Université de Montréal

Recherche de motifs structuraux dans les complexes acides ribonucléiques / protéines

par

Mathieu Drapeau

Département d'informatique et recherche opérationnelle

Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures

en vue de l'obtention du grade de

Maîtrise ès sciences (M.Sc.)

en informatique

Décembre, 2002

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé :

Recherche de motifs structuraux dans les complexes acides ribonucléiques / protéines

présenté par :

Mathieu Drapeau

a été évalué par un jury composé des personnes suivantes :

Nadia El-Mabrouk,

président-rapporteur

François Major,

directeur de recherche

Sébastien Roy,

membre du jury

Sommaire

L'analyse des complexes moléculaires composés d'acides ribonucléiques (ARN) et de protéines n'en est qu'à ses débuts. En outre, la compréhension des interactions ARN-protéine rendra possible le développement d'outils de prédiction automatique d'association. Une méthode exploratoire afin de déterminer les facteurs de redondance de structure est donc proposée.

Les structures tridimensionnelles et les relations symboliques composant les différentes informations contenues dans cette structure sont encodées dans un graphe relationnel non-dirigé. De là, les sous-unités intéressantes sont extraites et analysées afin de repérer de nouveaux motifs de structure. Ces régions sont potentiellement responsables de la reconnaissance et des associations spécifiques qui se produisent entre certains ARN et certaines protéines.

L'approche est basée sur un nouvel algorithme de recherche des sous-graphes maximaux locaux, qui repose sur une modification d'une recherche en profondeur. Comparativement aux méthodes utilisées antérieurement, cet algorithme s'avère être plus performant. Cette efficacité vient en majeure partie de l'omission volontaire des calculs d'isomorphisme de graphe qui ralentissent grandement l'exécution de l'analyse.

Cette méthode d'analyse est confirmée par la découverte de nouveaux motifs intéressants au sein des complexes ARN-protéine. De nouvelles caractéristiques présentes dans la structure récemment produite de la sous-unité 30S du ribosome sont dévoilées. De nombreuses autres utilités ont également été attribuées à cette méthode.

Mots clés : Complexes ARN / protéine, interaction ARN / protéine, recherche de motifs, analyse structurale, algorithme sous-graphes communs maximaux

Abstract

The analysis of the molecules composed of ribonucleic acids (RNA) and amino acids (protein) is at its beginning. The developments in this field of interest will open the way to new tools that predict which interaction may occur between proteins and RNAs. An exploratory method to determine the structure redundancy of the protein-RNA molecular complex is discussed.

The information contained in structures such as the type of residue and the chemical relation between them are encoded in a bidirectional relational graph. Computations are performed on these graphs to query for interesting information and to discover new structural motifs. These regions or sub-structure are potentielly the main factor of influence in the specific associations of a given RNA to particular proteins.

The method we developed is based on a new algorithm that finds the maximal local subgraphs. It is a modification of a recursive depth-first search algorithm. By comparison with the other approches, graph isomorphism is not calculated. This allows our algorithm to be of higher efficiency.

This research results in the discovery of new motifs in RNA-protein complexes. New structural characteristics are presented from studies of the 30S ribosomal subunit. We also discuss other utilities that this approach now allows to perform more easily.

Key words : RNA / protein complex, RNA / protein interaction, structural motif, structural pattern, maximum common subgraph algorithm

Table des matières

Liste des Figures vii

Liste des Tables x

Chapitre 1 : Introduction et notions préliminaires 1

1.1 Constituants des cellules vivantes 2

1.2 Les ARN 3

1.3 Structure des polynucléotides 4

1.4 Les protéines 6

1.5 Structure des protéines 8

1.6 Les complexes ARN-protéine 9

Chapitre 2 : Stratégie 12

2.1 Motifs structuraux 12

2.2 Revue des méthodes existantes 13

2.3 Définition du problème 14

2.4 Influence des travaux antérieurs 14

2.5 Représentation des structures 15

2.6 Recherche de motifs récurrents **(Notre stratégie employée)** 17

Chapitre 3 : Nos algorithmes 19

3.1 Description de l'algorithme implanté 22

3.2 Analyse et tri des sous-graphes générés 25

Chapitre 4 : Résultats **28**

4.1 Analyse statistique des points d'interaction 28

4.2 Disposition des acides aminés autour des nucléotides 29

4.3 Analyse de la sous-unité ribosomale 30S et de ses protéines 31

4.4 Motif conservé retrouvé chez 1ASZ et 1B7F 35

4.5 Autres motifs conservés 39

4.6 Compariason de la rapidité d'exécution 39

Chapitre 5 : Discussion **41**

Chapitre 6 : Conclusion **43**

Références **45**

Liste des Figures

1.1 Figure illustrant les différents groupements qui composent les résidus d'ARN et les liaisons qui peuvent apparaître entre eux. Les résidus *i* et *i-1* sont attachés par un lien phosphodiester. Les résidus *i* et *j* partagent une laison de type Watson-Crick entre leurs bases azotées. 3

1.2 Représentation des différents niveaux de complexité de l'ARN. La structure primaire (en **a**) est une séquence linéaire de l'ARN. Les repliements simples en 2D où sont représentés les laisons Watson-Crick (en **b**) sont caractéristique de la structure secondaire. La structure tertaire (en **c**) fait référence à l'organisation spatiale de l'ARN. 5

1.3 Représentation de la structure des acides aminés et les différents groupements qui y sont attachés. Les polyaminoacides se créent par la liaison entre le groupement amine d'un résidu et le groupement acide d'un autre résidu juxtaposé. 6

1.4 Représentation caractéristiques de la structure secondaire des acides aminés. La structure de cette protéine (code pdb : 1DUL) illustre les hélices α qui sont indiqués par les formes spiralées. 8

2.1 Représentation d'une structure sous forme d'un graphe relationnel. Les différents types de noeuds et d'arcs du graphe sont indiqués par des couleurs différentes. 15

2.2 Graphe relationnel associé à une structure tridimensionnelle spécifique. Le graphe relationnel (en **a**) est caractérisé par différent types d'arcs qui correspondent aux différentes liaisons que l'on retrouve dans la structure tridimensionnelle (en **b**). Les types d'arc retrouvés sont : [1] Adjacence (nucléotide - nucléotide), [2] Liaison Watson-Crick, [3] Liaison d'interaction (nucléotide - acide aminé), [4] Liaison structure secondaire (hélice *alpha* ou feuillet *beta*), [5] Adjacence (acide aminé - acide aminé). 17

3.1 Le sous-graphe maximal commun des deux graphes illustrés est représenté en bleu. 21

3.2 Schéma des différentes étapes de l'algorithme du calcul des sous-graphes communs maximaux. Les deux graphes sont comparés entre eux et l'on fait correspondre les noeuds d'un graphe avec ceux de l'autre ("graph matching"). **a)** Interaction de départ de l'algorithme (en vert) et chemins potentiels pouvant étendre le sous-graphe. **b)** Un chemin optimal est sélectionné (en rouge) et marqué afin de ne plus repasser par ce chemin. Dans ce cas-ci, le chemin optimal est le chemin possédant le plus de noeuds. L'algorithme recalcule les chemins disponibles de nouveau en évitant de repasser par le chemin conservé. **c)** Le chemin optimal est conservé (en rose) et les chemins potentiels sont recalculés. **d)** Tous les chemins conservés sont additionnés pour former deux sous-graphes communs maximaux. **e)** Sous-graphes maximaux obtenus par l'algorithme en 1. 23

4.1 Figure représentant la disposition de l'acide aminé arginine autour du nucléotide guanine retrouvé dans les structures analysées (voir tableau 4.2). Le résidu d'ARN représenté en jaune est un ensemble de guanines superposés autour desquels chaque arginine y étant liée figure en bleu. Cet ensemble d'interactions guanine-arginine ont été produit en les extrayant de différentes structures (voir tableau 4.1). 31

4.2 Structure de la sous-unité ribosomale 30S et les différentes protéines s'y liant. Chaque couleur identifie les protéines différentes fixées au ribosomes. L'ARN du ribosome est représenté dans les teintes de gris. 32

4.3 Emplacement des motifs conservés retrouvés dans la structures de la sous-unité 30S ribosomale. En rouge et vert, les résidus d'ARN sont représentés. Les résidus d'acides aminés (lysine et arginine) figurent également en noir. . 33

4.4 Vue stéréoscopique du motif retrouvé (celui en **a** représente le motif en rouge et celui en **b** représente le motif en vert) tel que démontré selon la figure 4.3. 34

4.5 Superpositon des deux structures du motif redondant retrouvé dans la sous-unité 30S du ribosome. 35

4.6 Emplacement des motifs similaires retrouvés dans les structures des molécules ASZ et B7F. Les résidus d'acide aminé sont représentés en couleur plus foncée tandis que le ribonucléotide du motif est en bleu plus pâle. 36

4.7 Vue stéréoscopique du motif présent chez ASZ (en **a**) et chez B7F (en **b**). . 37

4.8 Motif utilisé pour tester la rapidité de l'algorithme développé. 39

Liste des Tables

1.1 20 acides aminés différents et leur abréviation. 7

4.1 Structures utilisées pour l'analyse des complexes ARN-protéine 29

4.2 Tableau de la composition des interactions ARN-protéine (% du nombre d'acides aminés en fonction du nombre de résidus d'ARN répertorié) 30

4.3 Tableau des motifs découverts. 38

À mes parents,

Chapitre 1

Introduction et notions préliminaires

La compréhension des mécanismes biologiques de régulation des composés biochimiques fait partie des étapes importantes qui permettront la découverte du fonctionnement des cellules vivantes. Une étape primordiale sera certainement franchie au courant de l'année 2003 avec la complétion de la phase finale du séquençage de "haute-résolution" de l'ADN humain entrepris par le projet *HUGO* en 1990 [1]. Parallèlement, d'autres axes de recherche seront explorés suite à cette gigantesque prolifération de données et aux énigmes qu'elle engendre. De toutes ces séquences, il nous faut comprendre où se situent les gènes, quelles sont leur structure tridimensionnelle et comment ils interagissent entre eux. Outre le séquençage, la spectroscopie par résonance nucléaire et la cristallographie à rayons X nous permettent d'obtenir des structures tridimensionnelles de toutes ces molécules biochimiques. Notre étude s'intéresse à ces structures, et plus particulièrement à l'analyse des agencements qui existent entre deux ou plusieurs structures différentes.

Les protéines et les ARN (acides ribonucléiques) sont des molécules distinctes qui se retrouvent à l'intérieur de toute cellule vivante. Elles interagissent souvent ensemble pour former des complexes qui, une fois assemblés, permettent d'accomplir des réactions biochimiques qui n'auraient pas pu avoir lieu avec chacune des entités prise individuellement. Les fonctions découlant de ces interactions sont de diverses natures, par exemple, la contraction musculaire, la régulation hormonale, la digestion, le transport des nutriments, la régulation métabolique ou encore la réplication du matériel génétique [2]. Il est important de pouvoir prédire le rôle biochimique d'un complexe donné à partir de sa structure et de l'organisation spatiale entre ses composants biologiques. Dans le domaine pharmaceutique, l'étude

des complexes impliqués dans différentes maladies permet d'imaginer de nouvelles cibles thérapeutiques grâce à l'interdépendance retrouvée entre les molécules le composant. Pour qu'un complexe entre une protéine et un ARN puisse se constituer, diverses liaisons doivent se former. Les liaisons qui peuvent se former sont de trois types : ionique, hydrogène et van der Waals [3, 4, 5].

Il existe plusieurs méthodes expérimentales pour déterminer des interactions ARN-protéine, celles-ci utilisent majoritairement des composés radioactifs et localisent précisement les points de contacts entre l'ARN et les protéines [6, 7, 8]. D'autres méthodes employées dans les laboratoires de biologie moléculaire utilisent les techniques de mutagénèse dirigée [9]. À partir des structures tridimensionnelles de ces assemblages protéines/ARN, nous pouvons explorer des conformations spatiales qui demeurent constantes entre les modèles et en définir des sous-structures récurrentes[1]. Notre étude porte donc sur la recherche de représentations spatiales constantes au sein d'un ensemble de complexes ARN-protéine.

1.1 Constituants des cellules vivantes

La cellule est l'unité fondamentale qui constitue chaque être vivant. Les formes de vie simplistes sont constituées de cellules isolées qui se propagent en se divisant en deux. Ce mécanisme est appelé "mitose". Par contre, les organismes supérieurs sont des cités cellulaires où certaines peuvent accomplir des fonctions spécialisées. Les cellules sont donc à mi-chemin entre la molécule et l'être humain. Afin de coordonner leur fonctionnement, les cellules vivantes ont besoin de petites molécules organiques telles que *les acides aminés, les nucléotides, les sucres* et *les acides gras*. Ces molécules peu complexes peuvent ensuite s'associer pour former des polymères[2] de grandes dimensions. Un acide aminé peut se lier

[1]Le terme "récurrence" se rapporte à la notion de motif structural qui sera défini à la section 1.3.

[2]Deux molécules peuvent être jointes ensemble par une liaison particulière, ce qui en résulte un dimère (dans le cas des acides aminés, le terme "dipeptide" est utilisé, tandis que pour les acides nucléiques on utilise

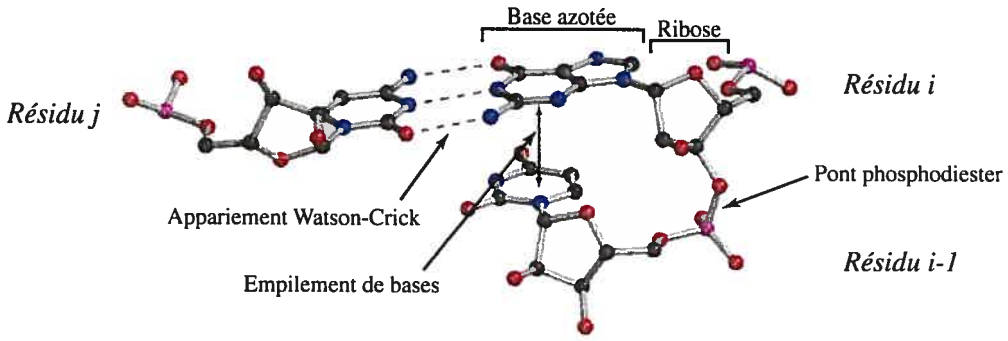


FIG. **1.1. Figure illustrant les différents groupements qui composent les résidus d'ARN et les liaisons qui peuvent apparaître entre eux. Les résidus *i* et *i-1* sont attachés par un lien phosphodiester. Les résidus *i* et *j* partagent une laison de type Watson-Crick entre leurs bases azotées.**

à un autre en formant une liaison peptidique, alors que deux nucléotides peuvent s'associer par un pont phosphodiester [10] (voir la figure 1.1). La répétition de ces réactions conduit à des polymères que l'on nomme **polypeptide** et **polynucléotide**. Les polypeptides, communément appelés *protéines*, seront traités aux section 1.4 et 1.5.

1.2 Les ARN

Les polynucléotides se retrouvent sous deux formes distinctes : *les acides ribonucléiques (ARN)* et *les acides désoxyribonucléiques (ADN)*. L'information génétique de tous les organismes vivants est entreposée dans l'ADN (sauf pour les *virus à ARN*). Un segment d'ADN qui contient l'information requise pour la synthèse d'un produit biologique fonctionnel (protéine ou ARN) est appelé un gène. Plusieurs classes d'ARN sont retrouvées dans les cellules, chacune possédant une fonction distincte. **L'ARN ribosomal** (ARNr)

---

le terme "dinucléotide".). Un oligomère est obtenu en joignant un faible nombre de molécules ensemble (oligopeptide pour les acides aminés et oligonucléotide pour les acides nucléiques). Quand il y a assemblage de plusieurs molécules, on nomme ce produit "polymère" ('polypeptide" ou "polynucléotide" dans le cas des acides aminés et acides nucléiques). Les unités composant ces structures sont souvent appelées **résidus**.

est la composante structurelle des ribosomes, les complexes qui élaborent la synthèse des protéines. **L'ARN messager** (ARNm) correspond aux acides nucléiques qui ont été générés par une batterie de molécules biochimiques à partir de l'ADN, et qui contiennent l'information d'un ou de plusieurs gènes. **Les ARN de transfert** (ARNt) sont les molécules qui traduisent l'information contenue sur les ARNm en une séquence spécifique d'acides aminés. Les molécules d'ARN et d'ADN sont constituées d'un enchaînement de quatres types de nucléotides. Ces nucléotides (voir figure 1.1) se composent : (1) d'une base azotée, (2) d'une pentose ou sucre et (3) d'un groupement phophate. Les bases azotées sont dérivées de deux composés apparentés, les **pyrimidines** et les **purines**. Les résidus [3] d'ADN et d'ARN sont formés de deux bases purines, l'**adénine** (A) et la **guanine** (G). L'ADN et l'ARN contiennent également deux pyrimidines, l'une d'entre elles étant la **cytosine** (C). La différence existant entre les séquences d'ADN et d'ARN tient au fait que la seconde pyrimidine est la **thymine** (T) dans le cas de l'ADN et l'**uracile** (U) dans le cas de l'ARN.

1.3 Structure des polynucléotides

La structure tridimensionnelle adoptée par les acides nucléiques engendre de vastes fonctions spécifiques et différentes qui caractérisent aussi bien l'ADN que l'ARN. La structure bicaténaire de l'ADN a été découverte par Watson & Crick en 1953 [11] et son organisation spatiale suggère inévitablement la façon dont s'effectue le transfert de l'information [12]. La structure de l'acide nucléique est souvent décrite en termes de niveaux hiérarchiques de complexité (primaire, secondaire, tertiaire). La figure 1.2 illustre clairement les différences entre les différents niveaux de complexité retrouvé dans l'ARN. La structure primaire d'un acide nucléique fait référence à la séquence des nucléotides sous forme d'une structure de liaisons covalentes. Toute structure stable et régulière où l'on re-

[3] Le terme résidu est employé comme synonyme à nucléotide et permet de désigner la sous-unité constituant les polynucléotides. Il peut également être utilisé dans le cas des protéines pour désigner les molécules la constituant.

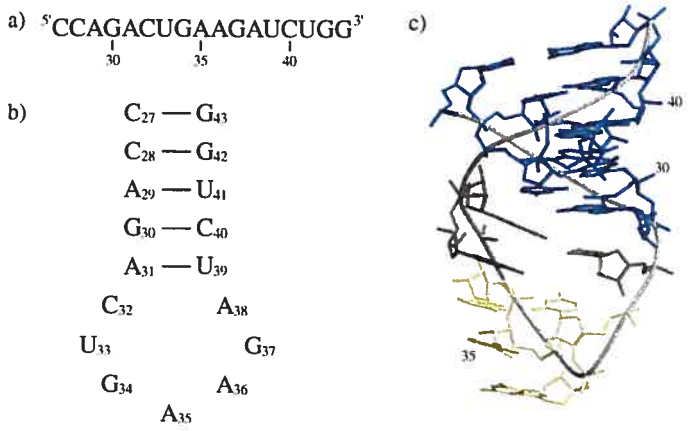


FIG. **1.2. Représentation des différents niveaux de complexité de l'ARN. La structure primaire (en a) est une séquence linéaire de l'ARN. Les repliements simples en 2D où sont représentés les laisons Watson-Crick (en b) sont caractéristique de la structure secondaire. La structure tertaire (en c) fait référence à l'organisation spatiale de l'ARN.**

trouve une ou plusieurs liaisons entre les nucléotides peut faire référence à une structure secondaire. C'est à ce niveau que l'on distingue de très importants appariements entre les bases appelées Watson-Crick qui se retrouvent autant dans les ADN que les ARN. Les repliements complexes de la structure secondaire sont généralement considérés comme la structure tertiaire. Dans ce mémoire, nous nous intéresserons plus particulièrement à la structure des ARN. Peu importe la classe d'ARN synthétisé, le produit de la transcription[4] est toujours un ARN simple brin. Cette nature moléculaire ne signifie pas que la structure des ARN est aléatoire. Les simples brins tendent à prendre une conformation hélicoïdale "right-handed" qui est dominée par des interactions de stabilité d'empilements de bases. L'empilement[5] est beaucoup plus intense entre deux purines qu'entre une purine et une

[4]Dans une cellule vivante normalement constituée, la traduction réfère à la copie du message génétique codé dans l'ADN en une molécule d'ARN complémentaire (ARNm) [13].

[5]L'empilement de base, plus communément appelé "base stacking" en anglais, réfère aux interactions de stabilisation qui prennent place entre les bases azotées de nucléotides. Les forces de dispersion de London et

FIG. **1.3. Représentation de la structure des acides aminés et les différents groupements qui y sont attachés. Les polyaminoacides se créent par la liaison entre le groupement amine d'un résidu et le groupement acide d'un autre résidu juxtaposé.**

pyrimidine ou encore qu'entre deux pyrimidines. L'ARN peut apparier ses bases avec un autre brin complémentaire. La règle standard d'appariement des bases de l'ARN (qui est identique à celle de l'ADN) consiste à apparier la cytosine avec la guanine et l'adénine avec l'uracile (la thymine dans le cas de l'ADN). La structure tridimensionnelle de plusieurs ARN, comme celle des protéines que nous traiterons dans la prochaine section, est complexe et unique. Les interactions faibles, spécialement l'empilement de bases, les interactions hydrophobes et les liaisons marginales jouent un rôle majeur dans la stabilité des structures [15].

1.4 Les protéines

Les protéines sont des chaînes d'acides aminés. Chacun de ces acides aminés est attaché à son voisin par une liaison covalente. En combinant uniquement les vingts différents acides aminés mentionnés dans le tableau 1.1, les cellules vivantes peuvent produire une variété innombrable de protéines ayant diverses propriétés et activités. Les différents produits résultant de ces constructions prennent la forme d'enzymes, d'hormones, d'anticorps, de poisons, d'antibiotiques, de pilosité et d'une foule d'autres substances[12]. Tous les 20 acides aminés ont en commun un groupement carboxyl et un groupement amine at-

les interactions entre les charges partielles des anneaux aromatiques adjacents [14]

Type d'acide aminé	Abréviation 3-lettre	Abréviation 1-lettre
A	Ala	alanine
C	Cys	cystéine
D	Asp	aspartate
E	Glu	glutamate
F	Phe	phénylalanine
G	Gly	glycine
H	His	histidine
I	Ile	isoleucine
K	Lys	lysine
L	Leu	leucine
M	Met	méthionine
N	Asn	asparagine
P	Pro	proline
Q	Gln	glutamine
R	Arg	arginine
S	Ser	sérine
T	Thr	thréonine
V	Val	valine
W	Trp	tryptophane
Y	Tyr	tyrosine

TAB. **1.1. 20 acides aminés différents et leur abréviation.**

tachés à un même carbone appelé le carbone α. Ils diffèrent par leur chaîne latérale (appelée également le groupement R) (voir la figure 1.3 pour la structure des acides aminés). Grâce à ce groupement, la structure, la taille, la charge électrique et la solubilité dans l'eau des acides aminés sont différentes [16]. La compréhension des propriétés chimiques des acides aminés est fondamentale pour analyser les interactions possibles de ceux-ci avec d'autres composés. Nous pouvons simplifier ce processus en regroupant les acides aminés par classes. Pour ce faire, il suffit de se baser sur les propriétés de leur groupement R, en particulier leur polarité[6]. Il y a cinq grandes classes d'acide aminés [17] :

- groupement R non-polaire et aliphatique : cette classe contient les résidus hydrophobe **alanine, glycine, isoleucine, leucine, proline et valine**.
- groupement R polaire non-chargé : cette classe contient les résidus hydrophile **asparagine, cystéine, glutamine, méthionine, sérine et thréonine**.
- groupement R aromatique (généralement non-chargé) : cette classe contient les résí-

[6]Tendance d'une molécule à interagir avec l'eau à un pH proche de pH 7.0.

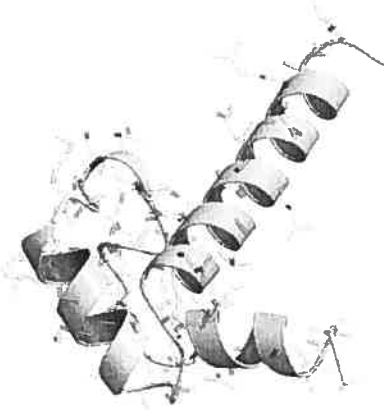


FIG. 1.4. **Représentation caractéristiques de la structure secondaire des acides aminés. La structure de cette protéine (code pdb : 1DUL) illustre les hélices α qui sont indiqués par les formes spiralées.**

dus ayant une chaîne latérale aromatique **phénilalanine, tryptophane et tyrosine**.

- groupement R chargé négativement : cette classe contient les résidus possédant une charge négative nette (à pH 7.0) **aspartate et glutamate**.
- groupement R chargé positivement : cette classe contient les résidus possédant une charge positive nette (à pH 7.0) **arginine, histidine et lysine**.

1.5 Structure des protéines

La structure moléculaire d'une protéine dicte la spécificité de ses fonctions et, par conséquent, indique les interactions possibles avec d'autres biomolécules. Dans le but de caractériser les éléments essentiels aux interactions de la protéine, il est donc nécessaire d'étudier la structure tridimensionnelle de celle-ci. Conceptuellement, la structure tridimensionnelle des protéines peut être constituée de quatres niveaux d'organisation que l'on nomme structure primaire, secondaire, tertiaire et quaternaire. Les trois premiers niveaux sont comparables à ceux des acides nucléiques (voir section 1.3). **La structure primaire** est définie par la séquence des acides aminés liés par liaisons covalentes. La structure est

dictée également par les contraintes stériques et les faibles interactions inter-moléculaires qui permettent des arrangements plus stables que d'autres. **La struture secondaire** est organisée en patrons réguliers, ce qui minimise l'énergie libre de la chaîne. Il y a quelques types communs de structures secondaires, tels les hélices α (figure 1.4)et les feuillets β. **La structure tertaire** permet de former des unités compactes et rapprochées appelées domaines. Ces entassements ont tendance à contraindre les portions hydrophobes à l'intérieur et à polariser la surface des protéines. Celles-ci deviennent ainsi plus stables dans un solvant comme l'eau. La stabilité de cette organisation est également favorisée par les interactions qui se créent entre certains résidus éloignés en séquence, mais spatialement proches grâce au repliement. Les protéines constituées de plusieurs chaînes polypeptidiques ont un niveau supplémentaire : **la structure quaternaire**, qui réfère aux relations spatiales de chacune des chaînes par rapport aux autres. Il est à noter que la structure d'une protéine n'est jamais définitive. Elle est plutôt dynamique et a tendance à s'adapter en fonction de plusieurs facteurs, comme le pH du solvant dans lequel elle baigne (pour de plus amples détails sur la structure des protéines se référer à [16]). Ceci porte à croire que dans certaines conditions, la structure peut offrir un site d'interaction qui pourrait être inaccessible dans une autre situation. Cet élément complexifie énormément l'analyse de la structure des protéines. Un niveau de complexité supplémentaire apparaît également lorsqu'il y a formation d'un complexe entre une protéine et un ARN appariés par diverses liaisons inter-moléculaires.

1.6 Les complexes ARN-protéine

Il existe une grande quantité d'ARN qui peuvent former des associations avec des protéines. Ces interactions sont essentielles à plusieurs processus biologiques, comme la biosynthèse des protéines, l'épissage de l'ARN et la réplication des virus (pour plus de détails sur les rôles joués par ces complexes, voir [18]. Notre compréhension des mécanismes d'inter-relation entre la séquence et la structure permettant de former des éléments susceptibles d'engendrer un complexe ARN-protéine est encore à ses balbutiements. Constam-

ment, de nouvelles structures tridimensionnelles de complexe ARN-protéine sont élucidées et nous permettent ainsi de faire des analyses comparatives approfondies des patrons de reconnaissance récurrents entre celles-ci. Ces patrons, que l'on peut nommer motifs structuraux, sont une construction régulière habituellement associée à des fonctions précises. Dans notre cas, ce sont des unités de reconnaissance entre les structures de l'ARN et de la protéine. Il existe plusieurs motifs découverts à ce jour qui semblent posséder des propriétés pour lier l'ARN de façon spécifique. Mais jusqu'à maintenant, la communauté scientifique n'a pu apporter de confirmation sur ce sujet [19, 20, 21, 22, 23].

Les protéines peuvent reconnaître et interagir avec certaines localisations spécifiques sur des éléments de la structure d'un ARN. Premièrement, la forme générale (ou le repliement) de la molécule est très importante. La forme de l'ARN doit permettre un bon positionnement de la protéine pour lui permettre d'accomplir ses fonctions. L'encombrement stérique[7] de l'ARN ne doit également pas emcombrer ces sites fonctionnels. Les éléments qui permettent un bon positionnement spatial de la protéine face à l'ARN et sa fixation sur la molécule d'ARN se retrouvent au niveau de la structure secondaire de cette protéine. Ils sont les résidus aromatiques individuels (qui créent des interactions de *Van der Waals* et des interactions hydrophobes), les résidus polaires ainsi que les résidus chargés. Les chaînes latérales aromatiques s'empilent sur les bases nucléotidiques, tandis que les résidus chargés positivement se lient avec le squelette[8] sucre-phosphate [24]. Deuxièmement, la protéine interagit avec des éléments spécifiques dans la séquence du nucléotide. Ces contacts sont formés grâce à des ponts hydrogène entre les résidus nucléotidiques et les chaînes latérales ou des portions du squelette des protéines [25]. Troisièmement, l'ARN est une molécule flexible et peut adopter une structure locale non-canonique. Cet aspect important de la reconnaissance des acides nucléiques par la protéine est appelé la déformation séquence-

[7]L'encombrement stérique est la configuration spatiale dans laquelle les atomes d'une molécule se repoussent par gène volumique. La réactivité d'un groupement fonctionnel au sein d'une molécule peut ainsi être atténuée par la présence de groupements voisins.[10]

[8]appelé également le *backbone*

dépendante. Cette caractéristique qu'adopte l'ARN favorise l'énergie libre du complexe [21].

En se basant sur les structures de complexe ARN-protéine existantes, il existe différentes stratégies de reconnaissance spécifique de sites d'ARN par des protéines. Les protéines à homéodomaine, les protéines "OB-fold" , les protéines de la famille "PNPase" et les protéines à motif RNP possèdent leurs propres caractéristiques de reconnaissance de l'ARN [26, 27, 28].

Malgré le peu de travaux qui a été réalisé jusqu'à maintenant dans le domaine des complexes ARN / protéine. Ce domaine deviendra de plus en plus important étant donné que ces complexes deviendront de nouvelles cibles thérapeutiques à certaines maladies pour l'instant incurables. Il devient essentiel de développer de nouvelles méthodes d'analyse des interactions moléculaire, de ce fait, la compréhension des mécanismes d'interaction occasionnera la prédiction précise des complexes potentiels entre les ARN et les protéines. Pour l'instant, le problème demeure entier et toutes les avenues de recherche qui sont exploitées offriront peut-être quelques réponses.

Chapitre 2

Stratégie

Depuis quelques années, le nombre de structures tridimensionnelles, autant de protéines que d'acides nucléiques, a augmenté de façon exponentielle [29]. Il y a maintenant un besoin de classer et d'exécuter des recherches efficacement à l'intérieur de toutes les sources d'informations [30]. De telles méthodes sont nécessaires afin d'aider les biologistes structuraux qui, il y a une dizaine d'années, devaient mémoriser les détails des quelques structures résolues à cette époque. Mais, avec la croissance du nombre de ces structures résolues, il est maintenant impossible de fonctionner d'une telle façon. De plus, l'information struturelle est désormais utilisée par une foule de scientifiques comme les biologistes moléculaires, les généticiens et les chimistes. Ces professionnels sont peu familiers avec les concepts complexes de la biologie structurale. Il arrive bien souvent que leur intérêt se porte sur l'analyse de petites portions des structures plutôt qu'au niveau du repliement général d'une structure.

2.1 Motifs structuraux

Des portions restreintes et localisées ayant possiblement un rôle biologique peuvent porter le nom de **motif**. Cette recherche vise la découverte de motifs structuraux récurrents bordant les points d'interaction entre la structure d'un acide ribonucléique (ARN) et d'une protéine, formant un complexe ribo-nucléique. Une grande similarité structurale suggère la possibilité qu'il s'agisse d'une région fonctionnelle [31, 16]. Un *motif* est une appellation générale donnée à une région possédant un caractère spécifique lié à une fonction ou simplement à une redondance. Ces motifs peuvent être retrouvés au niveau de

la séquence d'une biomolécule (plus communément appelé "pattern") ou, comme nous le verrons par la suite, à la structure secondaire ou tertiaire des ARN et des protéines. Les motifs peuvent être utilisés pour prédire une propriété fonctionnelle ou structurelle conservée d'une protéine. Ils peuvent également décrire le repliement non-trivial d'une structure qui permet de la différencier [32].

2.2 Revue des méthodes existantes

Les dernières années ont permis l'apparition et le développement de méthodes et d'outils sophistiqués permettant la comparaison de structures, la reconnaissance du repliement, la comparaison et la classification de différents repliements et ce, autant au niveau des protéines qu'au niveau des acides nucléiques [33, 34, 35, 36, 37]. Plus spécifiquement, les programmes RNAMOT [38] et SPASM [39] se rapprochent davantage de nos travaux. Le programme RNAMOT permet la recherche de portions conservées des séquences d'ARN. En d'autres termes, ce programme recherche des motifs planaires se retrouvant dans la structure secondaire des ARN. Il nécessite une connaissance du motif que l'on désire retrouver. L'exploration ne peut donc pas se faire de manière aveugle. Dans le cas de SPASM, qui s'applique spécifiquement aux protéines, la recherche se fait au niveau d'une configuration spatiale conservée et également prédéfinie. La recherche de similitude au niveau des interactions ARN-protéines est, encore aujourd'hui, à un niveau peu poussé en bioinformatique. Étant donné qu'il existe encore peu de structures résolues offrant la structure d'un ARN interagissant avec une protéines (environ 160 structures sont répertoriées), le besoin de développer des outils informatiques d'analyse demeure faible. En effet, de nos jours, celle-ci se fait manuellement ou par mémorisation. Les seuls développements qui ont été réalisés dans ce domaine se situent au niveau de l'analyse statistique des interactions [40]. Il existe également un visualisateur d'interactions structurelles entre une structure bidimensionnelle d'un ARN et les acides aminés s'y liant appelé *NUCPLOT* [41]. Tout reste encore à découvrir et si on se fie à l'explosion du nombre des bases de données de struc-

tures de protéines, le nombre de complexes ARN-protéine aura centuplé d'ici quelques années. D'autre part, peu d'efforts scientifiques ont été consentis dans le domaine avancé qu'est l'identification de motifs structuraux à l'aide de graphes relationnels (à l'exception des travaux qui sont effectués dans notre laboratoire hôte).

2.3 Définition du problème

Dans le cadre de ce mémoire, nous avons voulu répondre au problème biologique à savoir s'il existe des facteurs de reconnaissance qui permettent de prédire l'agencement des ARN et des protéines entre eux. La formation de complexes ARN-protéine semblent se réaliser grâce à des portions spatiales compatibles qui forment des ensembles d'intéractions spécifiques. Il n'existe pas de moyen automatique de recherche de ces ensembles d'interactions (que nous nommerons motifs) dans les structures résolues jusqu'à présent. Nous avons tenté de solutionner ce problème en élaborant un programme qui **détermine les régions structurales répétitives** au sein d'un groupe de complexes ARN-protéine.

2.4 Influence des travaux antérieurs

Pour accomplir le but fixé (voir la section précédente), nous nous sommes inspirés du travail réalisé par P. Gendron [42] sur la recherche de motifs structuraux dans l'ARN. Nous avons modifié cette méthode afin de pouvoir annexer les protéines au modèle. Le "moteur" principal du programme, soit l'algorithme de construction de motif, a également été changé afin d'en améliorer son efficacité. Notre précurseur avait implanté son programme ainsi : La méthode de recherche de motifs se fait en deux étapes séquentielles qui consistent en l'énumération des sous-graphes possibles ainsi que la classification isomorphique de ces sous-graphes. Le travail de P. Gendron s'inspire de l'algorithme d'isomorphisme d'Ullman [43] pour identifier et compiler les motifs récurrents.

Dans les sections suivantes, nous traiterons exclusivement des modifications que nous avons apportées afin de réaliser notre programme. Ces modifications ont trait à l'ajout d'un

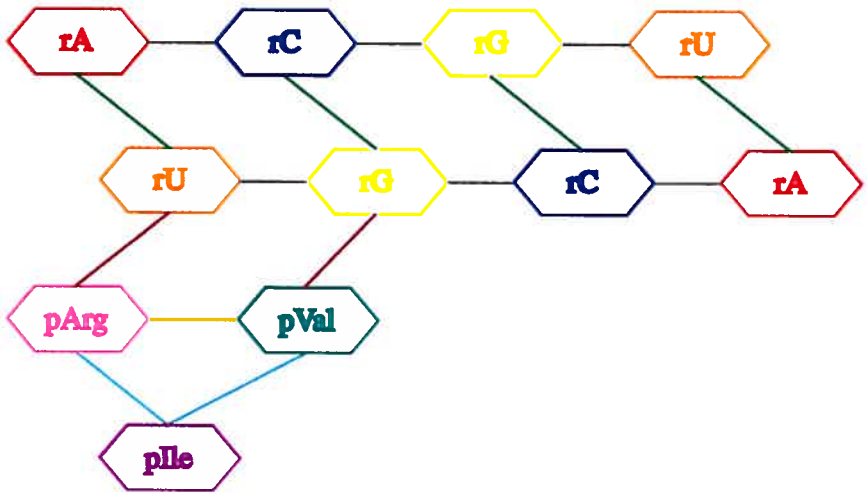


FIG. 2.1. **Représentation d'une structure sous forme d'un graphe relationnel. Les différents types de noeuds et d'arcs du graphe sont indiqués par des couleurs différentes.**

graphe de protéines pour la représentation de la structure, à l'implantation d'un tout nouvel algorithme de recherche de sous-graphe maximal commun et au classement des motifs obtenus.

2.5 Représentation des structures

La structure tridimensionnelle des unités est décomposée en un graphe relationnel non-dirigé à deux dimensions, G = (N,A), où les noeuds (N) sont représentés par les résidus des molécules, et les arcs (A) sont définis par les relations existants entre les résidus. Un exemple est illustré à la figure 2.1. À partir d'une structure tridimensionnelle, il est possible d'extraire les informations liées à sa structure secondaire et de les encoder dans un graphe. Nous considérons l'alphabet de l'ARN et celui des acides aminés comme l'ensemble des noeuds possibles du graphe. On utilisera la notation $p*$ pour symboliser un acide aminé et $r*$ pour symboliser un résidu d'ARN. Les relations entre les noeuds sont symbolisées par les différents types de liaisons hydrogène correspondant aux liens inter-résidus. Ces

informations sont tirées du fichier de la représentation structurale de la molécule ayant comme extension .pdb. Il existe plusieurs banques de données publiques qui entreposent ce type de fichier [44, 45, 46]. Les valeurs de ce fichier indiquent les coordonnées spatiales des différents atomes retrouvés dans la molécule. En utilisant la librairie *mccore* développée au laboratoire de F. Major (http ://sourceforge.net/projects/mccore/), il est ensuite possible de regrouper ces atomes en résidus. Notre tâche a été de représenter ces résidus sous la forme d'un graphe à deux dimensions. Les arcs sont indiqués différemment selon leur représentation : deux nucléotides adjacents en séquence (lien phosphodiester) ou deux nucléotides adjacents en appariement (liaison Watson-Crick).

La structure secondaire des protéines est également représentée dans le graphe, à savoir les liaisons hydrogène existant entre les acides aminés d'une hélice alpha ou d'un feuillet beta. Les données relationnelles de la structure tertiaire des protéines ont été extraites à l'aide du logiciel DSSP [47]. Ce programme permet d'extraire la structure secondaire des protéines à partir des propriétés géométriques de chacun des acides aminés et de la possibilité de liaison entre certains d'entres eux. Pour chaque résidu, le programme définit la possibilité d'une liaison avec un autre résidu en considérant un certain seuil d'acceptabilité.

Les arcs peuvent également être définis par une liaison hydrogène que nous dénotons différemment : interaction ARN-protéine, entre un acide aminé et un ribonucléotide (voir la figure 2.2). Les interactions ARN-protéine ont été déduites de la distance entre l'atome receveur et l'atome donneur de chacun des résidus selon une marge personnellement prédéfinie (en Armstrong) et généralement acceptée par la littérature [15]. En gardant en perspective que le but de cette recherche est purement exploratoire au niveau des interactions ARN-protéine, nous ne pouvons nous permettre d'omettre toutes les possibilités de liaisons intermoléculaires. De ce fait, il se peut que l'on forme des arcs croisés d'interactions entre une paire d'acides aminés et une paire de résidus d'ARN[1]. Il en découle alors la formation possible d'un graphe non-planaire. Ceci occasionne une restriction pour le choix d'un al-

[1] même s'il est impossible que ce phénomène se produise *in vivo*, nous devons considérer tous les cas possibles d'interaction.

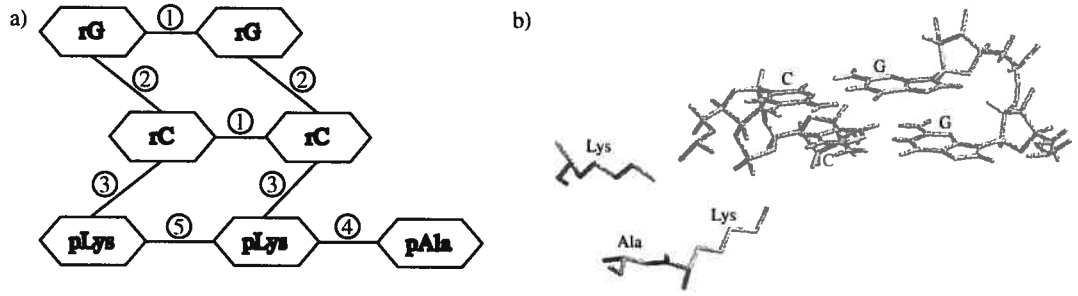


FIG. **2.2. Graphe relationnel associé à une structure tridimensionnelle spécifique. Le graphe relationnel (en a) est caractérisé par différent types d'arcs qui correspondent aux différentes liaisons que l'on retrouve dans la structure tridimensionnelle (en b). Les types d'arc retrouvés sont : [1] Adjacence (nucléotide - nucléotide), [2] Liaison Watson-Crick, [3] Liaison d'interaction (nucléotide - acide aminé), [4] Liaison structure secondaire (hélice *alpha* ou feuillet *beta*), [5] Adjacence (acide aminé - acide aminé).**

gorithme de recherche de motif qui s'implante uniquement sur des graphes planaires.

2.6 Recherche de motifs récurrents (Notre stratégie employée)

Le but algorithmique est de trouver des sous-graphes communs récurrents. Un sous-graphe d'un graphe G est un graphe dont les sommets et les arêtes sont des sommets et des arêtes de G. Ce sous-graphe possède un sous-ensemble n de noeuds dans G et un sous-ensemble a d'arcs reliant ces noeuds ($n \subseteq$ N et $a \subseteq n$ x n). Une paire de sous-graphes communs de deux graphes G1 et G2 consiste en un sous-graphe H1 de G1 et un sous-graphe H2 de G2, tel que H1 est isomorphe à H2. L'isomorphisme correspond à la possibilité que deux graphes se superposent de façon à ce que leurs structures soient semblables [48]. Le problème de l'isomorphisme de graphes et les méthodes pouvant le solutionner ont largement été discutés dans la littérature computationnelle [43]. Il est à noter que les graphes construits ici ne seront pas nécessairement planaires, comme spécifié précédemment. Les approches linéaires permettant de traiter l'isomorphisme n'ont donc

pu être utilisées [49]. Il en va de même avec l'approche élaborée par P. Gendron pour trouver les sous-graphes communs, étant donné que cette méthode calcule l'isomorphisme des graphes construits.

Plusieurs éléments stratégiques ont été apportés à notre programme afin de le rendre efficace et appliquable à notre contexte de recherche. Notre recherche de motifs au niveau des interactions ARN-protéine s'oriente vers une vision exploratoire, car très peu de résultats ont été obtenus dans ce domaine jusqu'à présent. Nous avons développé un algorithme qui ne délimite sa recherche de motifs qu'aux régions bordant les points de jonction entre l'ARN et la protéine. Il s'agit plutôt d'extraire tous les sous-graphes isomorphes à l'intérieur de différents graphes données. Ceci produira un ensemble de différentes paires de graphes. L'extraction se réalisera à l'aide d'un algorithme de calcul des sous-graphes maximaux (présenté au chapitre 3). La puissance de l'algorithme est issue par la modification d'un algorithme de recherche en profondeur. Il permet, par conséquent, d'obtenir tous les sous-graphes semblables d'un ensemble de graphes différents contenant des liens d'interaction entre un résidu d'ARN et un acide aminé. Selon notre implantation algorithmique, la méthode fixe plusieurs origines de départ et ensuite elle tente d'étendre la région au maximum en comparant toujours avec un autre endroit dans un autre graphe. Par la suite, ces sous-graphes maximaux sont extraits et une classification s'effectue afin de retrouver les graphes récurrents. Grâce à certains principes simples de la loi des ensembles, il est possible de regrouper les graphes de façon à permettre la découverte de régions conservées. Une comparaison au niveau de la RMSD [50] des paires de structures exclut certaines structures trop dissemblables. Le triage qui est exécuté au terme de l'analyse est traité à la fin du chapitre suivant.

Une partie de l'analyse doit se faire manuellement. Lorsque les ensembles de graphes sont rapportés sur la structure tridimensionnelle de la molécule et qu'ils sont analysés individuellement, il en ressort des récurrences structurales aux propriétés intéressantes. Une analyse plus poussée est alors nécessaire afin de retracer si ces sous-structures peuvent se retrouver à l'intérieur de simples structures d'ARN ou de protéine.

Chapitre 3

Nos algorithmes

Les algorithmes de correspondance de graphes (*graph matching*) ont été largement étudiés dans différents domaines. Ceux qui nous intéressent plus particulièrement dans le contexte de notre approche se réfèrent au problème de la détermination du sous-graphe maximal de deux graphes. Afin de clarifier la situation, notre approche algorithmique doit permettre d'extraire tous les sous-graphes communs d'un ensemble de graphes. Chaque sous-graphes correspondant au même sous-graphe commun doit contenir au moins une interaction ARN-protéine et le même nombre de noeuds. À la fin de l'exécution, nous obtenons tous les différents sous-graphes répartis en groupes possédant un sous-graphe commun. En ce sens, il s'agit d'implanter un algorithme qui recherche les sous-graphes communs maximaux ("locaux" aux interactions ARN-protéine) dans un ensemble de graphes et qui les regroupe selon les similarités qui les composent (noeuds et arcs).

Levi [51] et McGregor [52] ont déjà apporté quelques solutions à ce problème, mais leur solution est impratiquable dans notre situation. Grâce à leurs algorithmes, on obtient l'unique sous-graphe commun maximal. De plus, il devient très pénible d'utiliser ces algorithmes lorsque le nombre de noeuds $p1$ du premier graphe est largement supérieur au nombre de noeud $p2$ du deuxième graphe. Un autre inconvénient est que ces algorithmes fonctionnent comme à l'habitude, c'est-à-dire avec des matrices d'adjacence pour représenter les graphes (taille $p1 \; X \; p1$, $p2 \; X \; p2$, ... où pN représente le nombre d'éléments composant le graphe N). L'empilement de copies de matrices représentant les différentes structures à analyser (>100) devient alors fastidieux et impossible à réaliser lors de leur exécution. L'utilisation de ces algorithmes a été limitée au domaine de la chimie organique et au regroupement des mécanismes réactionnels possédant des "patterns" identiques.

Les algorithmes classiques ne permettent pas d'obtenir le même résultat que ceux obtenus avec notre algorithme, parce que leur but recherché est quelque peu différent du nôtre. Donnons la définition usuelle du terme sous-graphe maximal commun :

> Le sous-graphe maximal commun de deux graphes est le graphe contenant le plus grand nombre de noeuds communs possibles reliés par les même arcs (voir la figure 3.1).

Il est important de souligner qu'on ne calcule pas un sous-graphe maximal général entre deux graphes, mais bien plusieurs sous-graphes maximaux dans les localités des interactions ARN-protéine. Ce léger détail nous empêche d'implanter les méthodes traditionnelles. Nous devons également ajouter à la définition de sous-graphes maximaux un critère d'acceptation : Ceux-ci doivent posséder une paire de noeuds de départ spécifique. Cette paire de départ est composée d'une interaction formée d'un ARN et d'un acide aminé. Notre programme détermine préalablement les paires de noeuds d'interaction **équivalentes**[1] et les associe ensemble afin de démarrer l'algorithme à ces endroits précis.

Le principe fondamental de l'algorithme est d'augmenter le plus possible, en partant d'un point d'origine prédéterminé, le nombre de noeuds équivalents dans chacun des graphes à analyser. Pour chaque point de départ, l'algorithme génère deux sous-graphes (G1, G2) isomorphes ($G1 \simeq G2$). Il est à noter que même si les sous-graphes semblent adopter la même forme, cette représentation est faite en deux dimensions et la structure tridimensionnelle reportée de ces graphes peut être visuellement différente. Il ne suffit pas d'obtenir un sous-graphe commun pour déduire qu'il y a présence d'une structure semblable. L'équivalence au niveau des sous-graphes relationnels indique uniquement la possibilité que la structure tridimensionnelle associée soit spacialement correspondante.

Le problème consistant à trouver la correspondance entre les noeuds des deux graphes

[1]Le terme équivalent fait référence au domaine du "graph matching" indiquant que nous tentons de faire correspondre, à chaque itération, un noeud avec un autre. Dans le cas où l'étiquette des deux noeuds à faire correspondre et l'étiquette de l'arc les reliant sont semblables, il y a équivalence.

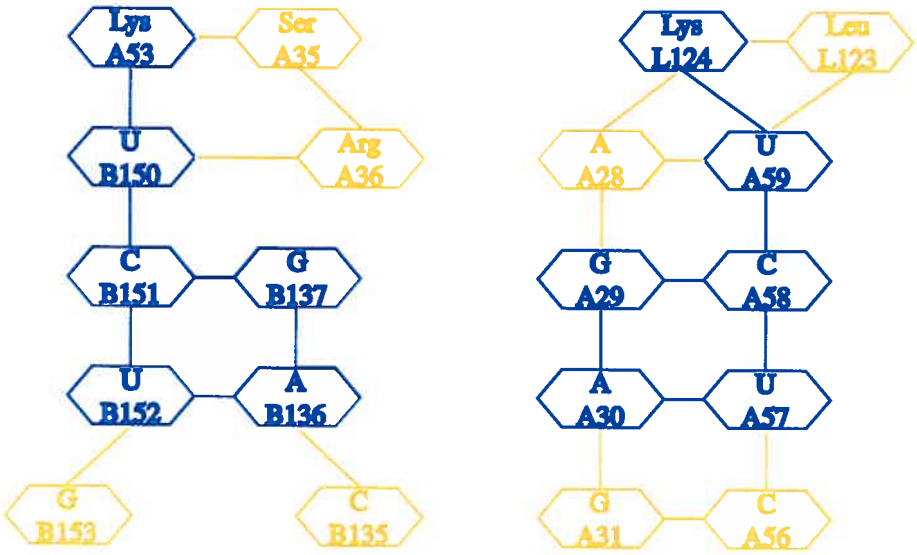


FIG. **3.1. Le sous-graphe maximal commun des deux graphes illustrés est représenté en bleu.**

(en satisfaisant leurs étiquettes[2]) a été élucidé par un algorithme de "backtracking" modifié. Les principales modifications de l'algorithme entraînent le parcours successif de certains segments du graphe au sein d'une même recherche, contrairement à un "backtrack" traditionnel où l'intégralité d'une représentation n'est parcourue qu'une seule fois.

Malheureusement, cet algorithme demeure inefficace dans la mesure où il est appliqué sur des noeuds et des arcs non étiquetés et que ces noeuds possèdent un grand nombre d'arcs pouvant relier d'autres noeuds. Il est utile de préciser que le facteur maximal de branchement observé dans nos situations est inférieur à six et que les noeuds et arcs des graphes sont fortement typés. Cette précision permet d'augmenter grandement l'efficacité de notre algorithme dans notre cas précis. Des analyses de performance sont calculées à la section **4.6**.

La figure 3.2 démontre les étapes essentielles de notre algorithme qui extrait le sous-

[2]Le terme étiquette signifie le type représentatif d'un noeud ou d'un arc.

graphe maximal de deux graphes simplistes. Le "backtrack" s'effectue sur une recherche en profondeur. L'algorithme génère, à chaque ajout de noeuds, les différentes listes de noeuds connexes potentielles. Ces listes de noeuds connexes, que l'on peut appeler chemins, sont formées d'un ensemble de noeuds interconnectés ne formant pas de boucles. Pour chaque liste, les chemins sont priorisés selon différents critères (la taille en noeuds du chemin et/ou la représentation tridimensionnelle du chemin). Une fois qu'un chemin est conservé, on recalcule une fois de plus la liste des chemins potentiels en s'assurant de ne pas passer par les noeuds du chemin conservé. Cette étape est répétée jusqu'à l'épuisement des possibilités. Le sous-graphe maximal est alors l'addition de tous les chemins conservés. Cette méthode s'assure d'extraire les sous-graphes maximaux possédant comme origine deux paires équivalentes d'interactions ARN-protéine en joignant plusieurs chemins optimaux et en évitant que ceux-ci se croisent.

3.1 Description de l'algorithme implanté

Nous présentons maintenant le pseudo-code de l'algorithme (voir l'algorithme 1) permettant d'extraire les sous-graphes maximaux locaux par récursion. Le premier appel de la fonction *MOTEURMOTIF* reçoit comme arguments deux origines de départ, correspondant chacune à une paire de noeuds (un résidu acide aminé et un résidu ARN relié par un arc d'interaction). Une recherche en profondeur s'effectue ensuite sur tous les noeuds adjacents aux noeuds courants. Il est essentiel de marquer les noeuds qui ont déjà été utilisés afin d'empêcher l'algorithme de boucler indéfiniment. Le test à la ligne **12** permet de mettre en relation les noeuds adjacents selon le facteur de comparaison recherché. Dans notre programme, l'usager peut définir le type de correspondance qu'il veut faire entre les résidus. Il existe plusieurs niveaux de correspondance, le plus strict étant la correspondance exacte entre résidus du même type, mais il peut également y avoir une correspondance entre

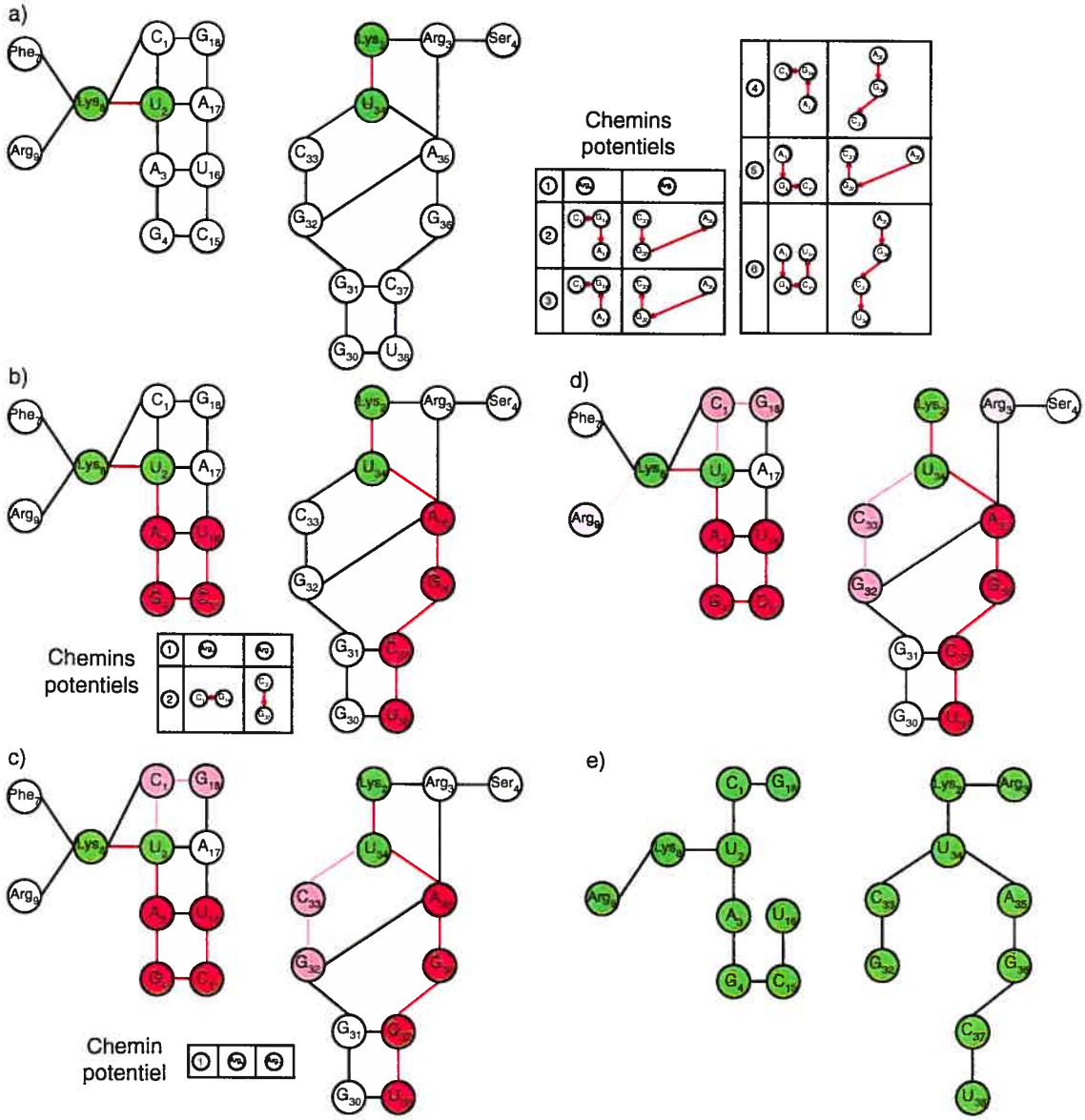


FIG. 3.2. **Schéma des différentes étapes de l'algorithme du calcul des sous-graphes communs maximaux. Les deux graphes sont comparés entre eux et l'on fait correspondre les noeuds d'un graphe avec ceux de l'autre ("graph matching"). a) Interaction de départ de l'algorithme (en vert) et chemins potentiels pouvant étendre le sous-graphe. b) Un chemin optimal est sélectionné (en rouge) et marqué afin de ne plus repasser par ce chemin. Dans ce cas-ci, le chemin optimal est le chemin possédant le plus de noeuds. L'algorithme recalcule les chemins disponibles de nouveau en évitant de repasser par le chemin conservé. c) Le chemin optimal est conservé (en rose) et les chemins potentiels sont recalculés. d) Tous les chemins conservés sont additionnés pour former deux sous-graphes communs maximaux. e) Sous-graphes maximaux obtenus par l'algorithme en 1.**

différentes familles de résidus[3].

Lorsque cette première étape de recherche en profondeur a été réalisée, nous obtenons une liste de chemins possibles pour chaque paire de résidus correspondants. À la ligne **21**, la méthode TSMP (trouver chemin au meilleur potentiel) permet d'indiquer quel chemin sera sélectionné préférentiellement parmi la liste de chemins possibles. Nous privilégions toujours les chemins possédant le plus grand nombre de noeuds et, selon le cas, celui qui correspond à la structure tridimensionnelle la plus semblable. Ce chemin au meilleur potentiel est ensuite marqué et la récursion s'exécute encore sur les mêmes résidus. Cette façon de faire permet de garantir la sélection des meilleurs chemins restants à chaque étape. L'algorithme se termine lorsque toutes les possibilités restantes (noeuds adjacents) ont été explorées et que, pour chacune, le meilleur chemin a été gardé, en évitant que celui-ci intersecte un chemin précédemment trouvé. Lorsqu'il y a dépilement, chaque chemin est additionné au précédent afin de trouver le sous-graphe contenant le plus de noeuds.

En utilisant cette approche, il devient alors facile de lancer l'algorithme sur toutes les combinaisons de paires d'interactions correspondantes possibles et d'en extraire les différents motifs. L'algorithme est également indifférent à la structure soumise et au nombre de noeuds à analyser. Ainsi, il est possible de lancer le programme sur la totalité des structures disponibles et de faire une analyse d'ensemble. Cela permet aussi d'analyser de façon intrinsèque les sous-graphes maximaux locaux présents dans un seul graphe.

Bien entendu, le calcul du chemin au meilleur potentiel est l'étape critique de notre approche. Nous aurions pu opter pour une autre façon de prioriser les chemins, mais cela aurait donné un tout autre comportement à l'algorithme. Nous avons essayé de prioriser les chemins aux nombres d'arcs correspondants plus grands, mais de cette façon, les résultats étaient moins significatifs. Notre but étant d'obtenir la plus grande région com-

[3]Nous avons implanté dans notre programme les regroupements les plus fréquemment utilisés. Ces familles font référence aux purines et pyrimidines pour les nucléotides [15] et aux groupements polaires, non-polaires, chargés positivement et négativement chez les acides aminés [10]. Il est également possible de différencer les atomes impliqués dans l'interaction et de permettre une discrimination à partir de ce détail.

mune possédant le plus de résidus possible, il faut par conséquent employer la méthode telle qu'elle a été décrite précédemment.

Il est possible d'accélérer significativement la recherche en passant outre certains points d'origine. Les origines qui étaient associées ensemble dans un motif précédemment trouvé permettront de trouver ce même motif. Il devient alors inutile de calculer le même graphe.

Nous avons découvert par la suite que notre algorithme nous permettait également de rechercher efficacement des sous-graphes prédéfinis à l'intérieur d'autres graphes. Nous avons utilisé cette modification afin de rechercher les occurences de certains sous-graphes préalablement trouvés à l'intérieur d'autres structures. Il s'agit de lancer la recherche sur les paires formées d'un noeud du premier graphe (un résidus du motif à rechercher) et des noeuds correspondants des graphes à analyser. Les sous-graphes possédant la même taille que le premier graphe sont conservés et forment ainsi des occurences différentes du motif recherché.

3.2 Analyse et tri des sous-graphes générés

Une étape essentielle se produit subséquemment à la découverte des sous-graphes. Le tri de ces derniers est effectué, dans le but de faciliter leur analyse manuelle. Étant donné que tous les sous-graphes sont appariés, il devient évident d'appliquer les lois simples des ensembles. Supposons que nous voulons comparer les sous-graphes G1 et G1' avec les sous-graphes G2 et G2'. Si le sous-graphe G1' contient les même noeuds que le sous-graphe G2, il est probable que la structure tridimensionnelle de ces quatres sous-graphes soit sensiblement semblable. De cette facon, il est possible de créer des regroupements entre les différents sous-graphes trouvés. Il faut comprendre qu'étant donné que nos sous-graphes sont transposés sur la structure tridimensionnelle, il ne devient plus nécessaire de vérifier l'isomorphisme des graphes. Nous ne sommes intéressés uniquement qu'aux noeuds qui le constituent.

Lorsque nous obtenons un nombre important de sous-graphes à vérifier manuellement,

il devient très utile de comparer les paires de structures tridimensionnelles à l'aide du calcul de distance basé sur la RMSD[4]. En deça d'un certain seuil, les structures sont rejetées, car elles demeurent trop dissemblables pour correspondre à un motif. Le calcul de la RMSD entre deux structures de même taille (possédant le même nombre d'atome) peut être décrit ainsi :

$$RMSD = \frac{\sum_{i=0}^{n} \left((x_i^{struct1} - x_i^{struct2})^2 + (y_i^{struct1} - y_i^{struct2})^2 + (z_i^{struct1} - z_i^{struct2})^2 \right)}{N}$$

où :

- n correspond au nombre d'atomes retrouvés dans chacune des structures.
- $x^{structX}$ représente les coordonnées en x de chacun des atomes correspondants sur leur structure respective (1 ou 2).
- $y^{structX}$ représente les coordonnées en y de chacun des atomes correspondants sur leur structure respective (1 ou 2).
- $z^{structX}$ représente les coordonnées en z de chacun des atomes correspondants sur leur structure respective (1 ou 2).
- N est le nombre d'atomes contenus dans la structure.

[4]La RMSD (*Root mean square deviation*) est une sommation des différentes distances entre chacun des atomes correspondants. Nous avons employé un algorithme qui permet d'aligner préalablement les atomes, afin de minimiser les distances avant de les additionner.

Algorithme 1: Algorithme des sous-graphes maximaux locaux

```
1  MOTEURMOTIF ( noeudCourant1, noeudCourant2 );
2  Données : listeDeChemin
3  Données : cheminRetour
4  si noeudCourant1 == noeudCourant2 alors retourner [cheminRetour];
5  pour chaque noeud adjacent au noeudCourant1 = noeudAdjacent1 faire
6      si noeudAdjacent1 est marqué alors Prendre le noeud adjacent suivant;
8      Marquer le noeud noeudAdjacent1;
9      pour chaque noeud adjacent au noeudCourant2 = noeudAdjacent2 faire
10         si noeudAdjacent2 est marqué alors Prendre le noeud adjacent suivant;
11         Marquer le noeud noeudAdjacent2;
12         si VérifieIdentité( noeudAdjacent1, noeudAdjacent2 );
           alors
13             Données : cheminPresent = chemin(noeudAdjacent1,noeudAdjacent2)
14             cheminPresent += MOTEURMOTIF (noeudAdjacent1,noeudAdjacent2);
15             Ajouter à la liste de chemin ( cheminPresent );
16
17         Effacer la marque sur noeudAdjacent2;
18
19     Effacer la marque sur noeudAdjacent1;
20
21 TSMP ( listeDeChemin ) ; Calculer le chemin ayant le meilleur potentiel ; Marquer
   le chemin Maximal atteint dans la liste de chemin;
22 cheminRetour = MOTEURMOTIF ( noeudCourant1, noeudCourant2 );
23 Effacer la marque sur chemin Maximal;
24 retourner [cheminRetour];
```

Chapitre 4

Résultats

Dans ce chapitre, il sera question de plusieurs aspects des interactions entre les ARN et les protéines. Nous avons tenté d'élargir notre champ de connaissance dans ce domaine, qui était quelque peu nouveau pour nous. Les interactions de contacts, les motifs qui en découlent et la comparaison de l'algorithme avec ce qui existe déjà seront les principaux points traités.

Dans l'ensemble des bases de données publiques, nous avons répertorié 163 complexes ARN-protéine. Toutes ces structures ont été résolues par crystallographie ou par RMN. Les analyses effectuées dans ce projet n'ont utilisé qu'une cinquantaine de ces structures. Nous avons rejeté certains complexes qui étaient redondants, c¡est-à-dire qui correspondent à des structures de la même molécule, résolus, par exemple, dans d'autres laboratoires. D'autres analyses structurales ne possédaient pas la résolution de définition limite que nous nous étions imposés (<3.2 A). Les codes PDB des structures conservées et utilisées sont indiquées dans le tableau 4.1.

4.1 Analyse statistique des points d'interaction

Tout d'abord, nous avons fait des analyses statistiques simples afin de déterminer la morphologie des interactions ARN-protéine. Nous avons extrait toutes les interactions des 50 structures et avons répertorié celles-ci de façon à rechercher des propriétés particulières. Comme l'indique le tableau 4.2, il semble que les types d'acides aminés préférés au contact avec les ribonucléotides sont Arg, Gly et Lys. Ceux qui sont le moins souvent retrouvés sont Asp, Cys, Ile, Met, Phe et Trp. De là, il est possible de retrouver les contributions des

1A34	1A4T	1A9N	1AJU	1AKX
1AQ4	1ARJ	1ASY	1ASZ	1AUD
1AV6	1B23	1B7F	1C9S	1CK5
1CK8	1CN9	1CVJ	1CWP	1CX0
1D6K	1D9F	1DFU	1DI2	1DK1
1DRZ	1DUL	1DZS	1EC6	1EIY
1EKZ	1ETF	1EUY	1EXD	1FJF
1G1X	1GTR	1KOC	1MMS	1QA6
1QF6	1QFQ	1QRS	1QTQ	1TTT
1URN	1ZDI	1ZDK	2A8V	2BBV

TAB. 4.1. **Structures utilisées pour l'analyse des complexes ARN-protéine**

différents groupes d'acide aminés : hydrophobe 20.53%, chargé 43.78%, polaire 28.57% et la glycine (Gly) 7.1%. Ces résultats sont confirmés par ceux obtenus par Westhof et al. [40]. Du côté de l'ARN, la guanine (G) est favorisée tandis que l'uracile (U) est délaissée dans la formation des interactions.

4.2 Disposition des acides aminés autour des nucléotides

En observant plus particulièrement le profil type d'une interaction spécifique, nous déduisons la place préférentielle qu'occupe un acide aminé autour d'un résidu d'ARN. Pour le cas le plus représenté d'interaction, le couple Arg-G, la morphologie de l'interaction ressemble à celle montrée a la figure 4.1 et semble indiquer que le groupement phosphate est préféré au ribose, qui est préféré à la base du nucléotide. Cela semble être causé par le fait que la majorité des résidus d'ARN sont retrouvés en appariement Watson-Crick. Dans le cas des autres résidus, cette morphologie semble également conservée, excepté dans le cas de Pro et Asn, qui préférent la base au ribose et au groupement phosphate. Les interactions semblent se lier sur le "backbone" de l'ARN sans discrimination de la base impliquée et

	Ribonucléotides				
Acides aminés	A	C	G	U	Total
ALA	1,08	0,85	1,39	0,39	3,71
ARG	4,40	8,57	7,64	3,78	24,40
ASN	1,00	1,00	1,31	0,62	3,94
ASP	0,23	0,39	0,54	0,46	1,62
CYS	0,23	0,08	0,31	0,15	0,77
GLN	0,46	0,77	1,62	0,62	3,47
GLU	0,69	0,93	0,77	0,15	2,55
GLY	1,62	1,70	2,78	1,00	7,10
HIS	0,77	0,85	1,54	0,39	3,55
ILE	0,46	0,31	0,85	0,46	2,08
LEU	0,77	0,85	1,24	0,62	3,47
LYS	3,63	4,56	5,25	1,78	15,21
MET	0,31	0,39	0,69	0,23	1,62
PHE	0,46	0,54	0,77	0,39	2,16
PRO	1,24	0,93	1,62	0,62	4,40
SER	1,08	1,85	1,85	0,62	5,41
THR	1,47	1,31	1,85	0,54	5,17
TRP	0,31	0,31	0,46	0,39	1,47
TYR	1,39	1,62	1,08	0,69	4,79
VAL	0,62	1,16	0,69	0,62	3,09
Total	22,2	28,9	34,2	14,52	100,00

TAB. **4.2. Tableau de la composition des interactions ARN-protéine (% du nombre d'acides aminés en fonction du nombre de résidus d'ARN répertorié)**

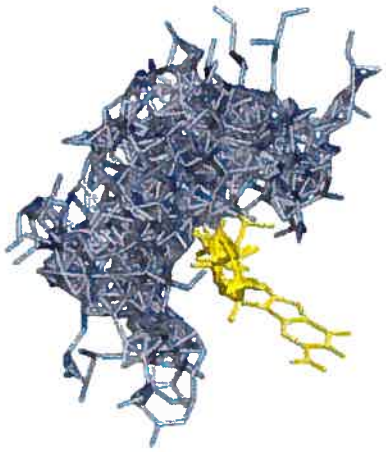


FIG. **4.1. Figure représentant la disposition de l'acide aminé arginine autour du nucléotide guanine retrouvé dans les structures analysées (voir tableau 4.2). Le résidu d'ARN représenté en jaune est un ensemble de guanines superposés autour desquels chaque arginine y étant liée figure en bleu. Cet ensemble d'interactions guanine-arginine ont été produit en les extrayant de différentes structures (voir tableau 4.1).**

indique le peu de spécificité des acides aminés pour les résidus d'ARN. Autrement dit, tous les résidus d'ARN possèdent un sucre et un groupement phosphate sur leur squelette où se lient majoritairement les acides aminés.

Le programme que nous avons élaboré permet la recherche de motifs selon une grande quantité de paramètres prédéfinis. La pécularité atomique des interactions, le type d'appariement (selon les différents groupes de résidus) et la taille maximale de recherche peuvent être adaptés. Le bon agencement de ceux-ci peut amener à la découverte de nouveaux motifs structuraux, car l'éventail de contextes de recherche est très large.

4.3 Analyse de la sous-unité ribosomale 30S et de ses protéines

Dans un premier temps, nous avons porté notre analyse sur l'énorme structure tridimensionnelle de la sous-unité ribosomale 30S [53] qui peut être observée avec ses protéines

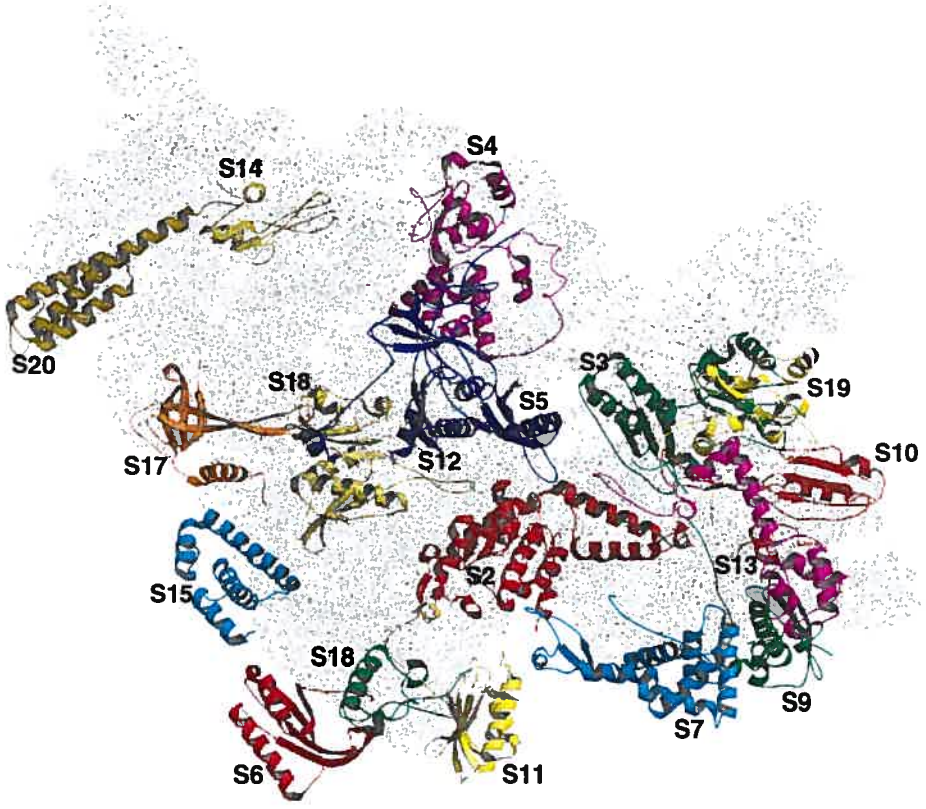


FIG. 4.2. **Structure de la sous-unité ribosomale 30S et les différentes protéines s'y liant. Chaque couleur identifie les protéines différentes fixées au ribosomes. L'ARN du ribosome est représenté dans les teintes de gris.**

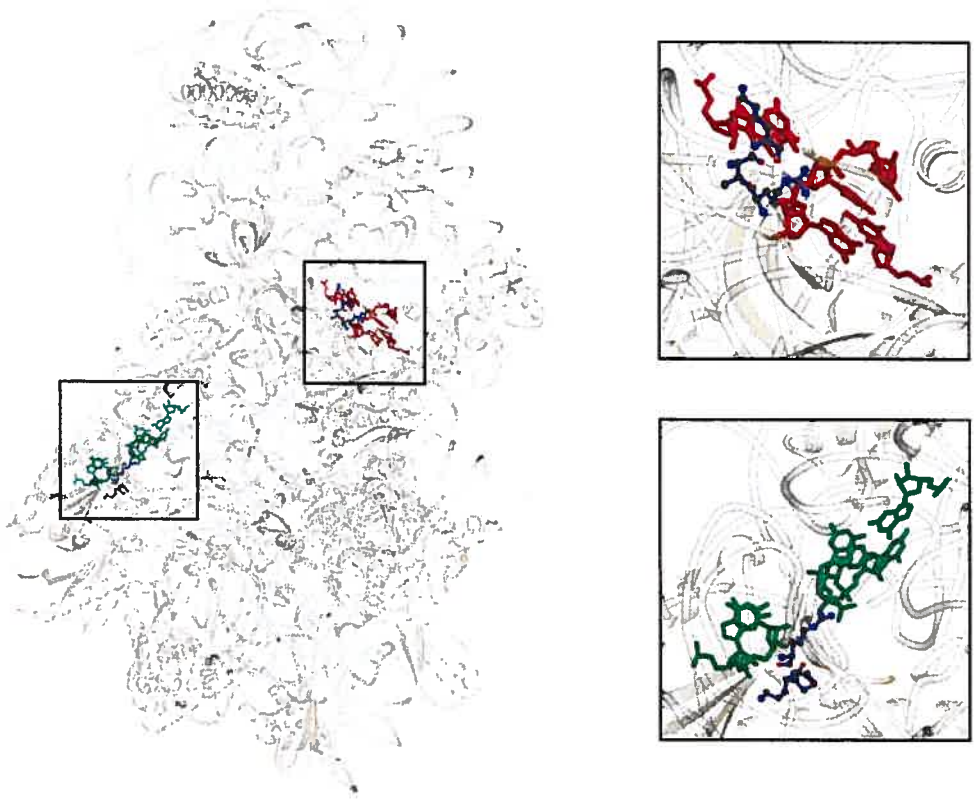


FIG. **4.3. Emplacement des motifs conservés retrouvés dans la structures de la sous-unité 30S ribosomale. En rouge et vert, les résidus d'ARN sont représentés. Les résidus d'acides aminés (lysine et arginine) figurent également en noir.**

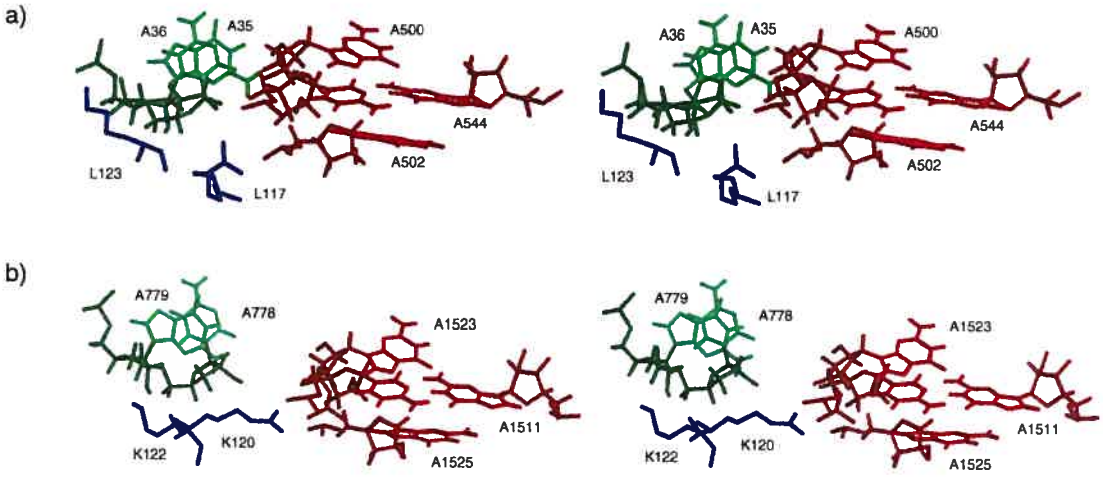


FIG. **4.4. Vue stéréoscopique du motif retrouvé (celui en a représente le motif en rouge et celui en b représente le motif en vert) tel que démontré selon la figure 4.3.**

ribosomales à la figure 4.2. Nous avons tenté de découvrir de nouveaux motifs de structure en élaborant une recherche très contraignante sur les types d'appariement entre les résidus et les résidus eux-même. Après une analyse minutieuse des résultats, nous avons extrait une structure conservée apparaissant deux fois au sein de la même structure globale (voir figure 4.3. Ce motif, qui ne ne semble ne pas avoir été documenté, est illustré à la figure 4.4. Nous pouvons observer qu'il fait interagir six nucléotides organisés en deux double-hélices qui se croisent perpendiculairement, ainsi que deux acides aminés présents dans une boucle protéique sans configuration particulière. Ces acides aminés se logent directement entre les double-hélices, suggérant la possibilité d'une interaction de stabilisation entre les deux molécules. Comme dans la plupart des cas, l'interaction se situe entre le squelette de l'ARN et la chaîne latérale des acides aminés (Arg et Lys). Ceci suggère le peu de spécificité que possède l'interaction au niveau des résidus. Cela porte à croire que la spécificité possède une caractéristique de niveau supérieur et donc les structures secondaire et tertaire auraient une influence déterminante dans ce motif.

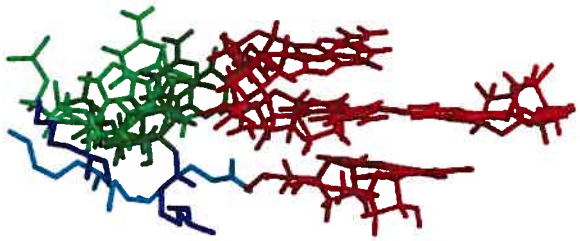


FIG. **4.5. Superpositon des deux structures du motif redondant retrouvé dans la sous-unité 30S du ribosome.**

En observant la superposition de ces deux occurences du motif, il est possible de croire que l'ARN est la partie fixe[1] où les acides aminés viendraient se loger et interagir. Comme le démontre la figure 4.5, la structure de l'ARN est peu modifiée. Cela met quelque peu en doute le principe selon lequel les ARN opèrent plus fréquemment une réorganisation spatiale pour s'ajuster aux conformations protéiques. Dans notre cas, cette analyse a porté sur le ribosome qui semble avoir tendance à se comporter autrement du reste des autres structures, car il est vrai que l'on connait encore peu de choses sur cette structure qui a été récemment publiée. À ce moment, il est encore difficile de trouver une utilité fonctionnelle à ce motif, mais il demeure certainement un facteur de stabilité et de liaison fondamental entre les deux protéines ribosomales qui le possèdent (S11 et S12) et l'ARN ribosomal.

4.4 Motif conservé retrouvé chez 1ASZ et 1B7F

Un autre motif intéressant, faisant interagir beaucoup plus d'acides aminés cette fois, a été retrouvé à l'intérieur de deux structures différentes. Les structures 1ASZ [54] et 1B7F [55] (voir la figure 4.6) sont pourvus d'un motif qui contient trois acides aminés juxtaposés sur un feuillet-beta protéique. Comme nous pouvons l'observer sur la figure 4.7, il semble que ce feuillet vient se superposer sur un uracile pour former des liaisons hydrogène. Il

[1] C'est-à-dire la plus stable, la moins flexible.

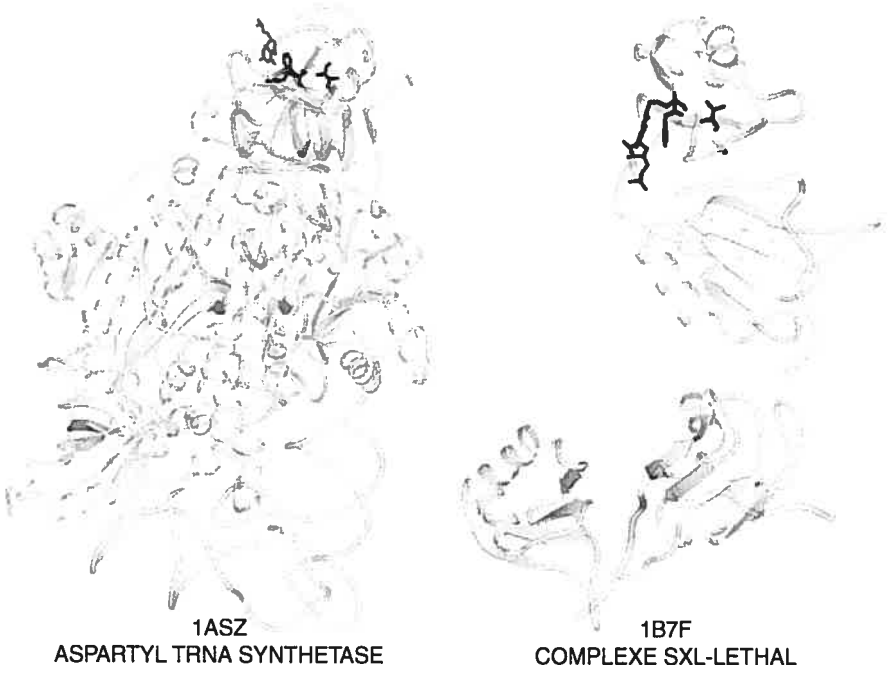


FIG. **4.6. Emplacement des motifs similaires retrouvés dans les structures des molécules ASZ et B7F. Les résidus d'acide aminé sont représentés en couleur plus foncée tandis que le ribonucléotide du motif est en bleu plus pâle.**

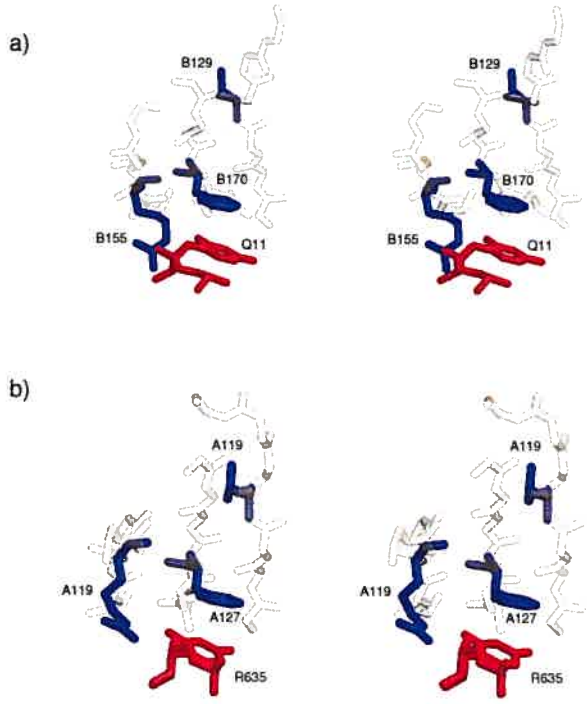
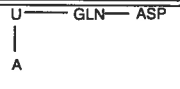
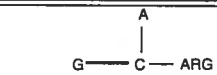
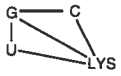
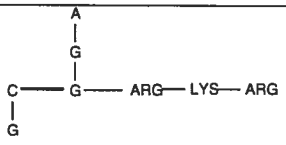
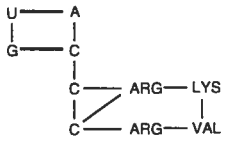
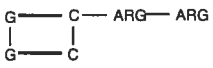
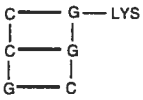
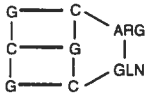
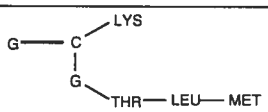
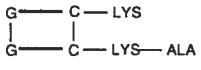
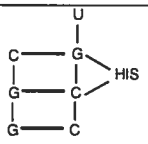
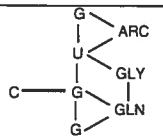
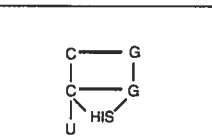
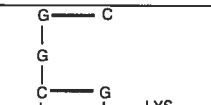
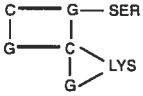


FIG. **4.7. Vue stéréoscopique du motif présent chez ASZ (en a) et chez B7F (en b).**

appert que cette interaction n'en soit pas une de stabilisation, car les résidus ne viennent pas s'intercaler dans la structure de l'autre molécule. En d'autres termes, même s'il s'agit possiblement de ponts hydrogène entre les deux molécules, ce motif ne génère pas une forte stabilité au complexe pour l'empêcher de se scinder, d'autant plus que le motif est placé en "extrémité" de la molécule. S'il était possible de donner un ordre d'importance de stabilité aux interactions, celui-ci figurerait parmi les derniers. Il s'agit plutôt d'un motif de reconnaissance spécifique qui permet à certaines protéines de discerner, parmi un ensemble de molécules différentes, l'ARN engendrant un appariement approprié et vice-versa.

Structure secondaire	Fichier PDB retrouvé	Structure secondaire	Fichier PDB retrouvé
	1DFU et 1EXD		1A4T et 1CK8
	1A9N et 1DFU		1FJF (3X)
	1FJF (3X)		1FJF (2X)
	1FJF (2X)		1FJF (2X)
	1FJF (2X)		1FJF (2X)
	1FJF (2X)		1FJF (2X)
	1FJF (2X)		1FJF (2X)
	1FJF (2X)		1FJF (2X)
	1FJF (2X)		1FJF (2X)
	1FJF (2X)		1FJF (2X)

TAB. **4.3. Tableau des motifs découverts.**

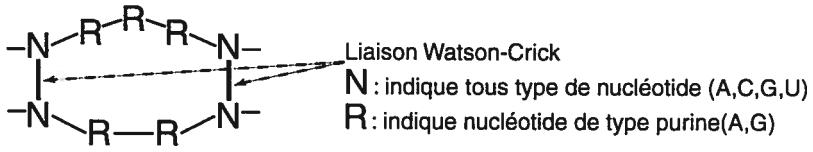


FIG. **4.8. Motif utilisé pour tester la rapidité de l'algorithme développé.**

4.5 Autres motifs conservés

D'autres motifs ont également été découverts à l'aide de notre méthode. La table 4.3 montre la structure secondaire de ceux-ci ainsi que les structures le possédant. Ces motifs sont un peu moins notables que les deux premiers soulignés, mais démontrent tout de même qu'il existe une certaine forme de récurrence au niveau des interactions ARN-protéine. Il est évident de voir que la majorité d'entre eux figure parmi la structure 1FJF (sous-unité 30S du ribosome). Celle-ci compte plus de 4000 résidus, comparativement à moins de 200 pour les autres structures.

4.6 Compariason de la rapidité d'exécution

Comme précédemment discuté, notre approche permet également de rechercher des structures de motif prédéfini dans un ensemble de structure donné. De cette façon, il devient beaucoup plus évident de comparer notre méthode avec celle antérieurement utilisée pour la recherche de motif [42]. Comme structure test, nous avons défini le motif de la figure 4.8 qui possède une taille relativement petite. Nous avons découvert trois emplacements semblables aux deux méthodes qui correspondent à un temps de *115.65 sec* pour la version antérieure et de *24.57 sec* pour la nôtre sur un PC Athlon 1300MHz. L'efficacité est nettement évidente lorsque des motifs de taille plus importante sont recherchés. À titre d'exemple, nous avons recherché un morceau aléatoire d'ARN de 200 nucléotides du ribosome. Le résultat que nous avons obtenu est de *0.73 sec* pour notre méthode comparativement à un temps indéterminé de plus de 24 heures pour l'approche d'isomorphisme de

graphe. En recherchant la même structure dans elle-même comme pour le tRNA[2], les temps respectivement obtenus ont été de *0.06 sec* contre plus de 24 heures. Ceci démontre donc les raisons qui nous ont poussés à modifier notre approche afin de la rendre plus efficace pour l'analyse des motifs de taille plus importante. La recherche en profondeur a donc été un incitatif d'implantation de notre algorithme.

[2]Les ARN de transfert sont de petites molécules d'ARN qui ont une longueur comprise entre 70 et 90 nucléotides. Celui qui a été utilisé pour notre essaie était conçu pour porter la phénylalanine comme acide aminé, c'est pourquoi il porte le nom de $tRNA^{phe}$ [56].

Chapitre 5

Discussion

De cette analyse des interactions ARN-protéine, il a été possible de découvrir certains aspects importants. Nos façons d'aborder le problème sont très spécifiques. Par contre, la morphologie des interactions ARN-protéine indiquent, au niveau atomique, le peu de spécificité de celles-ci. Les interactions se font sur le squelette électronégatif de l'ARN sans préférence du nucléotide impliqué. Il devient alors plus difficile de trouver des motifs récurrents lorsque les paramètres de recherche sont sélectifs sur les types de noeuds. Dans ce cas, certains motifs qui ont une forte ressemblance structurale, mais une faible similarité séquencielle, peuvent être oubliés lors de la recherche. Tout dépend de la définition que l'on donne d'un motif.

Un moyen de rendre notre recherche plus axée sur la structure des motifs serait de relâcher les contraintes d'appariement entre les différents types de résidus. Il serait ensuite nécessaire de détailler les arcs reliant les noeuds pour leur permettre de mieux définir les relations spatiales entre les résidus. Le travail de triage manuel demeurerait tout de même exigeant, car il permettrait cette fois de discerner les motifs possédant un *pattern* séquentiel régulier au sein de ces motifs nouvellement trouvés. Cette nouvelle paramétrisation allouerait également la rechercher des occurences dégénérées des motifs que nous avons trouvés auparavant. Cette dégénérescence pourrait se faire selon les différents groupes d'acides aminés ou d'ARN (purines et pyrimidines) ou uniquement selon le type de résidus impliqué (nucléotide ou acide aminé).

Il appert que la quantité de structures de complexes ARN-protéine dans les bases de données publiques soit insuffisante pour trouver une récurrence significative de motifs. Il sera donc intéressant de continuer l'analyse des nouvelles structures qui apparaîtront, afin

de trouver de nouveaux motifs ou de nouvelles occurences différents de ceux déjà trouvés. De cette observation, des comparaisons et des classifications selon les différentes familles de protéines pourront établir des liens de causalité pour l'appariement spécifique d'un ARN envers une certaine protéine et confirmer l'hypothèse d'un motif fonctionnel.

Il sera peut-être possible de confirmer l'hypothèse que nous avons élaborée auparavant au sujet du réarrangement moléculaire s'opérant au sein des protéines, au lieu du contraire fréquemment adopté par la communauté scientifique. Toutefois, il semble que le ribosome se comporte très différemment de l'ensemble des autres structures, il suffit de remarquer que nous n'avons pas retrouvé un motif contenu à la fois dans le ribosome et dans une autre structure.

Chapitre 6

Conclusion

Les travaux présentés dans ce mémoire ont tenté d'explorer différentes avenues dans le domaine encore peu connu des interactions ARN-protéine. Dans cette lancée, notre but était d'obtenir des sous-graphes maximaux bordant les liaisons inter-moléculaires entre l'ARN et la protéine associée. Les méthodes actuelles développées par *Gendron et al.* [42] ne nous permettaient pas d'effectuer la recherche de sous-structures efficacement. Par conséquent, nous avons voulu modifier l'approche isomorphique en développant un algorithme de recherche récursive.

Les chapitres 2 et 3 démontrent l'approche que nous avons considérée pour tenter de découvrir des structures récurrentes au sein des complexes ARN-protéine. L'algorithme qui a été développé s'inspire de la recherche en profondeur et permet de trouver les sous-graphes maximaux en comparant, par paires, l'appariement des structures soumises. Les sous-graphes qui en découlent sont ensuite comparés avec l'ensemble des sous-graphes afin d'effectuer des regroupements. Nous avons également trouvé d'autres utilités non négligeables à cet outil. Il a été possible de chercher rapidement, grâce à notre approche, la présence d'un motif connu à l'intérieur d'ensembles de structure. À l'aide d'une même molécule d'ARN possédant dans un cas une substance complexée (ex. antibiotique) et se retrouvant sous une forme "sauvage" dans l'autre cas, nous avons remarqué qu'il est possible d'extraire aisément les éléments de structure qui diffèrent d'un cas à l'autre.Ainsi, il devient facile de déterminer les résidus qui sont déplacés grâce à la substance étrangère et possiblement induire des propriétes fonctionnelles à certains agents.

Le chapitre 4 dévoile les motifs qui ont été extraits d'une cinquantaine de complexes soumis à l'analyse. La structure du ribosome, de par sa taille, a permis de découvrir le plus

de motifs récurrents. Le temps pour générer ces motifs démontre également l'efficacité de l'algorithme.

L'étude des interactions ARN-protéine en est encore à ses balbutiements, mais grâce aux développements importants d'outils comme celui que nous avons décrit, il sera possible de découvrir les facteurs essentiels qui lient une protéine à un ARN spécifique. Cette percée permettra éventuellement d'automatiser les comparaisons de complexes et amènera une prédiction automatique rapide des combinaisons structurelles pouvant se former à l'intérieur d'une cellule vivante.

Références

[1] Projet HUGO. www.oml.gov/hgmis/.

[2] J.D. Rawn. *Biochemistry*. De Boeck-Wesmael, 1990.

[3] I. Vidovic, S. Nottrott, Klaux Hartmuth, R. Luhrmann et R. Ficner. Crystal structure of the spliceosomal 15.5kd protein bound to a u4 snrna fragment. *Molecular Cell*, **6** :1331–1342, 2000.

[4] X. Shao et N.V. Grishin. Common fold in helix-hairpin-helix proteins. *NAR*, **28** :2643–2650, 2000.

[5] Y. Hou, X. Zhang, J.A. Holland et D. Davis. An important 2-oh group for an rna-protein interaction. *NAR*, **29** :976–985, 2001.

[6] L.B. Blyn, L.M. Risen ad R.H. Griffey et D.E Draper. The rna-binding domain of ribosomal protein l11 recognizes an rrna tertiary structure stabilized by both thiostrepton and magnesium ion. *NAR*, **28** :1778–1784, 2000.

[7] D. Dertinger, L.S. Behlen et O.C. Uhlenbeck. Using phosphorothioate-substitued rna to investigate the thermodynamic role of phosphates in a sequence specific rna-protein complex. *Biochemistry*, **39** :55–63, 2000.

[8] A. Akhtar, D. Zink et P.B. Becker. Chromodomains are protein-rna interaction modules. *nature*, **407** :405–409, 9 2000.

[9] C. Mazza, M. Ohno, A. Segref, I.W. Mattaj et S. Cusack. Crystal structure of the human nuclear cap binding complex. *mcell*, **8** :383–96, 2001.

[10] A.L. Lehninger, D.L. Nelson et M.M. Cox. *Principles of Biochemistry*. Worth Publishers, New York, second edition édition, 1993.

[11] J.D. Watson et F.H.C Crick. A structure for deoxyribose nucleic acid. *Nature*, **171** :737–738, 1953.

[12] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts et J.D. Watson. *Molecular Biology of the Cell*. Garland Pub, 1997.

[13] D.P. Snustad, M.J. Simmons et J.B. Jenkins. *Principles of Genetics*. John Wiley & Sons, Inc., 1997.

[14] Patrick Gendron, Sébastien Lemieux et François Major. Quantitative analysis of nucleic acid three-dimensional structures. *Journal of Molecular Biology*, **308** :919–936, 2001.

[15] W. Saenger. *Principles of Nucleic Acid Structure*. Springer-Verlag, New York, USA, 1984.

[16] C. Branden et J. Tooze. *Introduction to Protein Structure*. Garland Publishing, Inc., 1991.

[17] I.K. McDonald et J.M. Thornton. Satisfying hydrogen bonding potential in proteins. *jmb*, **238** :777–793, 1994.

[18] C.W. Muller et C. Wolberger. Protein-nucleic acid interactions. *Current Opinion in Structural Biology*, **12** :69–71, 2002.

[19] D.J. Klein, T.M. Schmeing, P.B. Moore et T.A. Steitz. The kink-turn : a new rna secondary structure motif. *The EMBO Journal*, **20** :4214–4221, 2001.

[20] P. Nissen, J.A. Ippolito, N. Ban, P.B. Moore et T.A. Steitz. Rna tertiary interactions in the large ribosomal subunit : The a-minor motif. *pnas*, **98** :4899–4903, April 2001.

[21] J.G. Arnez et J. Cavarelli. Structures of rna-binding proteins. *Quaterly Reviews of Biophysics*, **30** :195–240, 1997.

[22] N.B. Leontis et E. Westhof. A common motif organizes the structure of multi-helix loops in 16s and 23s ribosomal rnas. *jmb*, **283** :571–583, 1998.

[23] F. Bachand, I. Triki et C. Autexier. Human telomerase rna-protein interactions. *nar*, **29** :3385–3393, 2001.

[24] S. Jones, D.T.A Daley, N.M. Luscombe, H.M. Berman et J.M. Thornton. Protein-rna interactions : a structural analysis. *nar*, **29** :943–954, 2001.

[25] D.E. Draper. Themes in rna-protein recognition. *jmb*, **293** :255–270, 1999.

[26] D.E. Draper et Reynaldo L.P. Rna binding strategies of ribosomal proteins. *nar*, **27** :381–388, 1999.

[27] J.M. Pérez et G. Varani. Recent advances in rna-protein recognition. *Current Opinion in Structural Biology*, **11** :53–58, 2001.

[28] G. Varani et K. Nagai. Rna recognition by rnp proteins during rna processing. *Annu. Rev. Biophys. Struct.*, **27** :407–45, 1998.

[29] E.E. Abola, J.L. Sussman, J. Priluski et N.O. Manning. Protein data bank archives of three-dimensional macromolecular structures. *Methods Enzymol.*, **277** :556–571, 1997.

[30] S.P. Gardner et J.M. Thornton. The iditis relational database of protein-structure. *American chemical society journal*, **202** :32, 1991.

[31] D. Gusfield. *Algorithms on strings, trees and sequences*. Cambridge University Press, 1997.

[32] I. Eidhammer, I. Jonassen et W.R. Taylor. Structure comparison and structure patterns. *Journal of Computational Biology*, **7** :685–716, 2000.

[33] L. Holm et C. Sander. Searching protein structure databases has come of age. *Proteins*, **19** :165–173, Jul 1994.

[34] L. Falquet, M. Pagni, P. Bucher, N. Hulo, C.J. Sigrist, K. Hofmann et A. Bairoch. The prosite database, its status in 2002. *nar*, **30** :235–238, 2002.

[35] A.G. Murzin, S.E. Brenner, T. Hubbard et C. Chothia. Scop : a structural classification of proteins database for the investigation of sequences and structures. *jmb*, **247** :536–540, 1995.

[36] P.J. Artymiuk, A.R. Poirrette, H.M. Grindley, D.W. Rice et P. Willett. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *jmb*, **243** :327–344, 1994.

[37] B. Billoud, M. Kontic et A. Viari. Palingol : a declarative programming language to describe nucleic acids'secondary structures and to scan sequence databases. *nar*, **24** :1395–1403, 1996.

[38] D. Gautheret, F. Major et R. Cedergren. Pattern searching/alignement with rna primary and secondary structures : an effective descriptor for trna. *CABIOS*, **6** :325–331, 1990.

[39] G.J. Kleywegt. Recognition of spatial motifs in protein structures. *jmb*, **285** :1887–1897, 1999.

[40] M. Treger et E. Westhof. Statistical analysis of atomic contacts at rna-protein interfaces. *Journal of Molecular Recognition*, **14** :199–214, 2001.

[41] N.M. Luscombe, R.A. Laskowski et J.M. Thornton. Nucplot : a program to generate schematic diagrams of protein-nucleic acid interactions. *nar*, **25** :4940–4945, 1997.

[42] P. Gendron, D. Gautheret et F. Major. Structural ribonucleic acid motifs identification and classification. *High Performance Computing Systems and Applications*, 1998.

[43] J.R. Ullmann. An algorithm for subgraph isomorphism. *J. ACM*, **23** :31–42, 1976.

[44] R. Apweiler, A. Gateau et W. Junker. Swiss-prot and its computer-annotated supplement trembl : New developments in the linking of biological databases and computer-generation of annotation. *Folding and Design*, **1** :3–4, 1996.

[45] NCBI. www.ncbi.org.

[46] W.C. Barker, J.S. Garavelli, P.B. McGarvey, C.R. Marzec, B.C. Orcutt, G.Y. Srinivasarao, L.L. Yeh, R.S. Ledley, H. Mewes, F. Pfeiffer, A. Tsugita et C. Wu. The pir-international protein sequence database. *nar*, **27** :39–43, 1999.

[47] W. Kabsch et C. Sander. Dictionary of protein secondary structure : Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22** :2577–2637, 1983.

[48] M. Gondran et M. Minoux. *Graphs and algorithms*. John Wiley & Sons, Inc., 1984.

[49] S. Mitchell, T. Beyer et W. Jones. Linear algorithms for isomorphism of maximal outerplanar graphs. *Journal of the ACM*, **26**, October 1979.

[50] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica*, **A32** :922–923, 1976.

[51] G. Levi. A note on the derivation of maximal common subgraphs of two directed or undirected graphs. *J. Am. Chem. Soc.*, **99** :7668–7671, 1977.

[52] J.J. McGregor. Backtrack search algorithms and the maximal commun subgraph problem. *Software - Practice and experience*, **12** :23–34, 1982.

[53] A.P. Carter, W.M. Clemons, D.E. Brodersen, B.T. Wimberly, R. Morgan-Warren et V. Ramakrishnan. Functional insights from the structure of the 30s ribosomal subunit and its interactions with antibiotics. *Nature*, **407** :340, 2000.

[54] J. Cavarelli, B. Rees, M. Ruff, J.C. Thierry et D. Moras. The active site of yeast aspartyl-trna synthetase : structural and functional aspects of the aminoacylation reaction. *EMBO journal*, **13**, 1994.

[55] N. Handa, O. Nureki, K. Kurimoto, I. Kim, H. Sakamoto, Y. Shimura, Y. Muto et S. Yokoyama. Structural basis for tra mrna precursor recognition by the sex-lethal protein. *Nature*, **398** :579, 1999.

[56] L. Jovine, S. Djordjevic et D. Rhodes. The crystal structure of yeast phenylalanine trna at 2.0 a resolution. *JMB*, **301**, 2000.