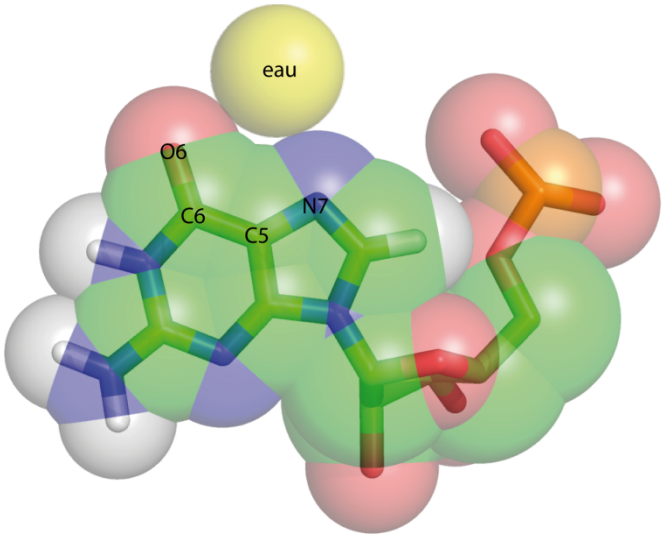


Figure 40. Surface accessible. La surface accessible est l'aire de la surface d'une molécule (ici le nucléotide G représenté par des boules de van der Waals) pouvant être accessible par un solvant, ici une molécule d'eau (en jaune). Dans cet exemple, la molécule d'eau a accès aux atomes O6 et N7, mais pas aux atomes C5 et C6. Les atomes d'azote sont en bleu ; oxygène en rouge ; carbone en vert ; phosphate en orange et hydrogène en blanc.

Distance euclidienne pondérée

La distance Euclidienne pondérée est comparable à la distance Euclidienne, voir **Eq. (5)**, mais permet l'introduction de coefficients w_i pour donner plus d'importance à la différence $x_i - y_i$ pour un i donné, voir **Eq. (6)**

$$\text{Distance Euclidienne}_{X,Y} = \sum_{i=1}^{n} w_i (x_i - y_i)^2 \qquad \textbf{(5)}$$

$$\text{Distance Euclidienne Pondérée}_{X,Y} = \sum_{i=1}^{n} w_i (x_i - y_i)^2 \qquad \textbf{(6)}$$

où X et Y sont deux points dans un espace à n dimensions, x_i (resp. y_i) est la valeur du point X (resp. Y) pour la dimension i.

La distance Euclidienne pondérée est considérée aussi comme une mesure de distorsion (voir **Figure 41**).

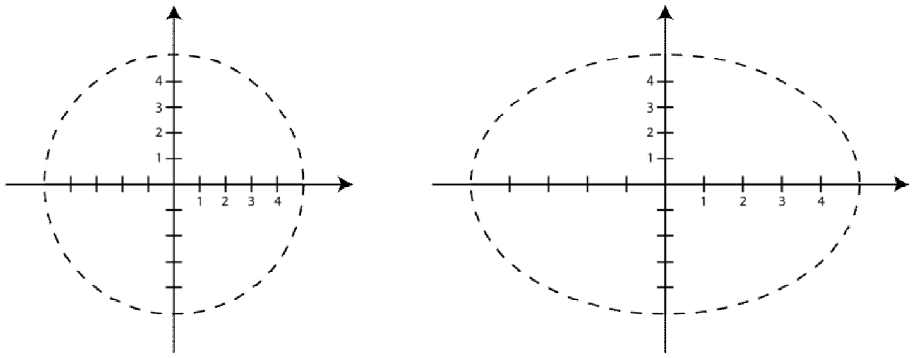


Figure 41. Distance euclidienne pondérée. Un exemple d'effet de distorsion. Gauche) Les axes ont le même poids, il s'agit de la distance euclidienne. Droite) Les axes ont des poids différents (dans ce cas-ci, l'axe horizontal a un poids plus grand que l'axe vertical), il s'agit de la distance euclidienne pondérée.

Méthode du plus proche voisin

C'est la méthode d'interpolation la plus simple, qui attribue à un point inconnu la valeur du point connu le plus proche. La méthode du plus proche voisin est associée aux diagrammes de Voronoï [95], qui consistent à partitionner le plan (voir **Figure 42**) de sorte que chaque point commence à grossir à la même vitesse, en définissant une région circulaire. Quand deux régions se rencontrent, une frontière linéaire se forme, le long de la médiatrice des deux points associés.

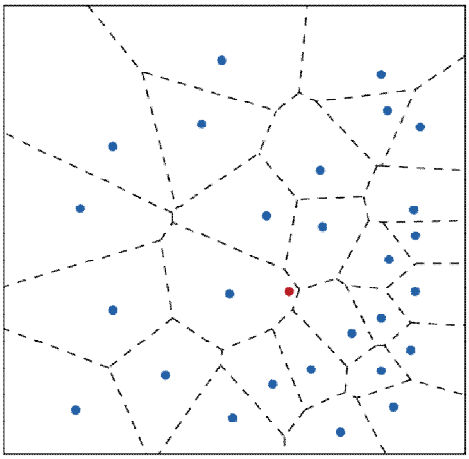


Figure 42. Plus proche voisin. Le plus proche voisin d'un nouveau point (rouge) est le point (bleu) associé à la case du diagramme de Voronoï où se situe le nouveau point (rouge).

LOOCV (Leave-One-Out Cross-Validation)

Le LOOCV (Leave-One-Out Cross-Validation) [88] est une technique permettant d'évaluer comment les résultats d'une analyse statistique vont se généraliser à un ensemble de données indépendantes. Il est principalement utilisé dans des contextes où le but est la prédiction, et où on veut estimer la précision d'un modèle prédictif.

Le LOOCV consiste à retirer une donnée du lot de données, de procéder à l'analyse sur l'ensemble des données restantes (données d'entraînement), et de valider l'analyse sur la donnée retirée (voir **Figure 43**). Cette procédure est répétée n fois (n est le nombre de données totale) afin de valider l'analyse sur chacune des données.

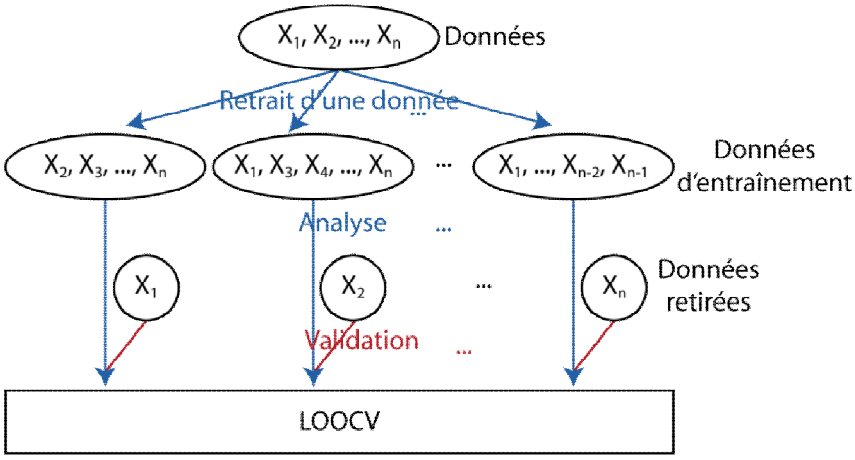


Figure 43. LOOCV. Pour une expérience de LOOCV, une donnée est retirée de l'ensemble de données initial pour former l'ensemble de données d'entraînement. À partir de l'analyse des données d'entraînement, la validation est effectuée sur la donnée retirée.