Université de Montréal

Modélisation tridimensionnelle des ARN par exploration de l'espace conformationnel et satisfaction de contraintes

par
Philippe Thibault

Département d'informatique et de 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

mars 2004

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :

Modélisation tridimensionnelle des ARN par exploration de l'espace conformationnel et satisfaction de contraintes

présenté par :

Philippe Thibault

a été évalué par un jury composé des personnes suivantes :

Bernard Gendron,	président-rapporteur
François Major,	directeur de recherche
Sylvie Hamel,	membre du jury

Mémoire accepté le : 19 avril 2004

Sommaire

Dans ce mémoire, nous traitons du problème de modélisation de la structure tridimensionnelle des ARN dans le cadre de conception du système de modélisation *MC-Sym*. Nous avons développé une toute nouvelle approche de modélisation où les éléments constitutifs d'un nucléotide – base azotée, phosphate et ribose – sont positionnés dans l'espace tridimensionnel de façon indépendante. D'abord, les bases azotées et les phosphates sont positionnés de façon rigide par transformation linéaire. Ensuite, leur interconnexion est réalisée par la construction artificielle des riboses. Formellement, cette nouvelle méthode élimine l'exploration du sous-espace des conformations statiques des nucléotides. Expérimentalement, elle réussit à explorer complètement l'espace conformationnel de structures simples en quelques minutes alors que le système *MC-Sym* original nécessite plusieurs jours pour réaliser la même exploration. Pour sa part, la méthode de construction artificielle atome par atome du ribose est paramétrée par les torsions caractéristiques de la structure du ribose. Les paramètres optimaux dans le contexte d'une modélisation sont déterminés par approximation numérique. Nous avons développé et comparé quelques techniques d'optimisation différentes. La plus performante est sans contredit notre méthode inédite d'estimation directe des paramètres. En effet, cette méthode nécessite toujours exactement deux itérations de construction, alors que les autres nécessitent une quantité variable d'itérations oscillant entre 20 et 120. Finalement, nous avons développé un algorithme probabiliste de résolution du problème de modélisation par satisfaction de contraintes. En effet, tout le système de modélisation s'encadre dans un CSP (*constraint satisfaction program*). Cet algorithme probabiliste, comparé à l'algorithme classique de retour arrière (*backtrack*) accroît grandement le taux de génération de solutions et permet de considérer un espace conformationnel de taille maximale, sans troncature.

Mots clés : bioinformatique, structure tertiaire, *MC-Sym*, modélisation discrète, mode de *puckering*, retour arrière, problème de satisfaction de contraintes, propagation des contraintes.

Abstract

In this document, we study RNA tertiary structure modeling inside *MC-Sym* modeling system framework. We developed a new modeling approach where nucleotide's inner components – nitrogen base, phosphate and ribose – are independently placed in three-dimensional space. First, nitrogen bases and phosphates are placed as rigid body by linear transformation. Then, they are interconnected by artificial construction of riboses. Formally, this new method eliminates nucleotides' static conformation sub-space exploration. Experimentally, it succeeds in the complete exploration of some simple structure's conformational space in a few minutes, where *MC-Sym*'s original system takes many days to do the same exploration. For itself, the ribose's artificial construction method is parameterized by the ribose structure's feature torsions. The optimal parameters inside a modeling context are determined by numerical approximation. We developed and compared different optimisation methods. The most efficient is undoubtfully our novel direct parameter estimation method. Indeed, this method always consumes exactly two construction iterations, where the other methods consume a variable iteration amount ranging between 20 and 120. Finally, we developed a probabilistic algorithm to solve the modeling problem by constraint satisfaction. In fact, the whole modeling system encapsulates itself inside a constraint satisfaction problem. This probabilistic algorithm, compared against a classical backtrack algorithm, greatly builds up the solution generation rate, and makes it possible to use a maximal conformational space, that is without pruning.

Key words: bioinformatics, RNA tertiary structure, *MC-Sym*, discrete modeling, puckering mode, backtrack, constraint satisfaction problem, constraint propagation.

Table des matières

Liste des tableaux

Liste des figures

Chapitre 1

Mise en situation

Depuis que la génétique a révélé le fonctionnement de la cellule, les biologistes moléculaires n'ont cessé d'investiguer ses mécanismes internes et de pousser le niveau de détails à la représentation atomique. Puisque la structure d'une molécule dicte sa fonction, plusieurs chercheurs ont concentré leurs intérêts sur la détermination de la structure atomique d'une molécule. Les mécanismes de la génétique passent par l'action des molécules d'acides désoxyribonucléiques (ADN), d'acides ribonucléiques (ARN) et des protéines. Les méthodes physiques couramment utilisées pour déterminer la structure atomique tridimensionnelle de ces molécules en particulier, soient la cristallographie aux rayons X et la résonance magnétique nucléaire (RMN), sont souvent freinées par les limites expérimentales et les coûts engendrés pour leur mise en œuvre [6]. C'est pourquoi le besoin d'une méthode artificielle de détermination de la structure s'est rapidement fait sentir. Ici entre en jeu la science étendue de l'informatique pour mettre au point un système de modélisation artificielle qui produit un modèle atomique précis et visualisable à l'écran, en se basant sur un ensemble restreint d'information biochimique sur la structure de la molécule. En 1991, un tel système a été développé pour les ARN par François Major et ses collaborateurs [27] : *MC-Sym*. Nous proposons d'utiliser le cadre de conception de *MC-Sym* pour développer une nouvelle approche de modélisation des ARN plus performante.

1.1 Le monde de l'ARN

Le dogme central de la biologie cellulaire établit le mécanisme de fonctionnement de base de la cellule, depuis le matériel génétique conservé précieusement à l'abri à l'intérieur de son noyau jusqu'aux protéines qui circulent librement pour effectuer les diverses tâches exprimées par les gènes. Le matériel génétique est composé d'ADN sous la forme d'une double hélice. L'ADN, tout comme la protéine, est un polymère formé de la répétition en chaîne d'unités moléculaires de base : les nucléotides. L'ADN est composé de nucléotides :

adénine (A), guanine (G), cytosine (C) et thymine (T). Pour sa part, la protéine est composée de vingt acides aminés différents. Une représentation abstraite de ces polymères étend la chaîne de lettres qui correspond directement à la succession des nucléotides ou des acides aminés. Par conséquent, il existe un mécanisme de traduction entre l'alphabet à quatre lettres de l'ADN vers celui à vingt lettres de la protéine. Ce mécanisme se divise en deux processus, où l'ARN joue le rôle clé de l'interprète.

Le premier processus est l'étape de *transcription*, où un brin de l'ADN est séparé de la double hélice et transcrit sur une molécule d'ARN, grâce à l'enzyme ARN-polymérase. Le résultat est un ARN messager (ARNm) qui contient une partie de l'information du gène. L'alphabet est le même que l'ADN, à l'exception de la thymine (T) qui est remplacée par l'uracile (U). L'étape subséquente, le processus de *traduction*, traduit littéralement le message codé en quatre lettres contenu dans l'ARNm en celui codé en vingt lettres des protéines. Ce processus implique le concours d'un ribosome (complexe ARN-protéines composé à 66% d'ARN) et de l'ARN de transfert (ARNt) : l'ARNm est lu par groupe de trois lettres par le ribosome, puis un ARNt se complémente sur ce mot de trois lettres pour relier l'acide aminé correspondant à la chaîne polypeptidique grandissante. La figure 1 illustre ces différentes étapes de la synthèse d'une protéine selon le dogme central. Une fois l'ARNm complètement lu par le ribosome, la protéine est libérée pour exécuter sa tâche, telle que dictée initialement par le gène.

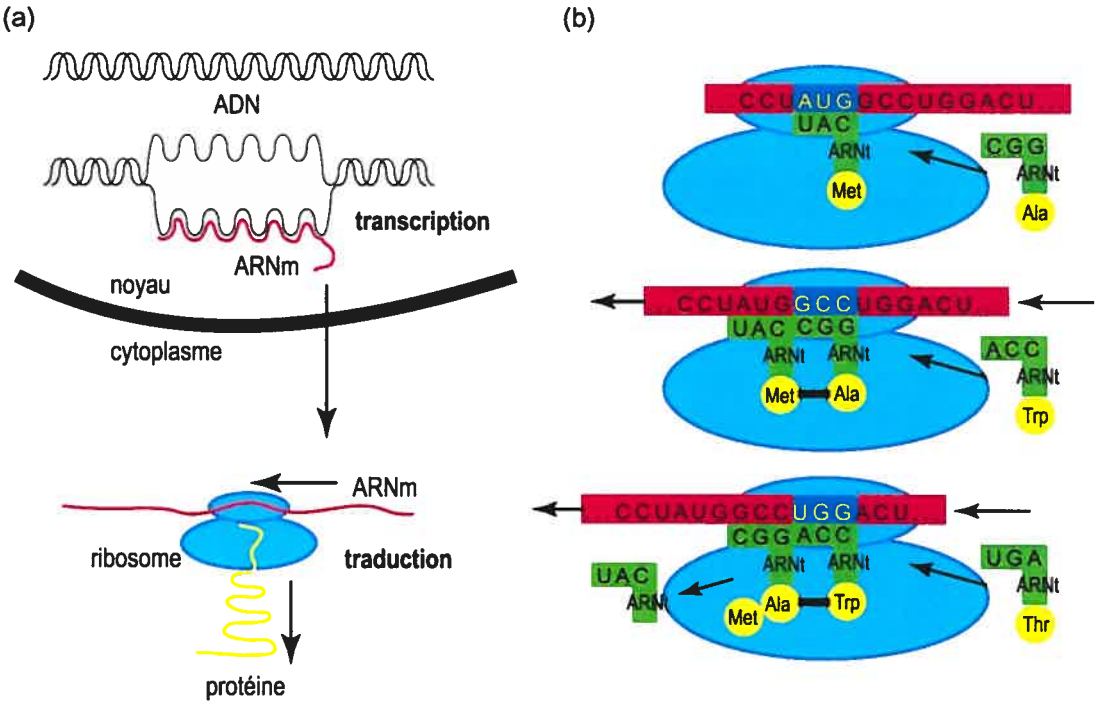


Figure 1 : Dogme central de la biologie cellulaire. (a) Aperçu général de l'étape de transcription de l'ADN (en noir) en ARNm (en rouge) dans le noyau de la cellule et de traduction de l'ARNm en protéine (en jaune) dans le ribosome (en bleu clair). (b) Détails de l'étape de traduction. À l'intérieur du ribosome, un ARNt (en vert) se complémente sur trois lettres (nucléotides) de l'ARNm. Cet ARNt transporte un acide aminé (en jaune) correspondant à la série de trois nucléotides complémentaires. Ensuite, un second ARNt vient se lier au triplet suivant sur l'ARNm, transportant lui aussi son acide aminé spécifique. Une liaison peptidique est alors créée entre les deux acides aminés dans le ribosome. De cette façon, la chaîne polypeptidique de la protéine est progressivement synthétisée par le décodage de l'ARNm.

Dans ce contexte, l'ARN semble jouer un rôle plutôt passif de messager bilingue entre l'ADN du gène et la protéine. Cependant, plusieurs recherches récentes ont mis à jour divers mécanismes sous-jacents au dogme central qui sont déterminants dans le contrôle de l'action des protéines dans la cellule. En 1989, le prix Nobel de chimie fut remis aux chercheurs Sidney Altmann et Thomas Cech pour leur découverte des propriétés catalytiques de l'ARN. Ainsi, l'ARNm peut agir lui-même sur son message, en s'auto-excisant par exemple, pour influencer la production d'une protéine en particulier. Dès lors, tout un horizon de recherche s'est ouvert sur les ARN dits non-codants, i.e. qui ont un tout autre rôle que celui de la simple traduction. Divers mécanismes de régulation et de contrôle de la qualité impliquant des ARN viennent agir directement sur la production ou sur l'action des protéines (voir [1] et [2] pour une revue des ARN non-codants et des mécanismes de régulation). Soudain, l'ARN devient

un membre actif dans le comportement général de la cellule. Certains chercheurs avancent même que la spécificité de la séquence des ARN les rend plus favorable à l'évolution que les protéines, et alimentent le concept du *Monde de l'ARN* (*the RNA World*) : époque primitive dans l'apparition de la vie sur Terre où la soupe des éléments primordiaux était gouvernée uniquement par des ARN spécialisés. Ainsi, les efforts de détermination de la structure tridimensionnelle d'un ARN deviennent tout à fait justifiés, en se rappelant que la structure dicte la fonction.

1.2 Formalisme de modélisation

D'un point de vue informatique, un système de modélisation d'ARN prend en entrée un ensemble d'information sur la structure d'un ARN et produit en sortie un ou plusieurs modèles tridimensionnels visualisables à l'écran. Dans le cadre de ce travail, une distinction précise s'établit entre *structure* et *modèle*. La *structure* d'un ARN est l'ensemble de ses propriétés biochimiques et stéréochimiques qui font se replier la chaîne polynucléotidique sur elle-même pour adopter dans l'espace une conformation particulière et identifiable, alors que son *modèle* consiste en une abstraction conceptuelle qui représente sa structure moléculaire artificiellement. Dans un système informatique, le modèle de l'ARN est constitué simplement de l'ensemble des coordonnées tridimensionnelles de chacun des atomes de la structure, conservées dans un fichier.

1.2.1 Nomenclature de la structure moléculaire

La structure moléculaire de la chaîne polynucléotidique d'un ARN se subdivise d'abord en nucléotides : A, C, G et U. À un niveau plus fin, un nucléotide se subdivise lui-même en trois parties moléculaires : la base azotée (ou simplement la base), le ribose et le groupement phosphate (ou simplement le phosphate). La figure 2 illustre cette nomenclature et numérote les atomes du nucléotide. La base azotée représente directement l'adénine, la guanine, la cytosine et l'uracile : les deux premières se regroupent dans la famille des purines, les deux dernières dans celle des pyrimidines. Dans tous les cas, la structure moléculaire de la base azotée est rigide et planaire, sa modélisation est donc statique. Le ribose et le groupement phosphate se répète de façon identique d'un nucléotide à l'autre et forment ensemble le

squelette (*backbone*) de l'ARN. Le ribose est composé d'un cycle furanosique dont la stéréochimie est variable; ainsi le squelette possède une certaine flexibilité. La stéréochimie du ribose est adressée par son mode de froissement (*puckering mode*) qui décrit plusieurs conformations particulières du furanose (voir [19] pour une définition formelle).

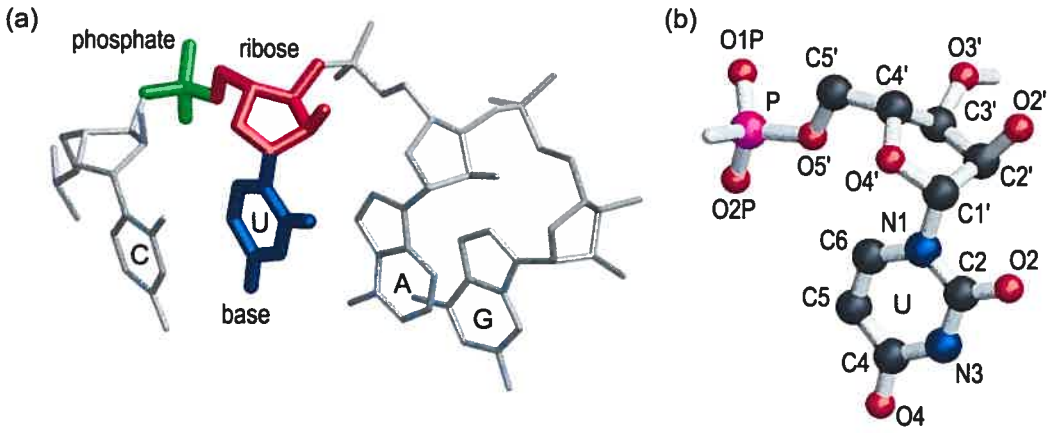


Figure 2 : Nomenclature d'une chaîne polynucléotidique et numérotation des atomes. (a) Une chaîne de quatre nucléotides est ici représentée, l'un a ses trois parties moléculaires mises en évidence : la base azotée (en bleu), le ribose (en rouge) et le groupement phosphate (en vert). La jonction successive des phosphates et des riboses forment le squelette de la structure. (b) Numérotation des atomes du nucléotide uracile. Pour les trois autres nucléotides, seule la numérotation de la base est différente. Images générées par *Molscript* [36] [44] et *Raster3D* [37] [45].

1.2.2 Appariement de bases et relation d'appariement

Le repliement de la chaîne polynucléotidique est dirigé par des ponts hydrogènes (ponts-H) qui se forment entre certaines bases azotées complémentaires. Un pont-H est une interaction intermoléculaire forte qui est le fruit de l'attirance électromagnétique entre un hydrogène (un proton) et une paire d'électrons libres, provenant habituellement de l'oxygène ou de l'azote. Un **appariement de bases** se caractérise donc par deux bases azotées positionnées dans l'espace sur le même plan par la stabilisation des interactions par pont-H. À cause de la forme orbitale du champ électronique, un pont-H est directionnel et favorise un agencement coplanaire des deux bases. En général, deux ponts-H sont nécessaires à la stabilisation de l'appariement, cependant il en existe aussi à un ou trois ponts-H [18]. En 1953, James D. Watson et Francis H. Crick ont découvert que l'ADN adopte la conformation de la double hélice grâce aux appariements entre les bases azotées G-C et A-T (A-U dans l'ARN) [3]. Cependant l'ARN adopte d'autres conformations plus complexes que la double hélice,

nécessitant une plus grande variété d'appariements de bases. Ainsi, la base azotée s'interface à des régions différentes, résultant en différentes conformations d'appariement. La figure 3a illustre les trois différentes interfaces. La nomenclature d'un appariement particulier se réfère aux interfaces respectives des bases azotées : la face Watson-Crick (*W*), la face Hoogsteen (*H*) et la face du sucre, c'est-à-dire du ribose (*S*). Par exemple, l'appariement Watson-Crick s'identifie par les interfaces *W* et *W*. La figure 3b montre une deuxième caractéristique des appariements de bases : l'orientation relative du squelette. L'identification d'un appariement de bases particulier par ses interfaces et par l'orientation du squelette devient dont claire et précise.

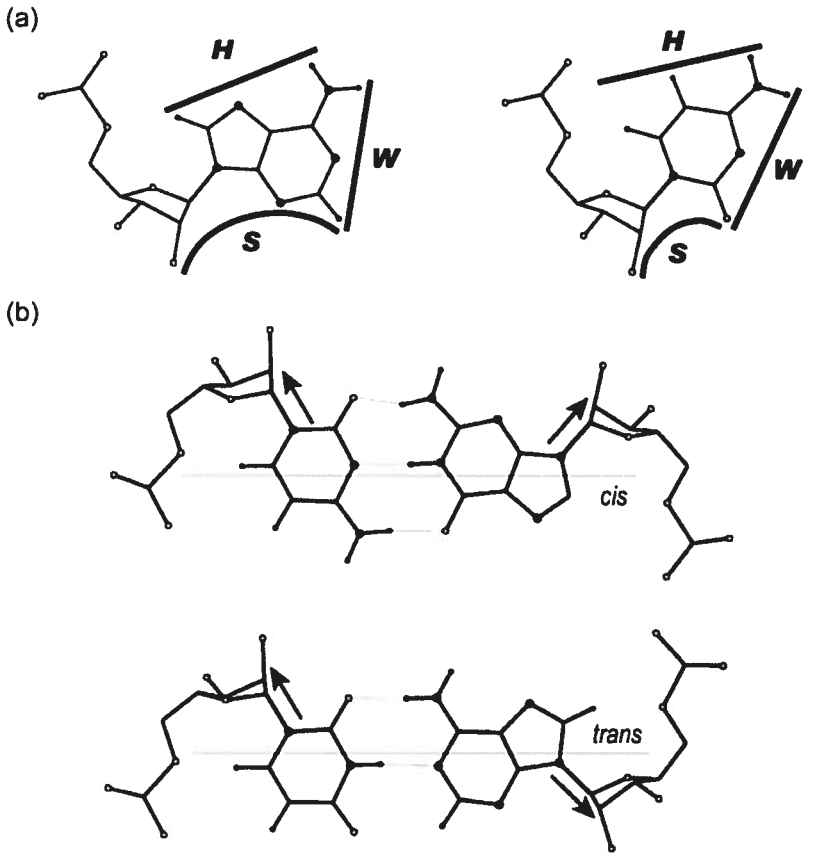


Figure 3 : Nomenclature des appariements de bases. (a) Trois différentes interfaces permettent des interactions par ponts-H. L'identification de l'appariement nomme simplement les interfaces en contact, par exemple *W* / *W* pour le type canonique Watson-Crick. (b) L'orientation relative du squelette ajoute un degré de spécificité : parallèle (*cis*) ou inversée (*trans*).

D'un point de vue strictement géométrique, un appariement de bases se modélise comme une composition de transformations linéaires – translation et rotation – qui exprime le mouvement nécessaire pour déplacer le plan d'une des deux bases azotées appariées en

superposition confondue avec celui de l'autre base, dans le référentiel de la première [29]. Ainsi, une **relation d'appariement** se définit comme une transformation linéaire exprimant la relation géométrique d'un appariement de bases dans le référentiel local d'une des deux bases. Donc, la relation d'appariement modélise le phénomène biochimique de l'appariement de deux bases azotées par pont-H et s'encode simplement par une matrice de transformation linéaire en coordonnées homogènes (4x4). L'expression de la transformation dans le référentiel d'une des deux bases permet de positionner l'autre base dans un contexte local, indépendant de la position globale de la base prise comme référence. Par ailleurs, le concept de relation d'appariement est plus général qu'un appariement de bases par pont-H. Celui-ci peut aussi modéliser le phénomène d'empilement de bases souvent remarqué entre les nucléotides adjacents d'une double hélice. Ce phénomène implique des dipôles induits qui sont des forces intermoléculaires plus faibles que des ponts-H, par conséquent leur géométrie relationnelle est plus variable que celle des appariements de bases. En fait, une relation d'appariement peut représenter le positionnement spatial relatif de toute paire de bases azotées, même si les bases ne sont pas reliées de façon mesurable dans la structure, soit par pont-H ou par empilement. Ainsi, une relation d'adjacence est une relation d'appariement mesurée entre deux bases adjacentes dans la chaîne polynucléotidique, qu'elles soient empilées ou non. Évidemment, une relation d'adjacence entre deux bases non-empilées est très variable, l'espace tridimensionnel couvert par les possibilités de positionnement d'une base par une telle relation est donc très vaste. Lors de la modélisation, l'utilisation d'une relation d'adjacence non-empilée pour la détermination d'un modèle est idéalement à proscrire si possible, sinon à minimiser. Elle est souvent la conséquence d'une faible connaissance préalable de la structure à modéliser, ou d'une structure peu contrainte (une boucle terminale, par exemple).

1.2.3 Niveaux d'abstraction de la structure

Dans le domaine de la modélisation de l'ARN, la structure de la chaîne polynucléotidique peut être représentée à trois différents niveaux d'abstraction : structure primaire, secondaire et tertiaire. Selon cet ordre, le degré de granularité de l'information structurale passe du plus grossier au plus fin. Il y a aussi une analogie avec la dimensionnalité du modèle représentant chaque niveau : respectivement unidimensionnel, bidimensionnel et tridimensionnel.

1.2.3.1 Structure primaire

La structure primaire représente le niveau le plus abstrait, soit simplement la séquence de lettres identifiant chacun des nucléotides au long de la chaîne. D'ailleurs, la structure primaire est habituellement appelée la séquence. Le modèle à ce niveau est une simple chaîne de caractères, d'où son unidimensionnalité.

1.2.3.2 Structure secondaire

La structure secondaire intègre l'information supplémentaire des appariements de bases qui permettent à la séquence de se replier à plat, en 2D. Le concept recherché ici est de pouvoir schématiser la structure à plat sur une feuille, sans détailler la structure moléculaire comme telle. Plusieurs éléments structuraux ressortent à ce niveau : tiges (aussi nommées hélices), boucles terminales et boucles internes reliant plusieurs tiges. Le modèle d'une structure secondaire est la séquence annotée : à chaque position de la séquence est associée zéro, une ou plusieurs autres positions signifiant qu'il y existe zéro, un ou plusieurs appariements entre ces bases. Un logiciel comme *RnaViz* [38] permet la visualisation, l'édition et l'impression d'un tel modèle. Tout un champ de modélisation s'intéresse uniquement à la prédiction de la structure secondaire d'un ARN à partir de sa séquence, en recherchant l'ensemble des appariements canoniques (Watson-Crick) potentiels qui permettent la formation d'une structure secondaire optimale, où le critère d'optimalité dépend d'une stabilité énergétique qui est fonction des éléments de structure secondaire mis en évidence (voir les logiciels *mfold* [7] [8] et *PKNOTS* [9]). Leontis et Westhof ont récemment proposé une nomenclature des appariements de bases dans les structures secondaires [30] qui suit celle des interfaces de contact des ponts-H et de l'orientation du squelette présentée précédemment. La figure 4 montre la structure secondaire d'un ARNt suivant la nomenclature de Leontis et Westhof.

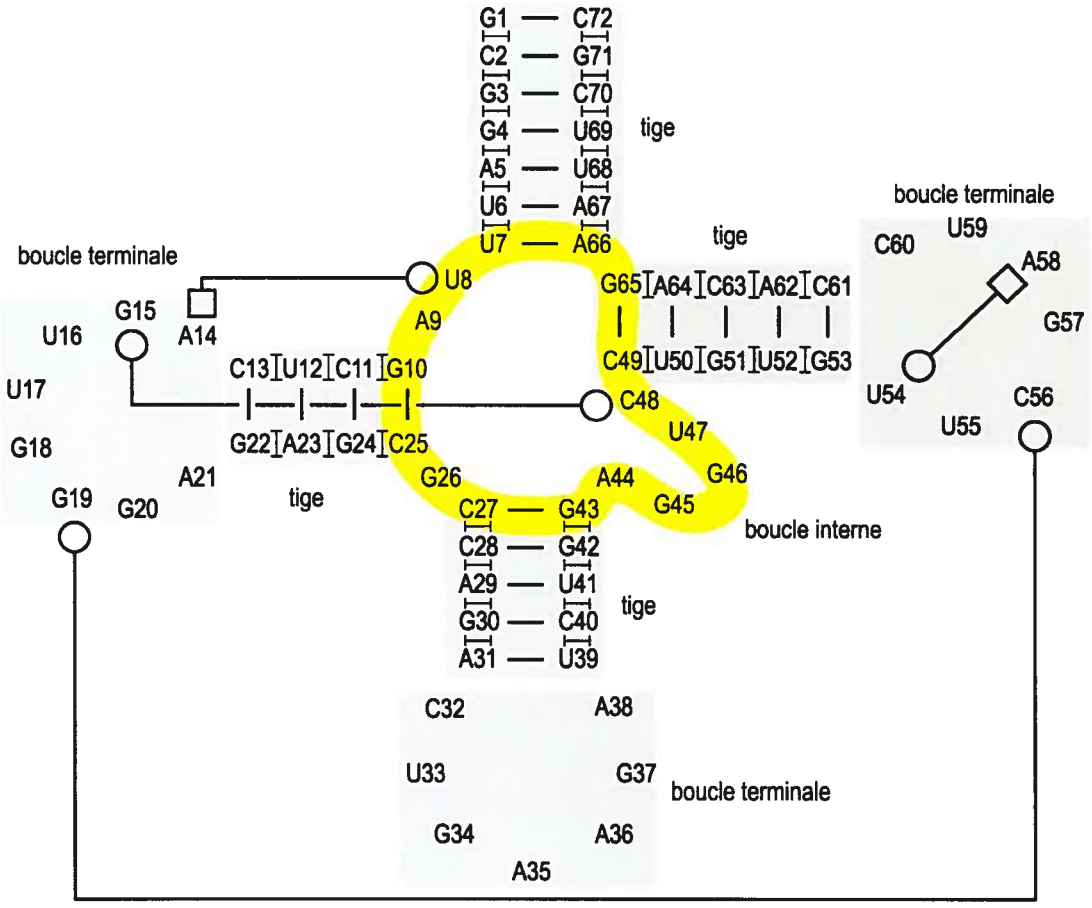


Figure 4 : Structure secondaire de ARNt^{Phe} provenant de la levure. Les appariements de bases suivent la nomenclature de Leontis et Westhof [30] (les cercles indiquent des interfaces *W* et les carrés des interfaces *H*). Les éléments de structure secondaire sont ombrés, soient trois tiges, trois boucles terminales et une boucle interne (en jaune). Les échafauds indiquent des empilements de base, cependant ils sont souvent omis dans les tiges puisqu'ils y sont implicites.

1.2.3.3 Structure tertiaire

La structure tertiaire représente l'ARN directement au niveau atomique, c'est-à-dire que le modèle présente la molécule en 3D. C'est le niveau d'abstraction le plus détaillé, en accord avec le modèle généralement accepté en chimie pour représenter une molécule. Selon cette définition, un modèle en plastique avec des boules pour les atomes et des tiges pour les liens covalents représente bien une structure tertiaire. Cependant, l'informatique s'intéresse plutôt à un fichier de coordonnées tridimensionnelles pour chaque atome, où l'atome est réduit à un seul point, de sorte à pouvoir visualiser le modèle interactivement en 3D et l'analyser en

utilisant l'outil logiciel adéquat. Dans le contexte de ce travail, toute référence à un modèle s'adresse implicitement à un modèle tertiaire.

Du point de vue de la prédiction de structure, la structure tertiaire est le but à atteindre. Souvent le problème se pose comme la détermination de la structure tertiaire d'un ARN en fonction de ses structures primaire et secondaire, c'est-à-dire de sa séquence et de la notion des appariements et empilements entre les bases azotées. Un système de modélisation prend donc cette information en entrée et produit en sortie un ensemble de modèles tridimensionnels représentant la structure tertiaire. *MC-Sym* [16] [40] est un tel système, mais il en existe d'autres qui se placent à différents pôles de conception de l'engin de modélisation. Par exemple, certains systèmes, comme *MANIP* [13] ou *ERNA-3D* [14], produisent artificiellement des éléments de structure de base, les tiges par exemple, et laissent le soin à l'usager de corriger le modèle 3D par manipulations interactives. En fait l'usager doit être un modélisateur chevronné puisque sa subjectivité et son talent prévalent à la vraisemblance du modèle. À l'opposé se retrouve *JUMNA* [10] [12] qui modélise la chaîne polynucléotidique par minimisation de l'énergie potentielle de la molécule. L'énergie de la molécule est fonction de sa conformation, plus précisément d'un ensemble de variables géométriques pour chaque nucléotide : torsions des liens covalents dans le ribose et géométrie hélicoïdale des bases azotées [11]. Une descente de gradients conjugués permet la modification de ces paramètres vers une conformation optimale. Si cette approche possède une solide fondation théorique, son application pratique est plus compliquée. Le problème majeur vient de la détermination de la conformation initiale pour le modèle; il faut avoir une connaissance préalable de la structure tertiaire à modéliser. Conceptuellement, *MC-Sym* permet une automatisation complète par modélisation discrète du processus de manipulation graphique présent dans *MANIP* et *ERNA-3D*. De plus, *MC-Sym* permet de trouver plusieurs modèles différents pour une structure en particulier. L'idée de base du système *MC-Sym* est de construire un modèle qui représente symboliquement la structure et de laisser le raffinement du modèle moléculaire à un système de minimisation d'énergie tel *JUMNA*.

1.2.4 Déviation RMS

La déviation RMS (*root mean square*) [34] [35] sert de métrique de distance entre deux modèles de la même structure tertiaire. Elle quantifie la similitude relative des deux modèles

par rapport à la structure tertiaire modélisée. Cette métrique calcule la moyenne des distances au carré entre chaque paire d'atomes homologues d'un modèle à l'autre. Soit $A = \{a_1, a_2, \ldots, a_n\}$ et $B = \{b_1, b_2, \ldots, b_n\}$ deux modèles constitués de n atomes, où chaque a_i et chaque b_i est un triplet contenant les coordonnées tridimensionnelles du i^e atome de chaque modèle. La déviation RMS est :

$$\sqrt{\frac{\sum_{i=1}^{n}(a_i - b_i)^2}{n}}$$

En modélisation moléculaire, l'unité de mesure de distance popularisée (au sens euclidien du terme) est l'Angstrom (Å), équivalent à 1×10^{-10} m. Pour obtenir la mesure adéquate de déviation RMS entre deux modèles, il faut préalablement les aligner l'un sur l'autre dans l'espace de façon optimale. Ainsi, la déviation RMS mesure la plus petite distance possible entre les deux modèles, peu importe leur orientation relative dans l'espace. Dans le reste de ce travail, toute référence à la déviation RMS s'adresse à cette métrique de distance particulière.

1.3 Présentation de *MC-Sym*

MC-Sym est un système de modélisation de la structure tertiaire de l'ARN. Cette section en présente les grandes lignes de fonctionnement pour paver la route vers le système de modélisation développé dans ce travail. Par ailleurs, nous décortiquerons sa conception au chapitre suivant.

1.3.1 Bases de données de modélisation discrète

MC-Sym se catégorise par sa méthode de modélisation dite discrète : un modèle est construit nucléotide par nucléotide par application successive des relations d'appariement connues entre les bases azotées. En effet, la transformation linéaire contenue dans la relation d'appariement permet de positionner dans l'espace une base azotée (ou un nucléotide entier) par rapport à une autre en reflétant l'appariement de bases modélisé (voir la sous-section 1.2.2). De plus, chaque nucléotide est discrétisé par un ensemble de modèles d'un nucléotide entier figé dans différentes conformations stéréochimiques particulières. L'idée est

d'apprendre un ensemble de conformations et de relations d'appariement en analysant des modèles de référence pour pouvoir ensuite réutiliser spécifiquement chacune de ces conformations et chacune de ces relations dans la création de nouveaux modèles. Ainsi, *MC-Sym* possède deux bases de données : une base de relations et une base de nucléotides.

1.3.1.1 Base de relations

La base de relations est créée par l'annotation automatique [29] des modèles d'ARN de référence qui proviennent des bases de données publiques *PDB* [46] et *NDB* [47]. Ces modèles de référence ont été déterminés par cristallographie aux rayons X ou par RMN. L'annotation automatique d'un modèle permet d'accoler une étiquette nommant chacune des relations d'appariement extraites entre chaque paire de bases azotées des modèles analysés. L'étiquette a deux sections : une concernant l'adjacence, l'autre concernant le positionnement spatial relatif des bases azotées. L'adjacence est binaire – présente ou non – dépendamment de la position relative des deux nucléotides dans la chaîne polynucléotidique. Le positionnement spatial relatif des bases se subdivise en deux catégories : empilées ou appariées (voir figure 5). Les bases empilées peuvent être inversées ou non, par rapport à la direction du squelette. Typiquement, les bases adjacentes dans une tige sont empilées de façon non-inversée, mais deux bases non-adjacentes peuvent très bien s'empiler de façon inversée ou non. Quant aux bases appariées, elles sont nommées par leurs interfaces en contact par pont-H et par l'orientation *cis* ou *trans*, tel que présenté à la figure 3.

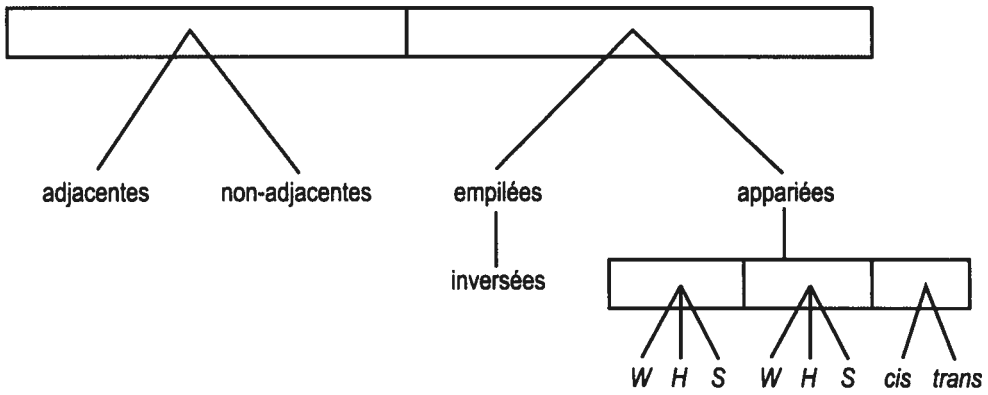


Figure 5 : Organigramme d'une étiquette de relation d'appariement entre deux bases azotées. Les niveaux représentent une restriction de plus en plus fine dans les qualificatifs. Le raffinement du qualificatif *appariées* créé une sous-étiquette spécifiant les interfaces et l'orientation du squelette.

De cette façon, la base de relations contient diverses relations d'appariement, chacune étant étiquetée par un ensemble de qualificatifs. La requête à la base de données mentionne un ou plusieurs de ces qualificatifs en utilisant des opérateurs de logique binaire tels l'union, l'intersection ou la négation. Le résultat est l'ensemble de toutes les relations d'appariement trouvées dans la base de données dont l'étiquette respecte l'équation logique des qualificatifs mentionnés. L'ensemble résultats peut être filtré selon un critère de similitude pour réduire sa taille et éliminer les redondances sans trop affecter sa couverture. C'est l'échantillonnage de l'ensemble résultats. Il se spécifie soit de façon absolue, i.e. par un décompte précis des relations de l'ensemble résultats, ou de façon proportionnelle à l'ensemble complet.

1.3.1.2 Base de nucléotides

La base de nucléotides est aussi créée par l'annotation automatique des modèles de référence de *PDB* et de *NDB*. Cette fois les coordonnées 3D de chaque nucléotide sont directement extraites des modèles, et annotées selon l'un des dix modes de *puckering* du ribose, l'une des deux orientations de la base azotée par rapport au ribose et l'un des quatre types de bases azotées. Les dix modes de *puckering* identifient la conformation particulière du furanose *endo* ou *exo* selon l'un des cinq atomes du cycle (C1', C2', C3', C4' ou O4'). L'orientation de la base par rapport au ribose est *syn* ou *anti*, selon l'angle de torsion du lien glycosyl (le lien covalent C1'-N9 dans les purines ou C1'-N1 dans les pyrimidines) [19]. Ainsi, la base de nucléotides contient divers modèles de nucléotides entiers, chacun étant étiqueté par le type de sa base azotée et par ses deux qualificatifs de conformation : mode de *puckering* et orientation glycosylique. Le fonctionnement des requêtes et l'échantillonnage de l'ensemble résultats est similaire à la base de relations.

1.3.2 Engin de modélisation

L'engin de modélisation est le cœur du fonctionnement de *MC-Sym*. Il relie l'information structurale donnée en entrée par l'usager aux bases de relations et de nucléotides pour construire des modèles par exploration des différentes combinaisons d'assignations de conformations de nucléotides et de relations d'appariement. La figure 6 schématise le flot de donnée dans le système *MC-Sym* complet. Ainsi, l'information structurale est représentée par un ensemble de requêtes aux bases de relations et de nucléotides dont les ensembles résultats

seront parcourus par l'engin de modélisation dans un ordre particulier pour construire les modèles. Cette information se représente habilement par l'annotation du graphe de relation de la structure à modéliser.

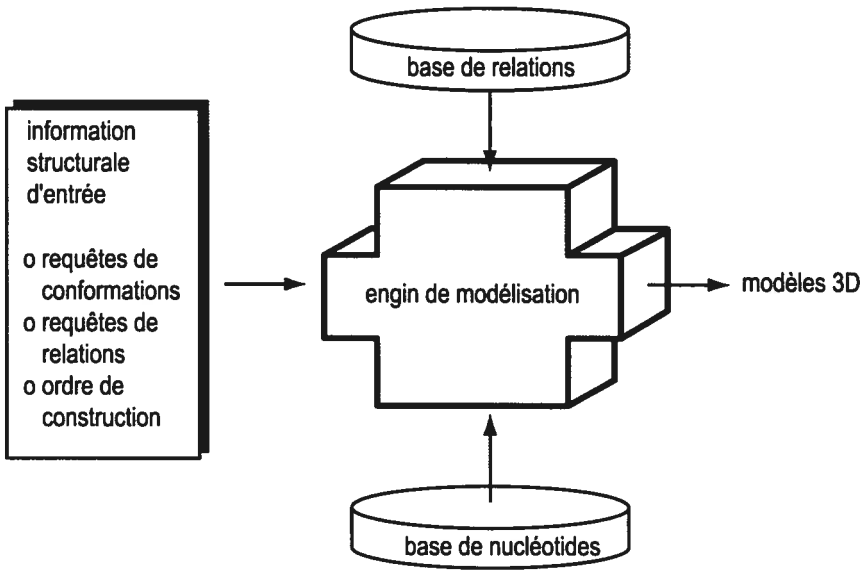


Figure 6 : Schéma conceptuel du flot d'information passant à travers le système *MC-Sym*. L'usager est responsable de l'information structurale en entrée, soient la spécifications de toutes les requêtes au bases de relations et de nucléotides avec leur échantillonnage respectif, ainsi que l'ordre d'application des relations. L'engin de modélisation se charge de créé l'espace de recherche en questionnant les bases de relations et de nucléotides, et de l'explorer pour construire les modèles.

1.3.2.1 Graphes de relation et format des données d'entrée

Un **graphe de relation** d'un modèle d'ARN est un graphe où les nœuds sont les nucléotides et les arêtes les relations d'appariement. Soit $G = (S, A)$ un tel graphe, où S est l'ensemble des nucléotides de la séquence et A est l'ensemble de tous les appariements de bases et les relations d'adjacence mesurables sous forme de relations d'appariement dans la structure, c'est-à-dire de toutes les relations spatiales internucléotides. À partir du graphe de relation, le modèle décrit peut être reconstruit artificiellement. En effet, les relations d'appariement, c'est-à-dire les arêtes du graphe de relation, permettent de positionner les bases azotées par paires : sachant la position spatiale d'une des deux bases, l'autre se positionne par l'application de la transformation contenue dans la relation d'appariement dans le référentiel local de la base originale. Ainsi, par le positionnement d'une première base du modèle à l'origine du référentiel global, les bases suivantes se positionnent les unes par rapport aux

autres par application successive des relations d'appariement qui les interconnectent dans le graphe de relation. Par conséquent, un arbre de recouvrement R du graphe de relation G, $R=(S,A')$ où $A' \subset A$, sélectionne un sous-ensemble des relations d'appariement qui est nécessaire et suffisant au positionnement successif de toutes les bases du modèle. De plus, un parcours préfixe de cet arbre depuis une racine arbitraire ordonnance l'application des relations de façon adéquate. Puisque plusieurs arbres de recouvrement existent pour un graphe donné et que la racine de chaque arbre peut être située à une position arbitraire, alors plusieurs ordonnancements différents construisent adéquatement le modèle selon le même graphe de relation.

Si un graphe de relation peut décrire un modèle en particulier, il peut tout aussi bien décrire une structure tertiaire à modéliser. Cependant, les éléments du graphe de relation d'une structure, à la différence de ceux du graphe de relation d'un modèle, sont des ensembles : un nœud est un ensemble de conformations décrivant le même nucléotide et une arête est un ensemble de relations d'appariement décrivant la même relation spatiale internucléotide. Plusieurs modèles différents peuvent donc être construits à l'aide du graphe de relation de la structure en sélectionnant un élément particulier pour chaque nœud et pour chaque arête et en parcourant un arbre de recouvrement du graphe, de la même façon qu'avec le graphe de relation d'un modèle. Dans le contexte de ce travail, toute référence à un graphe de relation concerne cette définition générale du graphe de relation de la structure tertiaire.

MC-Sym utilise le graphe de relation de structure pour représenter l'information structurale à modéliser. En nommant chaque nœud et chaque arête selon la syntaxe des requêtes aux bases de nucléotides et de relations, les ensembles résultats qui en sont retirés sont directement associés au graphe. La spécification de l'échantillonnage pour chaque requête fixe la taille de chaque ensemble dans le graphe. Avec la spécification d'un arbre de recouvrement et de sa racine, c'est-à-dire d'un ordre de construction en particulier, l'engin de modélisation parcourt le graphe de relation pour construire les différents modèles possibles. La figure 7 montre un exemple de graphe de relation pour la structure de la tige-boucle D d'un ARN de transfert. Le graphe est annoté selon la syntaxe des bases de données de *MC-Sym*. Cette annotation complète dans laquelle est aussi incluse la spécification de l'arbre de recouvrement choisi pour ordonnancer la construction, consignée dans un fichier texte, constitue le script d'entrée donné à *MC-Sym* par l'usager. Dans cette syntaxe, l'énoncé *residue* contient les requêtes de conformations pour chaque nucléotide. La requête entre accolades s'adresse à tous les

nucléotides numérotés dans l'intervalle spécifié. L'échantillonnage d'une requête en particulier est indiqué à la droite de la requête. Ensuite, l'énoncé *connect* contient les requêtes de relations d'appariement entre toutes les bases azotées adjacentes, c'est-à-dire les relations d'adjacence. Normalement, une relation, c'est-à-dire une arête du graphe, est indiquée par les numéros des deux nucléotides impliqués. Cependant, pour faciliter la spécification de la même requête de relation d'adjacence le long d'une chaîne, seulement les numéros des nucléotides aux extrémités de la chaîne sont nécessaires. Par exemple, la numérotation « 49 53 » spécifie les relations d'adjacence pour les quatre arêtes successives 49-50, 50-51, 51-52 et 52-53. Enfin, l'énoncé *pair* contient les requêtes pour toutes les autres relations d'appariement entre bases azotées strictement non-adjacentes. Évidemment, dans cet énoncé, la numérotation « 49 53 » spécifie directement l'arête reliant les nucléotides 49 et 53, et non la suite des nucléotides adjacents de 49 à 53. Dans la précédente description, tout comme dans le reste de ce travail d'ailleurs, il est important de remarquer que toute notion d'adjacence réfère uniquement à la position relative des nucléotides dans la chaîne polynucléotidique, et non à la définition d'adjacence de la théorie des graphes. Ainsi, dans l'exemple précédent, si les nœuds 49 et 53 sont adjacents par le fait qu'une arête les relie dans le graphe, les nucléotides 49 et 53 ne le sont certainement pas dans la chaîne polynucléotidique. Malheureusement, l'abus de langage est inévitable.

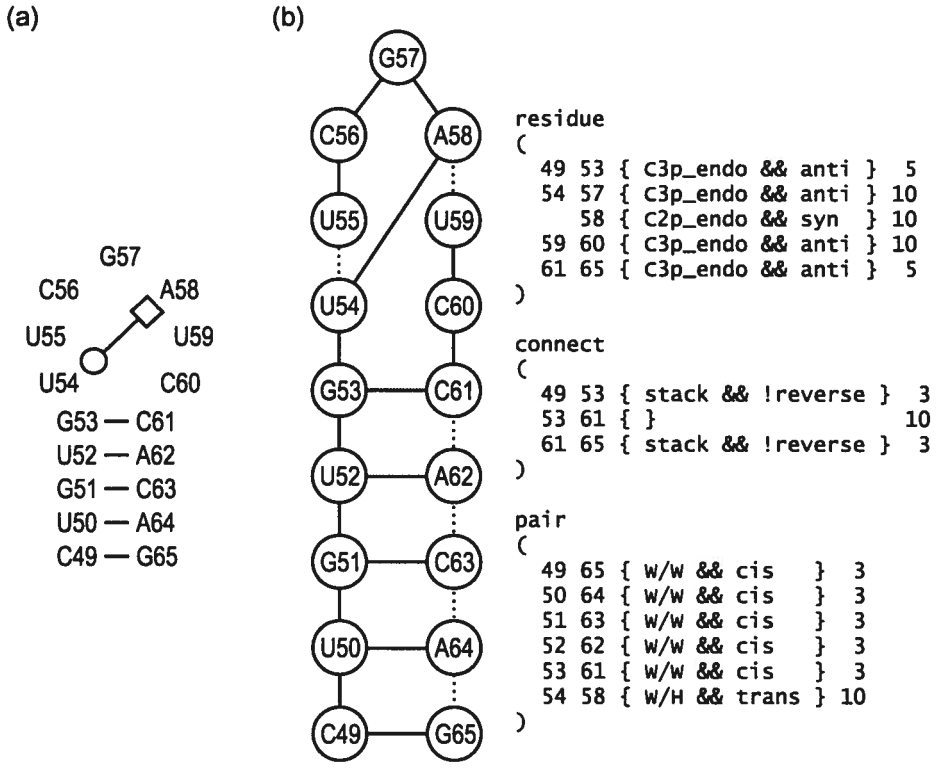


Figure 7 : Exemple de graphe de relation pour la structure de la tige-boucle D d'un ARN de transfert (a) Structure secondaire de la tige-boucle D selon la nomenclature de Leontis & Westhof [30]. La numérotation suit celle de la structure secondaire de l'ARN de transfert au complet. (b) Graphe de relation annoté. L'arbre de recouvrement spécifiant l'ordre de construction est représenté par des arêtes en trait plein, les arêtes en trait pointillé sont donc implicitement représentées. L'annotation des résidus et des relations est présentée selon la syntaxe d'entrée de *MC-Sym*.

Toutes ces spécifications permettent à l'usager une grande flexibilité dans le contrôle du déroulement du processus de modélisation, et du même coup une grande responsabilité lui incombe. D'abord, l'échantillonnage des requêtes a un impact combinatoire direct sur les performances de l'engin de modélisation lors du parcours du graphe et de la construction des modèles. Ensuite, la spécification d'un ordre de construction par l'usager entraîne le contrôle des relations d'appariement qui seront explicitement représentées et de celles qui le seront implicitement. En effet, les arêtes du graphe omises par l'arbre de recouvrement représentent des relations d'appariements qui ne serviront pas explicitement à la construction des modèles. Les bases reliées par une relation implicite seront plutôt positionnées par d'autres relations, donc leur positionnement relatif sera indépendant de la spécification de l'ensemble de relations d'appariement qui les relie dans le graphe. Bref, l'édition du script d'entrée à

MC-Sym par l'usager s'avère un processus assez méticuleux où la philosophie de travail par essais et erreurs règne.

1.4 Présentation du mémoire

Ce mémoire se divise en cinq chapitres. Le présent chapitre étant le premier, il a servi à mettre en contexte la modélisation des molécules d'ARN dans le domaine de la recherche en bioinformatique et à jeter certaines bases théoriques et conceptuelles utiles à la compréhension des prochains chapitres. Plus particulièrement, le premier chapitre a introduit le système de modélisation *MC-Sym* dont le cadre de conception sera utilisé dans notre implantation d'un nouveau système de modélisation. Le chapitre 2 définit en détails notre nouvel engin de modélisation en le comparant à l'engin original de *MC-Sym* à la fois conceptuellement et expérimentalement. Les chapitres 3 et 4 se veulent tous deux des approfondissements de deux modules de notre nouvel engin. Le chapitre 3 analyse la méthode de construction artificielle du ribose que nous avons développée pour résoudre la conformation spatiale du squelette de la structure d'un ARN. Le chapitre 4 étudie l'aspect algorithmique de la modélisation par satisfaction de contraintes et présente un algorithme de recherche probabiliste que nous avons développé pour accroître les performances de modélisation. Le chapitre 5 apporte une conclusion générale à notre travail. En considérant le premier chapitre comme acquis, les chapitres suivants sont présentés de façon indépendante, chacun ayant ses résultats et sa conclusion qui lui sont spécifiques.

Chapitre 2

Engin de modélisation des ARN par relations d'appariement

2.1 Introduction

Avec une notion de la structure d'une molécule d'ARN, le système de modélisation *MC-Sym* crée un espace discret de conformations dont l'exploration construit des modèles tridimensionnels de la structure tertiaire de l'ARN étudié. Plus précisément, la structure à modéliser est construite par l'application successive des relations d'appariement connues entre les bases azotées. Ainsi, dans la conception de l'engin de modélisation du système *MC-Sym*, il existe une méthode formelle pour construire un modèle par l'application successive de relations d'appariement. D'abord, la méthode employée par l'engin original de modélisation de *MC-Sym* sera présentée et critiquée, puis nous décrirons un nouvel engin de modélisation basée sur une nouvelle approche de construction, toujours dans le cadre du système *MC-Sym*. Contrairement à la méthode originale de construction, notre nouvelle méthode sépare la modélisation en deux parties traitées différemment : le placement des bases azotées et des groupements phosphates, et la construction artificielle des riboses. Cette nouvelle méthode de construction implique donc une redéfinition complète de l'engin de modélisation de *MC-Sym*. Les performances de ce nouvel engin seront comparées à l'engin original par deux expérimentations de modélisation controllées : la modélisation de l'anticodon d'un ARN de transfert et la modélisation d'un motif de boucle interne. Finalement, le nouvel engin sera mis à l'épreuve sur la modélisation complète de la structure d'un ARN de transfert.

2.2 Engin de modélisation original de *MC-Sym*

De façon théorique, le système de modélisation *MC-Sym* se divise en trois sections : l'espace conformationnel, la méthode de construction et l'algorithme de recherche. L'espace conformationnel contient l'ensemble des diverses conformations structurales discrètes adoptables par l'ARN modélisé. La méthode de construction régit la façon d'assembler entre eux les éléments de l'espace conformationnel pour obtenir un modèle valide. La validité d'un modèle, partiel ou complet, est assurée par la satisfaction d'un ensemble de contraintes structurales. Quant à l'algorithme de recherche, il applique la méthode de construction de façon itérative pour explorer l'ensemble des conformations valides atteignables dans l'espace conformationnel, et produire des modèles complets de la structure tertiaire.

2.2.1 Espace conformationnel

Dans la définition d'un espace conformationnel, la caractéristique primordiale est la définition de l'élément de base de l'ensemble, le résidu, sur lequel la méthode de construction s'applique directement. Pour *MC-Sym*, un résidu est un nucléotide complet coupé au lien O3'-P (voir figure 8). L'espace conformationnel est partitionné en deux sous-espaces : le sous-espace des relations d'appariement et le sous-espace des conformations statiques.

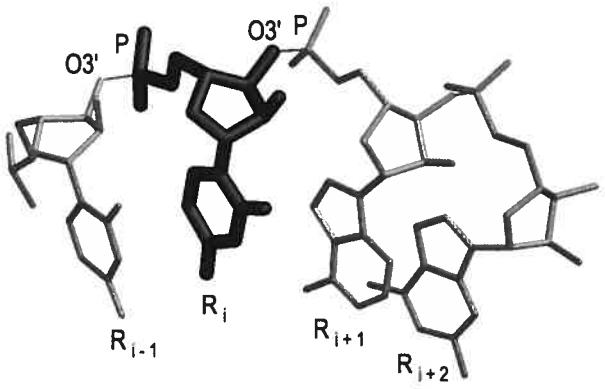


Figure 8 : Subdivision de la chaîne polynucléotidique en résidus et numérotation. L'élément résidu est mis en évidence. La séparation se fait en coupant le lien O3'-P (en trait fin). Images générées par *Molscript* [36] [44] et *Raster3D* [37] [45].

Le sous-espace des relations d'appariement permet à la méthode de construction de positionner un résidu par rapport à un autre par application de la transformation linéaire contenue dans la relation d'appariement. Or, cette transformation de corps rigide n'implique

que les bases azotées; le squelette du résidu, c'est-à-dire son ribose et son phosphate, n'est pas concerné par cette transformation. Dans l'engin de modélisation original, la solution est de déplacer le résidu en entier – base, ribose et phosphate – lors de l'application de la transformation sur la base. Cependant, si les bases azotées peuvent être considérées comme des corps rigides, ce n'est pas le cas du squelette qui peut adopter diverses conformations structurales lors du repliement de la chaîne polynucléotidique. Pour représenter cette diversité, l'espace des conformations du squelette est discrétisé simplement en conservant des modèles de résidu complet adoptant diverses conformations. C'est le contenu du sous-espace des conformations statiques. Ces deux sous-espaces sont en lien direct avec les bases de relations et de nucléotides de *MC-Sym* (voir la sous-section 1.3.1). De cette façon, l'espace conformationnel est créé par les résultats des requêtes à ces bases de données, selon l'information structurale spécifiée par l'usager dans le script d'entrée. Dans ce contexte, la jonction des bases de relations et de nucléotides forme l'espace conformationnel universel, contenant toutes les conformations atteignables par tous les modèles de référence qui ont été analysés pour construire les bases de données. Ainsi, toute spécification particulière restreint la recherche à un sous-ensemble de cet espace universel.

2.2.2 Méthode de construction et contraintes

La méthode de construction définit la façon de positionner un résidu dans l'espace tridimensionnel par rapport à un autre résidu déjà placé qui est considéré comme la référence. La validité de cette construction est assurée par la satisfaction d'un ensemble de contraintes géométriques qui seront décrites plus bas. De façon formelle, soit R_a un résidu déjà positionné dans l'espace tridimensionnel. La méthode de construction qui positionne le résidu R_b sur la référence R_a se divise en quatre étapes :

1. Choisir dans l'espace conformationnel une conformation pour R_b, c'est-à-dire un modèle décrivant entièrement le résidu R_b dans une conformation statique particulière.
2. Positionner R_b sur R_a. Cette étape se subdivise en deux tâches :
    a. Choisir dans l'espace conformationnel une relation d'appariement entre R_a et R_b
    b. Appliquer la transformation linéaire contenue dans la relation choisie sur le résidu complet R_b dans le référentiel de R_a.
3. Vérifier les contraintes applicables.
4. Valider la construction si et seulement si toutes les contraintes sont satisfaites.

Cette méthode de construction est rigoureusement appliquée pour chaque relation de l'ordre de construction spécifié pour la modélisation. Ainsi, en parcourant les arêtes de l'arbre de recouvrement représentant l'ordre de construction (voir la sous-section 1.3.2.1), un modèle est construit résidu par résidu de façon interrelative. L'étape 3 de la méthode permet de conserver une construction toujours valide. Dans ce contexte, les contraintes à valider sont typiquement de deux natures : collisions stériques et distances de fermeture d'adjacence.

Une collision stérique survient lorsque deux atomes de deux résidus différents se retrouvent à une distance inférieure à un seuil critique, représentant l'encombrement stérique atomique. Puisque dans un modèle l'atome est représenté par un simple point, i.e. une coordonnée tridimensionnelle, la distance entre deux atomes se calcule par la longueur du segment reliant les deux points représentatifs. Formellement, une contrainte de collision implique deux résidus : un qui vient d'être positionné par la méthode de construction et un qui l'est déjà. Elle est validée si et seulement si chacun des atomes d'un résidu sont distancés de chacun des atomes de l'autre résidu d'une longueur supérieure au seuil. Le seuil de la distance interatomique représente le paramètre libre d'une contrainte de collision; il est typiquement ajusté à 1Å.

Pour sa part, la contrainte de distance de fermeture d'adjacence vérifie la longueur du lien entre les atomes O3' et P qui relient deux résidus adjacents dans la chaîne polynucléotidique. En effet, ce lien covalent est implicitement créé lors de la construction, dû à la séparation du résidu. Sa longueur doit donc être contrôlée par une contrainte. Tout comme la contrainte de collision, cette contrainte implique un résidu nouvellement positionné et un résidu déjà

positionné, seulement ici la contrainte s'applique uniquement dans le cas où ces deux résidus sont adjacents. Elle est validée si et seulement si la longueur du lien O3'-P reliant les deux résidus est à l'intérieur d'un intervalle prédéterminé. Cet intervalle représente le paramètre libre de la contrainte. En considérant que, dans un modèle de référence, la longueur du lien covalent O3'-P est d'environ 1.6Å, ce paramètre est ajusté en fonction de l'écart accepté pour la longueur mesurée.

Bref, à l'étape 3 de la méthode de construction, si k résidus sont déjà positionnés, k contraintes de collision impliquant le résidu nouvellement positionné et chacun des k autres résidus sont à vérifier. De plus, si le résidu positionné est adjacent à un des k résidus déjà positionnés, une contrainte de distance de fermeture d'adjacence impliquant ces deux résidus adjacents est à vérifier à cette étape.

Le paramétrage de ces contraintes est spécifié par l'usager dans le script d'entrée. Bien que la vérification des contraintes de collision et de fermeture d'adjacence soit à elle seule suffisante à la validation du modèle construit, d'autres contraintes propres à une modélisation en particulier peuvent être ajoutées : par exemple une distance précise entre deux atomes particuliers.

2.2.3 Algorithme de recherche

Au cœur de l'engin de modélisation, l'algorithme de recherche applique systématiquement la méthode de construction décrite à la sous-section précédente pour naviguer d'un résidu à l'autre dans l'espace conformationnel. Cette définition de l'engin de modélisation où chaque résidu est positionné dans l'espace tridimensionnel itérativement selon un ensemble fini de possibilités, tout en respectant un ensemble de contraintes, cadre parfaitement avec la définition plus générale d'un problème de satisfaction de contraintes. Au chapitre 4, nous analyserons en détails cet aspect conceptuel. La résolution classique d'un problème de satisfaction de contraintes utilise un algorithme de retour arrière (*backtrack*) pour explorer systématiquement chaque combinaison des éléments de l'espace conformationnel et valider les contraintes à chaque étape. Ainsi, lors du positionnement du $i^{ème}$ résidu dans l'ordre de construction, chacune de ces conformations statiques conjointement à chacune de ces relations d'appariement avec sa référence sont tentées. Si une combinaison

conformation/relation particulière positionne le $i^{ème}$ résidu en respectant toutes les contraintes applicables, alors le $(i + 1)^{ème}$ résidu peut à son tour tenter de se positionner, sinon une autre combinaison pour le $i^{ème}$ résidu est tentée. S'il n'en existe plus, alors la recherche retourne en arrière pour positionner le $(i - 1)^{ème}$ résidu d'une autre façon, et poursuivre la recherche. Si le dernier résidu est correctement positionné, alors un modèle complet est obtenu et la recherche se poursuit en repositionnant le dernier résidu. De cette façon, l'algorithme de retour arrière explore l'espace conformationnel de façon exhaustive et y construit tous les modèles possibles en fonction du paramétrage des contraintes.

En se référant à la méthode de construction décrite à la sous-section précédente, si c_b est la quantité de conformations possibles pour le résidu R_b et r_b est la quantité de relations applicables entre la référence R_a et le résidu R_b, alors il y a $c_b r_b$ façons différentes de positionner R_b sur R_a. Ainsi, pour une chaîne polynucléotidique de n nucléotides (donc n résidus), la taille de l'espace conformationnel est $\prod_{i=1}^{n} c_i r_i$, où $r_1 = 1$ puisque le premier résidu est placé sans application de relation d'appariement; il représente la référence globale. Cette taille est donc exponentielle de second degré : un degré en terme de la taille du sous-espace des conformations et un degré en terme de la taille du sous-espace des relations. Puisque la recherche par retour arrière explore l'espace conformationnel au complet, sa taille a un impact direct sur le temps de recherche. Une alternative probabiliste au retour arrière positionne itérativement chaque résidu en choisissant une combinaison conformation/relation valide de façon aléatoire. Cette approche parvient à naviguer plus aisément dans les différentes régions de l'espace conformationnel même lorsqu'il est très vaste; cependant la recherche est non-exhaustive et potentiellement redondante. Le chapitre 4 analysera et comparera différents algorithmes de recherche.

2.3 Nouvel engin de modélisation

Dans le contexte du système *MC-Sym*, nous avons défini une approche différente de modélisation. Cette approche s'appuie sur un raisonnement simple, basée sur deux faits et une hypothèse :

Fait #1 : Les bases azotées sont rigides et planaires, leur modélisation est statique.
Fait #2 : Le ribose est flexible, son cycle furanosique peut adopter différentes conformations stéréochimiques.
Hypothèse : Le repliement tridimensionnel de la chaîne polynucléotidique est dirigé par des interactions par ponts-H entre les bases azotées, c'est-à-dire par des appariements de bases.
Conclusion : La conformation adoptée par le squelette est une *conséquence* des appariements de bases, et non une *cause*.

L'engin original de modélisation, de par sa méthode de construction, ignore un tel raisonnement. En effet, lors du positionnement d'un résidu, le choix de sa conformation est fait de façon indépendante à sa relation d'appariement avec le résidu référence. Par conséquent, la relation doit satisfaire à la conformation statique choisie. Cependant, la relation appliquée positionne correctement les bases des résidus reliés, mais s'intéresse peu à la conformation de leur ribose. Cette incohérence fait en sorte qu'il devient difficile de trouver une conformation statique pour le résidu qui concordera avec le positionnement par relation entre les bases azotées. C'est d'ailleurs la raison pour laquelle l'engin original doit se prémunir d'une contrainte de distance de fermeture du lien covalent O3'-P entre les résidus adjacents. Or, nous avons démontré expérimentalement que cette contrainte est rarement validée. L'expérimentation consiste à utiliser la méthode de construction de l'engin original de modélisation pour positionner des résidus adjacents en utilisant comme espace conformationnel toutes les conformations des nucléotides conjointement à toutes les relations d'appariement entre deux bases adjacentes qui existent dans les bases de données de *MC-Sym*. Pour chaque combinaison conformation/relation, la longueur du lien implicite O3'-P mesurée par la contrainte de fermeture d'adjacence est rapportée. La figure 9 montre l'histogramme de la distribution de ces mesures de distance d'adjacence. Ainsi, si l'erreur acceptée est de ±1Å (sur un lien optimal de 1.6Å), seulement 10.4% des constructions valident la contrainte. Une relaxation de cette contrainte à ±2Å valide seulement 23.4% des constructions. En fait, près de 50% des constructions présentent une distance de fermeture d'adjacence supérieure à 5Å. Cette proportion est énorme, d'autant plus en se rappelant que toutes les relations d'appariement utilisées dans l'expérimentation ont été extraites de modèles de référence où la distance de fermeture d'adjacence est toujours optimale.

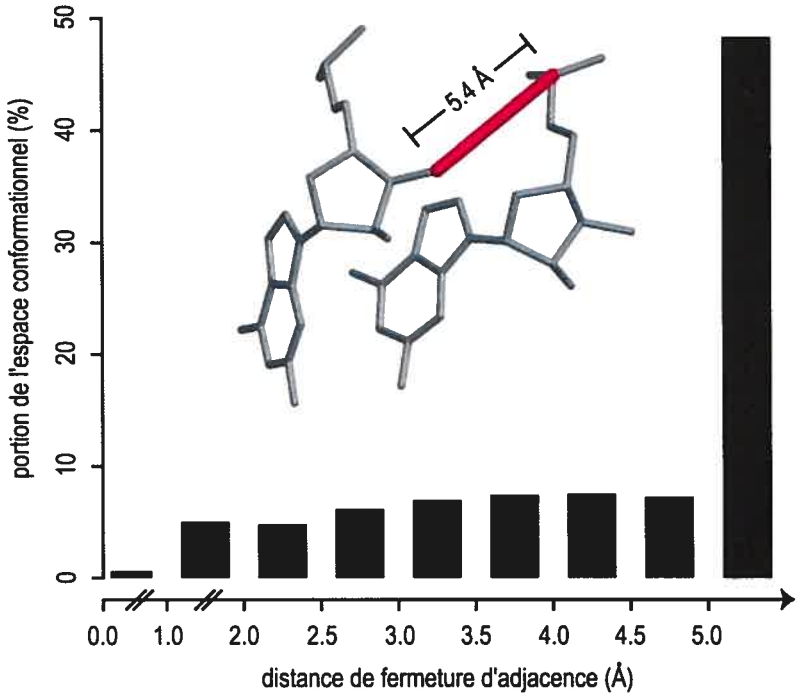


Figure 9 : Distribution des distances de fermeture d'adjacence pour l'engin original de modélisation. La construction d'un nucléotide adjacent a été tentée pour toutes relations d'appariement adjacentes et pour toutes conformations de nucléotides, et la distance du lien implicite O3'-P a été mesurée sur chaque construction.

Donc, l'engin original de modélisation présente une forte tendance à rejeter des modèles dont les bases seraient correctement placées mais où le squelette morcelé en conformations statiques de ribose n'arrive pas à se relier d'un résidu à l'autre, alors que sa flexibilité théorique devrait faire en sorte qu'il puisse s'adapter à un ensemble de bases cohérentes. Le problème avec l'approche où un nucléotide complet est positionné par application d'une relation d'appariement est qu'elle suppose faussement que la conformation spatiale du ribose et du phosphate est indépendante de la position relative de la base azotée, ignorant le raisonnement exposé précédemment.

Notre nouvelle approche de modélisation met donc en premier plan le placement cohérent des bases azotées par relation d'appariement. Seule la base azotée est déplacée par transformation. Le positionnement du squelette se fait par approximation numérique de sa conformation spatiale. Pour réaliser une telle approximation, nous avons développé une méthode de construction artificielle d'un ribose. Cette méthode construit un ribose atome par atome depuis une base azotée jusqu'aux deux phosphates adjacents de part et d'autre. Le chapitre suivant couvrira en détails la description et l'analyse de cette méthode. Bref, la

nouvelle méthode de construction de l'engin de modélisation se résume par les trois étapes suivantes : 1) positionnement des bases, 2) positionnement des phosphates et 3) construction des riboses. En plus de proposer une nouvelle méthode de construction des résidus, cette nouvelle approche redéfinit aussi l'espace conformationnel en scindant le nucléotide en trois parties indépendantes : base azotée, phosphate et ribose. Ainsi, un nouvel engin de modélisation est créé pour *MC-Sym*.

2.3.1 Redéfinition de l'espace conformationnel

Placer de façon indépendante bases, phosphates et riboses, c'est redéfinir l'élément de base de l'espace conformationnel : le résidu. Dans notre nouvel engin de modélisation, trois types de résidus sont considérés : la base azotée, coupée au lien glycosyl (N9-C1' dans les purines et N1-C1' dans les pyrimidines), le groupement phosphate, coupé aux liens C3'-O3' et C5'-O5', et finalement le ribose, ce qui reste du nucléotide (voir figure 10a). Les riboses sont construits artificiellement alors que les bases et les phosphates sont considérés comme des structures rigides de conformation stéréochimique unique. Le sous-espace des conformations des résidus est donc réduit à un seul exemplaire de modèles pour une base A, G, C ou U et pour un groupement phosphate. De ce fait, ce sous-espace est retiré de l'espace conformationnel puisqu'il ne fait plus partie de la recherche. Le sous-espace des relations d'appariement est conservé, cependant un supplément d'information est nécessaire pour le positionnement des phosphates. Au moment de la création de la base de relations dans le système *MC-Sym*, lors de l'extraction d'une relation d'appariement entre deux bases azotées adjacentes dans la chaîne polynucléotidique du modèle de référence analysé, la transformation linéaire qui exprime le positionnement spatial du groupement phosphate qui se retrouve entre les deux bases est aussi extraite, dans le même référentiel. La figure 10b illustre l'extraction de la transformation du phosphate conjointement à la transformation de la base azotée. Toutes deux sont conservées dans la relation d'appariement qui sera stockée dans la base de relations. Ainsi, lorsque la méthode de construction applique une relation d'appariement adjacente, elle positionne à la fois une base azotée sur une autre de référence et un groupement phosphate entre les deux. Bref, le nouvel espace conformationnel est uniquement peuplé de relations d'appariement augmentées par la transformation du phosphate.

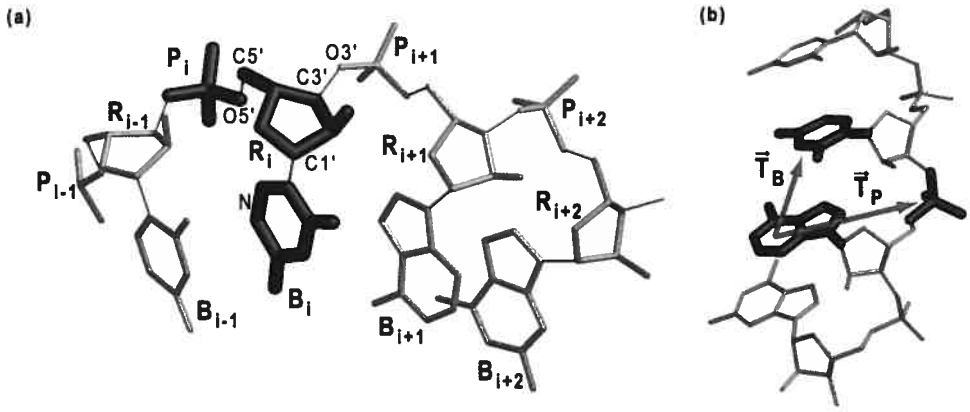


Figure 10 : Redéfinition du résidu et extraction de la transformation vers le phosphate. (a) Nouvelle subdivision de la chaîne polynucléotidique en trois types de résidus et numérotation. Les différents types de résidus – base azotée (B), groupement phosphate (P) et ribose (R) – sont mis en évidence. Les coupures de séparation sont présentées en trait fin. (b) Extraction de la transformation du phosphate ($\vec{T}_P$) dans le même référentiel que la transformation de la base azotée ($\vec{T}_B$). Chaque relation d'appariement adjacente contient maintenant ces deux transformations. Images générées par *Molscript* [36] [44] et *Raster3D* [37] [45].

2.3.2 Redéfinition de la méthode de construction

La redéfinition de la notion de résidu dans l'espace conformationnel implique nécessairement la redéfinition de la méthode de construction. Dans ce nouvel engin de modélisation, les bases et les phosphates sont d'abord positionnés selon les diverses relations d'appariement disponibles dans l'espace conformationnel, puis ils sont interconnectés par la construction artificielle de riboses. La méthode présentée ici construit le ribose dès qu'il est possible de le faire, c'est-à-dire dès que la base azotée et les deux phosphates qu'il interconnecte sont positionnés, pour assurer un maximum de cohérence à tout moment lors de la construction progressive des modèles par l'algorithme de recherche. De façon générale, la nouvelle méthode de construction se divise en deux cas particuliers : le positionnement d'une base adjacente à sa référence et le positionnement d'une base non-adjacente à sa référence. Avant de décrire le fonctionnement de ces deux méthodes particulières, il convient de décrire deux sous-méthodes de construction utilisées dans les deux cas : la *fermeture* et la *construction artificielle du ribose.*

2.3.2.1 Fermeture

Avant de décrire formellement les méthodes de construction, il convient de remarquer une faille dans l'approche de modélisation où les bases azotées et les phosphates sont positionnés par relation d'appariement et interconnectés par la construction artificielle de riboses. Pour qu'un phosphate en particulier soit positionné par la méthode de construction, il faut nécessairement que la relation d'adjacence qui contient sa transformation soit utilisée, c'est-à-dire que les bases adjacentes de part et d'autre du phosphate soient effectivement positionnées par application de la relation d'appariement qui les relie. Or, dans la spécification de l'ordre de construction par l'usager, seulement un sous-ensemble des relations d'appariement du graphe de relation est sélectionné pour former un arbre de recouvrement. Par conséquent, il est possible que certaines relations d'appariement adjacentes ne soient pas parcourues par un arbre de recouvrement en particulier. La figure 11 reprend l'exemple du graphe de relations de la structure de la boucle D d'un ARN de transfert pour illustrer les phosphates qui sont explicitement positionnés par l'application d'une relation d'adjacence, et ceux qui ne le sont pas. Le rôle de la sous-méthode de fermeture est de positionner ces phosphates manquants.

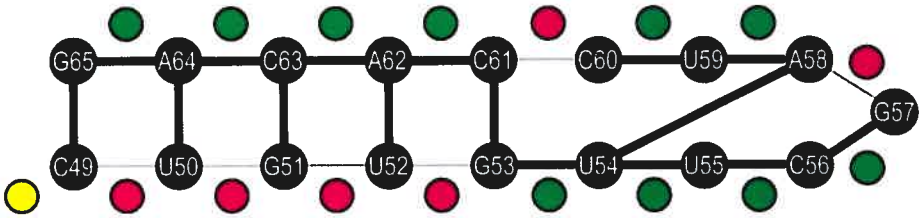


Figure 11 : Les phosphates à placer pour un ordre de construction de la structure de la tige et la boucle D de ARNt^{Phe}. L'arbre de recouvrement du graphe de relations est présenté en traits gras. Entre chaque paire de bases azotées adjacentes, un phosphate doit être placé. Les phosphates en vert sont explicitement positionnés par application de la relation d'appariement adjacente incluse dans cet arbre de recouvrement. Les phosphates en rouge ne sont pas positionnés explicitement puisque l'arête du graphe correspondant à leur relation d'adjacence n'est pas incluse dans l'arbre de recouvrement. Ces phosphates sont positionnés par la sous-méthode de fermeture. Le phosphate en jaune est un cas particulier puisque sa base azotée associée est la première de la chaîne. Il est donc placé d'une façon statique par la construction artificielle du ribose avoisinant.

Bref, une sous-méthode de fermeture s'applique à tout moment de la construction dès qu'une base est positionnée par une méthode de construction et qu'une des deux bases adjacentes à celle-ci (ou les deux à la fois) avait déjà été positionnée préalablement, en excluant le cas où cette dernière est la référence du positionnement. En fait, la fermeture est nécessaire dès que

deux bases adjacentes se retrouvent positionnées dans la construction du modèle sans que leur relation d'adjacence n'ait été explicitement utilisée. Le rôle de la fermeture est de positionner le phosphate manquant dû à l'inutilisation de cette relation. L'idée est d'extraire la relation d'appariement implicite formée par le positionnement indépendant des deux bases adjacentes et de la comparer à l'ensemble des relations qui a été spécifié dans l'espace conformationnel pour représenter la relation d'adjacence inutilisée. De cette comparaison ressort la relation de l'espace conformationnel la moins distante de la relation implicite extraite. Cette relation contient la transformation pour placer le phosphate manquant. La notion de distance de relation d'appariement est une métrique définie par Gendron et al. [29]. Elle est basée sur une pondération de l'apport en translation et en rotation de la transformation linéaire, fonction de la longueur de la translation et de l'angle de rotation. De cette façon, la sous-méthode de fermeture résout le problème des phosphates non explicitement positionnés.

Cependant, à chaque appel à cette sous-méthode de construction, chaque relation d'appariement de l'ensemble qui représente la relation d'adjacence inutilisée par la spécification de l'ordre de construction est comparée à la relation implicite formée par les bases adjacentes dans leur position courante. Par conséquent, la spécification de l'échantillonnage de cette relation, de la même façon qu'avec les relations explicitement utilisées par l'ordre de construction, a un impact direct sur l'efficacité de la recherche. D'un point de vue exhaustif, l'ensemble complet de toutes les relations d'appariement de la base de relations de *MC-Sym* devrait être considéré pour positionner correctement le phosphate. Par contre, une réduction de l'échantillonnage viendrait accélérer directement chaque appel à la sous-méthode de fermeture. L'idée est d'obtenir l'échantillonnage minimal qui conserve l'exhaustivité de la recherche. Pour ce faire, nous avons réalisé une expérimentation de modélisation. La structure à modéliser consiste en une boucle terminale de cinq nucléotides. Pour chaque modélisation, un échantillonnage différent de la relation d'adjacence inutilisée qui ferme la boucle a été spécifié. Les critères de comparaison sont la quantité de modèles trouvés et le temps écoulé pour compléter la recherche. Le tableau 1 résume les performances de ces modélisations. Selon ces résultats, un échantillonnage de 25% présente une accélération de la recherche de 77% par rapport à un échantillonnage complet. Le coût de cette accélération est la perte de 3% des modèles, ce qui est plutôt raisonnable. Par contre, une réduction supplémentaire de l'échantillonnage implique de trop grande perte de modèles

lorsque l'exhaustivité de la recherche est souhaitée, même si l'accélération du temps de recherche y gagne beaucoup.

échantillonage (%)	temps de recherche (sec)	accélération de la recherche (%)	modèles trouvés	modèles perdus (%)
100	6142	—	237	—
50	3019	51	235	1
25	1407	77	231	3
5	348	94	191	19
1	175	97	160	32

Tableau 1 : Résultats de la modélisation d'une boucle terminale avec utilisation de la sous-méthode de fermeture à différents échantillonnages de la relation implicite. Tous les résultats sont relatifs à ceux de la première ligne qui représente la modélisation utilisant l'échantillonnage complet. La ligne ombrée représente le meilleur choix d'échantillonnage.

2.3.2.2 Construction artificielle du ribose

Un ribose est construit par approximation numérique des positions de ces atomes. Le ribose est construit à l'intérieur de trois points d'ancrage fixes : la base associée et les deux phosphates reliés de part et d'autre de la base. L'idée est de positionner un à un les atomes du ribose depuis l'azote du lien glycosyl de la base (N9 pour les purines, N1 pour les pyrimidines) jusqu'aux atomes de phosphore des groupements phosphates. Cette méthode de construction artificielle est paramétrée selon la pseudorotation du cycle furanosique [20] et la torsion du lien glycosyl. Un algorithme réalise l'approximation numérique des paramètres qui optimisent la qualité du ribose construit, mesurée par l'erreur du positionnement artificiel des phosphates. Au chapitre suivant, nous exposerons en détails cette sous-méthode de construction et analyserons sa représentativité et ses performances. Cette sous-méthode s'applique à tout moment de la construction dès qu'une base et que deux phosphates peuvent être interconnectés par un ribose. Il y a le cas particulier des bases situées aux extrémités de la chaîne polynucléotidique. Dès qu'une base débutant ou terminant la chaîne ainsi que le phosphate adjacent interne à la chaîne sont placés, le ribose se construit sans tenir compte du phosphate manquant à l'extérieur de la chaîne, puisque celui-ci ne sera jamais positionné.

2.3.2.3 Méthode de construction d'une relation d'adjacence

Cette méthode régit le positionnement d'une base selon une relation d'adjacence. Puisque cette relation d'appariement est adjacente, elle contient aussi la relation pour positionner le phosphate entre la base positionnée et celle de référence. Positionner la base et le phosphate par application des transformations contenues dans la relation est l'étape de base de la méthode. Ensuite, puisque de nouveaux résidus ont été ajoutés à la construction du modèle, il faut vérifier l'application de deux règles concernant respectivement l'intervention des sous-méthodes de fermeture et de construction artificielle du ribose. La première règle vérifie si la base adjacente à la base positionnée par cette méthode de construction avait déjà été positionnée, le cas échéant impliquant le positionnement du phosphate manquant par fermeture. La deuxième règle vérifie si certains riboses peuvent être construits à ce moment à cause de l'ajout de ces nouveaux résidus. Un ribose est construit si la base et les deux phosphates qu'il interconnecte sont tous trois positionnés. Ici, trois riboses peuvent être potentiellement construits, dépendamment des phosphates avoisinants : un sur la base de référence, un sur la base positionnée par la relation et un sur la base successive, s'il y a eu fermeture. La figure 12 schématise et formalise une méthode d'ordonnancement des différentes sous-constructions exigées par ces deux règles. Cette méthode d'ordonnancement est traitée légèrement différemment dépendamment du sens de la relation d'adjacence par rapport à celui de la chaîne polynucléotidique. Dans les deux cas, les règles ci-haut mentionnées s'appliquent de façon conceptuellement identique, seule la numérotation des résidus diffère. Le cas où une base est située à l'extrémité de la chaîne est aussi traité.

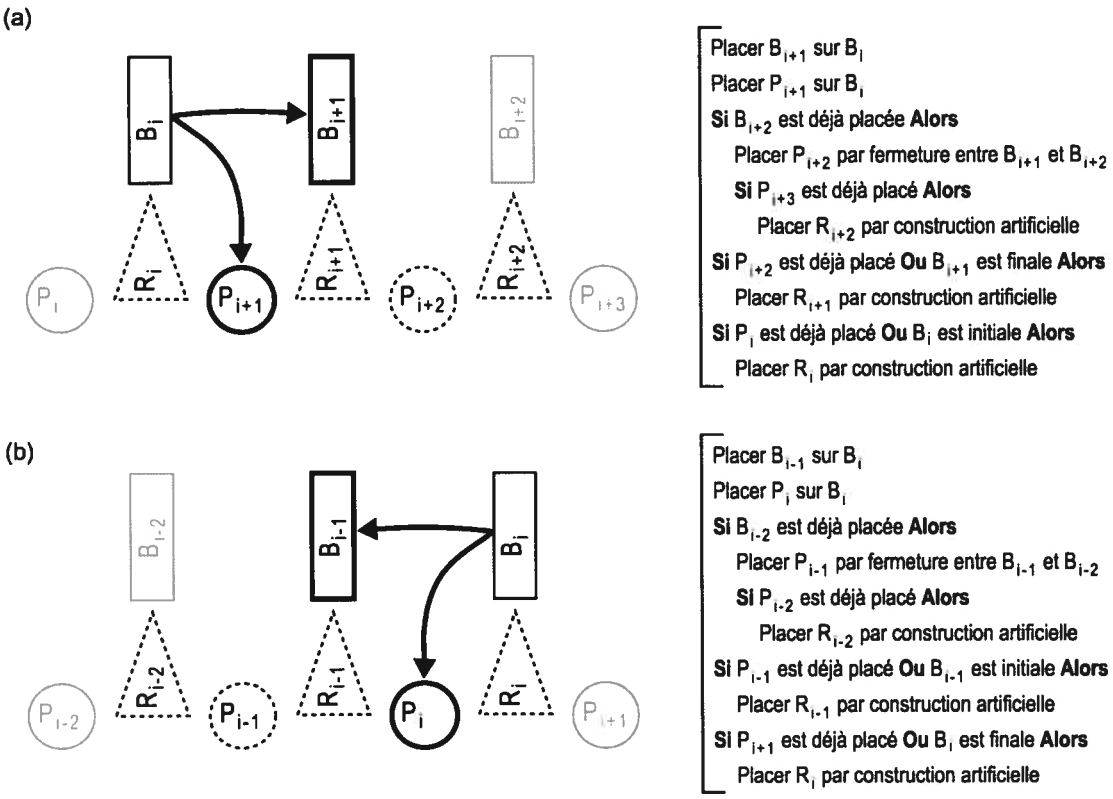


Figure 12 : Schéma de la méthode de construction d'une relation d'adjacence et méthode d'ordonnancement. Les résidus en trait gras sont placés directement par application de la relation d'appariement selon la base azotée de référence en trait normal. Les résidus en trait pointillé sont placés conditionnellement à la présence des résidus en trait ombré; leur construction est ordonnancée par la méthode formalisée à droite. En (a) est présenté le cas particulier où la relation d'adjacence est dans le même sens que la chaîne polynucléotidique (de B_i à B_{i+1}), alors qu'en (b) est présentée le cas où la relation est dans le sens inverse (de B_i à B_{i-1}). La différence dans le traitement est uniquement au niveau de la numérotation des résidus, le concept d'ordonnancement reste le même.

2.3.2.4 Méthode de construction d'une relation de non-adjacence

Cette méthode régit le positionnement d'une base selon une relation strictement de non-adjacence. Pour une relation de non-adjacence, la règle d'application de la sous-méthode de fermeture diffère. En effet, la non-adjacence de la relation implique que jusqu'à deux méthodes de fermeture sont applicables, dans chacune des directions d'adjacence de la base placée. Par conséquent, jusqu'à trois riboses pourront être construits : un sur la base positionnée et un sur chacune des deux bases adjacentes de part et d'autre de la base positionnée par la relation. La règle d'application de la construction du ribose est la même : il

est construit dès que la base et les phosphates associés sont positionnés. La figure 13 schématise et formalise la méthode d'ordonnancement des résidus potentiellement positionnés.

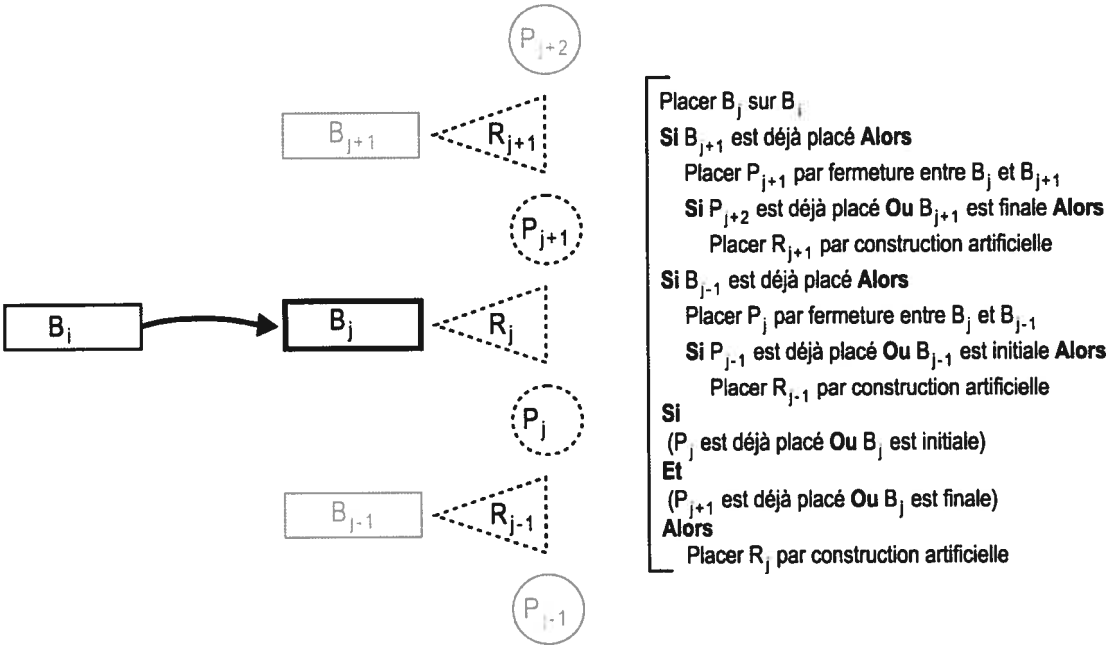


Figure 13 : Schéma de la méthode de construction d'une relation de non-adjacence et méthode d'ordonnancement. Seule la base en trait gras est placée directement par application de la relation d'appariement selon la référence en trait normal. Les résidus en trait pointillé sont placés conditionnellement à la présence des résidus en trait ombré; leur construction est ordonnancée par la méthode formalisée à droite.

2.3.3 Redéfinition des contraintes

Dans notre nouvel engin de modélisation, deux contraintes sont nécessaires et suffisantes à la validation de la construction : la contrainte de collision et la contrainte de qualité de la construction du ribose. De plus, nous avons étudié l'application d'une contrainte supplémentaire : la contrainte de fermeture.

2.3.3.1 Contrainte de collision

La contrainte de collision vérifie l'encombrement stérique des atomes. Cette contrainte est traitée de la même manière qu'avec l'engin original de modélisation : un seuil sur la distance

interatomique est vérifié chaque fois qu'un nouveau résidu est ajouté à la construction. Une amélioration a été apportée au processus de vérification des collisions : le résidu phosphate est considéré comme un unique atome centré sur le phosphore. Le seuil est ajusté selon le rayon de la sphère centrée au phosphore et englobant les oxygènes du groupement (voir figure 14). Pour les trois types de résidus, six combinaisons sont possibles pour le traitement de la contrainte de collision, dont trois font intervenir le résidu phosphate. La combinaison base/phosphate et ribose/phosphate bénéficie d'une réduction de 80% des calculs de distances interatomiques impliqués par la validation de la contrainte lorsque le phosphate est traité de cette manière (de cinq fois la quantité d'atomes dans une base ou un ribose à une fois la même quantité). Pour sa part, la combinaison phosphate/phosphate bénéficie d'une réduction de 96% (de vingt-cinq calculs de distance à un seul). Évidemment, les trois autres combinaisons possibles – base/base, base/ribose et ribose/ribose – ne bénéficient d'aucune réduction de la quantité de calculs de distance puisqu'elles n'impliquent pas le phosphate. Donc, en moyenne, la simplification du groupement phosphate à un seul atome dont le seuil stérique créé une sphère englobante réduit de 43% la quantité totale de calculs de distances interatomiques impliqués par la validation des contraintes de collisions.

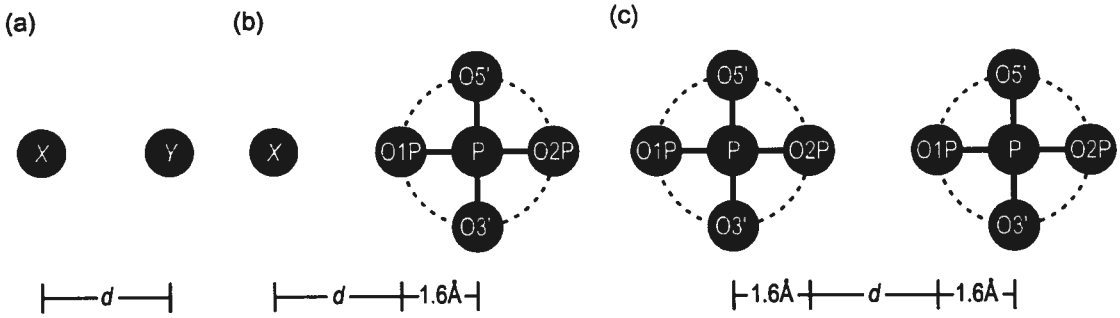


Figure 14 : Seuil sur la distance entre atomes et groupements phosphates. Le seuil stérique du groupement phosphate est une sphère centrée au phosphore et englobant le groupement entier (en trait pointillé. (a) Seuil entre deux atomes. Habituellement d = 1.0Å. (b) Seuil entre un atome et un groupement phosphate. (c) Seuil entre deux groupements phosphates.

2.3.3.2 Contrainte de qualité de la construction du ribose

La contrainte de qualité de la construction du ribose vérifie l'erreur de fermeture du ribose artificiellement construit depuis la base azotée jusqu'aux deux phosphates adjacents de part et d'autre. En effet, la construction artificielle relie implicitement les phosphates au ribose. L'erreur encourue est quantifiée par la mesure des liens covalents C5'-O5' et C3'-O3'

implicitement créés. Cette contrainte remplace la distance de fermeture d'adjacence O3'-P entre les nucléotides, contrainte utilisée par l'engin original.

2.3.3.3 Contrainte de fermeture

D'un point de vue structural, les deux contraintes ci-haut mentionnées sont à elles seules suffisantes à la validation de la construction du modèle. Cependant l'ajout de contraintes supplémentaires peut aider à accélérer le processus d'exploration de l'espace conformationnel. Dans cet ordre d'idée, nous avons implanté une contrainte préalable à l'application de la sous-méthode de fermeture. Le rôle de cette contrainte de fermeture est d'inhiber l'application complète de la sous-méthode de fermeture lorsque les bases adjacentes sont positionnées dans l'espace d'une telle façon qu'elles deviennent trop distantes pour espérer les réunir adéquatement lors de la construction ultérieure des riboses. L'idée est de quantifier l'éloignement entre deux bases adjacentes et de mettre un seuil d'acceptation de l'adjacence implicite. Cette contrainte de fermeture doit être vérifiée préalablement à l'application de la sous-méthode de fermeture. Pour obtenir une métrique d'éloignement de deux bases adjacentes, nous avons étudié diverses mesures sur les bases azotées adjacentes des modèles de références qui sont entrés dans la fabrication de la base de relations de *MC-Sym*. La mesure de la distance entre les atomes C1' respectifs des deux bases adjacentes montre une distribution de valeur très concentrée, comme le témoigne la figure 15 qui montre la fonction de distribution empirique cumulative des mesures de distance C1'-C1'. Ainsi, l'intervalle [20,100] $Å^2$ contient pratiquement toutes les mesures. Donc, une mesure extérieure à cet intervalle invalide la contrainte de fermeture avec un haut degré de certitude.

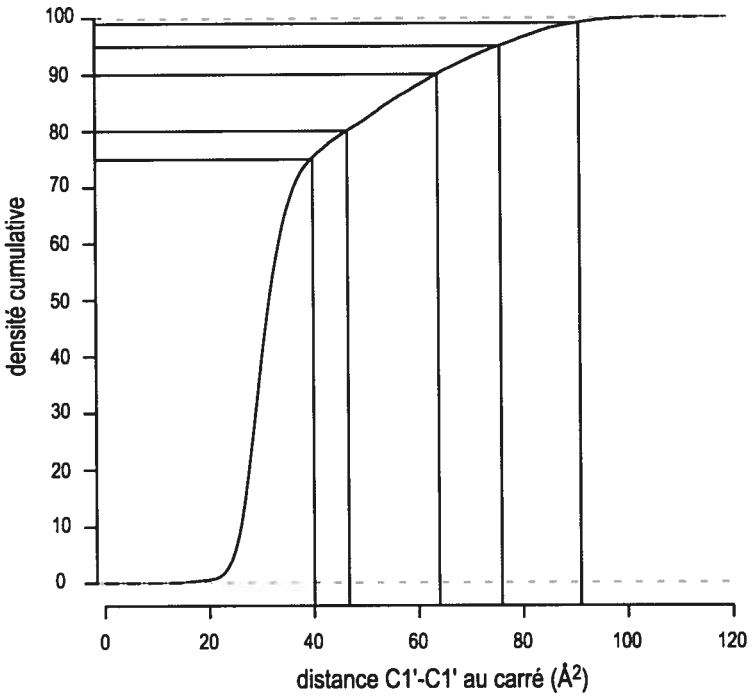


Figure 15 : Estimation de la courbe de densité empirique cumulative pour la mesure de la distance entre les atomes C1' de deux bases azotées adjacentes. 112 564 paires de bases adjacentes ont été mesurées. La distance est mesurée au carré pour des raisons d'efficacité calculatoire. Le seuil minimal est clairement de 20Å². Différents seuils maximaux sont acceptables dépendamment du degré de certitude recherché lors de la validation de la contrainte de fermeture. Par exemple, un seuil fixé à 65Å² assure qu'un peu plus de 90% des relations d'adjacence sont représentées. L'estimation de la courbe de densité empirique a été réalisée à l'aide du logiciel de calcul statistique *R* [43].

Cependant, le seuil maximal peut être abaissé tout en conservant une bonne couverture des mesures rapportées, comme l'indique la figure 15. Ainsi, la contrainte de fermeture devient plus stricte, mais elle s'expose au rejet de bases adjacentes positionnées aux limites de l'acceptable. Il existe donc un équilibre où la contrainte de fermeture témoignerait d'un taux de rejet suffisant pour avoir un impact sur l'efficacité globale de la recherche tout en minimisant les rejets de bonnes conformations, c'est-à-dire les faux négatifs. Pour trouver cet équilibre, nous avons mis au point une petite expérimentation de modélisation qui applique cette contrainte de fermeture à divers degré de certitude. Une simple boucle terminale à cinq nucléotides a été modélisée, puisqu'elle nécessite l'application de la sous-méthode de fermeture pour fermer la boucle à l'endroit choisi par l'ordre de construction spécifié. Les divers degrés de certitude utilisés dans la modélisation réfèrent directement à la couverture de la distribution des mesures de référence rapportée à la figure 15. Le tableau 2 détaille les résultats de ces modélisations par rapport à une modélisation de référence où aucune

contrainte de fermeture n'a été utilisée. Les critères de comparaison sont l'accélération de la complétion de la recherche, la perte de solutions et la diminution à la fois des appels de méthode de construction du ribose et des échecs de la contrainte de qualité du ribose construit. En général, peu importe le degré de certitude, l'utilisation de la contrainte de fermeture accélère grandement la recherche et réduit du même coup les échecs de construction de ribose. Cependant, le prix à payer est le rejet de nombreux modèles autrement valides. Lorsque la vitesse de recherche est de première importance, par exemple lorsqu'un seul modèle valide est recherché sans égard à la complétion de la recherche, la réduction du degré de certitude reste très bénéfique. Cependant, dans un contexte d'exploration complète de l'espace conformationnel, le degré de certitude doit rester élever : à 99% le rejet de solutions est acceptable tout en conservant une accélération marquée de la recherche.

degré de certitude (%)	temps de recherche (sec)	accélération recherche (%)	solutions	solutions rejetées (%)	riboses construits	riboses construits (%)	échec riboses	échec riboses (%)
—	1418	—	231	—	1 042 846	—	531 577	51.1
100	62	95.6	217	6.1	99 341	9.5	36 385	36.6
99	45	96.8	194	16.0	92 731	8.9	32 385	35.5
95	29	98.0	135	41.6	84 586	8.1	28 663	33.9
90	26	98.2	97	58.0	81 627	7.8	27 127	33.2
80	17	98.8	49	78.8	77 947	7.5	25 229	32.4
75	16	98.9	40	82.7	76 836	7.4	24 652	32.1

Tableau 2 : Résultats de la modélisation d'une boucle terminale avec utilisation de la contrainte de fermeture à divers degrés de certitude. Tous les résultats sont relatifs à ceux de la première ligne qui représente la modélisation sans l'utilisation de la contrainte de fermeture, à l'exception de la proportion d'échecs de construction de ribose qui réfère plutôt à la quantité totale de riboses construits. La ligne ombrée représente le meilleur choix de degré de certitude lorsque la recherche complète est considérée dans la modélisation.

2.4 Comparaison des engins de modélisation

Cette section est consacrée à l'analyse comparative des performances du nouvel engin de modélisation au cœur du système *MC-Sym* qui a été décrit dans les sections précédentes. Nous avons réalisé chaque expérimentation de modélisation d'un côté par l'engin original de modélisation et de l'autre par le nouvel engin. Tout d'abord, la première expérimentation compare la cohérence locale de la méthode de construction utilisée par chacun des engins.

Par cohérence locale, nous entendons l'habilité à modéliser deux nucléotides adjacents par application d'une relation d'adjacence. Les expérimentations suivantes sont basées sur deux modélisations complètes et veulent comparer les performances globales de modélisation des deux engins. La première modélise la structure connue de la boucle anticodon de $ARNt^{Phe}$ (levure). Les modèles seront comparés à la structure cristalline connue. La seconde modélise un motif présentant deux tiges reliées par une boucle interne.

2.4.1 Cohérence locale de la méthode de construction

Puisque l'application d'une relation d'adjacence lors de la construction implique la vérification d'une contrainte de distance de fermeture d'adjacence, il est intéressant de mesurer l'habilité démontrée par l'engin de modélisation à respecter cette contrainte dans ce cas particulier. En effet, puisque la relation d'appariement provient directement de la base de relations, c'est-à-dire d'un modèle de référence, la méthode de construction devrait être en mesure de relier correctement le squelette sur les bases azotées positionnées par cette relation. C'est cette cohérence locale de la construction qui est évaluée par cette expérimentation qui consiste à modéliser la structure de deux nucléotides adjacents, en utilisant toutes les relations d'adjacence de la base de relations, et à mesurer la distance de fermeture d'adjacence pour chaque construction. Pour l'engin original, le positionnement de deux bases adjacentes implique directement les nucléotides entiers, la distance de fermeture mesure le lien implicite O3'-P entre les deux nucléotides. Cette expérimentation pour l'engin original a déjà été présentée au début de la section 2.3 (voir figure 9). Pour le nouvel engin, le positionnement de deux bases adjacentes implique le positionnement d'un phosphate et la construction des deux riboses associés aux bases et se reliant à ce phosphate. Ici, la distance de fermeture d'adjacence est vérifiée par la qualité de construction de ces riboses, i.e. par la mesure des liens implicites C5'-O5' et C3'-O3'. Pour se comparer à la fermeture d'adjacence de l'engin original, la moyenne de ces deux mesures est considérée. La figure 16 montre l'histogramme de la distribution des distances de fermeture d'adjacence pour les deux engins. Le nouvel engin est clairement plus cohérent que l'engin original : 91.8% des relations d'adjacence sont construites dans un intervalle de distances de fermeture entre 1.0Å et 2.5Å, comparativement à seulement 10.4% dans le cas de l'engin original. Pour une distance s'étirant jusqu'à 3.0Å, cette proportion s'élève à 98.5% pour le nouvel engin comparativement à 16.5% pour l'engin original. Donc, cette expérimentation affirme

clairement la cohérence locale de la méthode de construction utilisée par le nouvel engin tout en soulignant l'incohérence de celle de l'engin original.

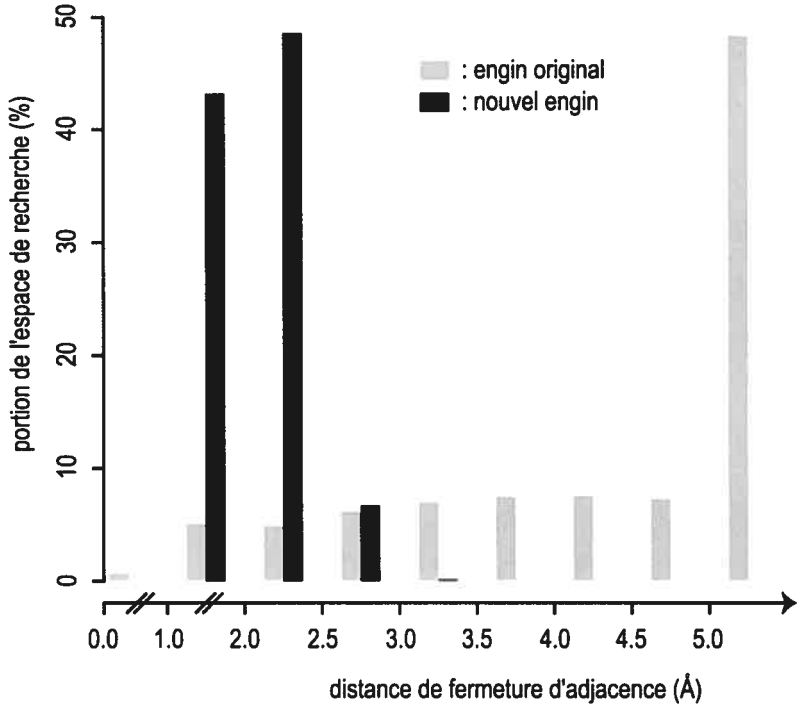


Figure 16 : Comparaison des distributions des distances de fermeture d'adjacence entre l'engin original de modélisation et le nouvel engin. La construction d'un nucléotide adjacent a été tentée pour toutes relations d'appariement adjacentes et pour toutes conformations de nucléotides (dans le cas de l'engin original), et la distance de fermeture d'adjacence a été mesurée pour chaque construction à l'aide des deux engins.

2.4.2 Modélisation de la boucle anticodon de ARNt^{Phe}

Pour comparer les performances globales des deux engins de modélisation, nous avons modélisé la structure de la boucle anticodon de ARNt^{Phe} (levure). Cette structure est bien connue : une tige se termine par une boucle de sept nucléotides dont deux sont empilés du côté 5' (le début de la chaîne polynucléotidique) et cinq du côté 3' (la fin de la chaîne). La figure 17 montre la structure secondaire de l'anticodon et l'information structurale donnée en entrée aux deux engins de modélisation. Évidemment, les spécifications des conformations des nucléotides ne sont utilisées que par l'engin original. Ainsi, le nouvel engin se démarque déjà de l'engin original en éliminant de l'exploration le sous-espace des conformations dont la taille combinatoire est de l'ordre de 10^{12} (16 conformations en 5 exemplaires et une en 10 exemplaires, donc $5^{16} \times 10 = 1.53 \times 10^{12}$). Cependant, il faut remarquer que le nouvel engin

parcourt les ensembles de relations d'adjacence inutilisés par la construction lors de l'application des sous-méthodes de fermeture, ainsi leur taille a un impact combinatoire sur la recherche, ce qui n'est pas le cas pour l'engin original. Nous avons fixé l'échantillonnage de chacun de ces ensembles de relations à 25%, tel que justifié à la sous-section 2.3.2.1.

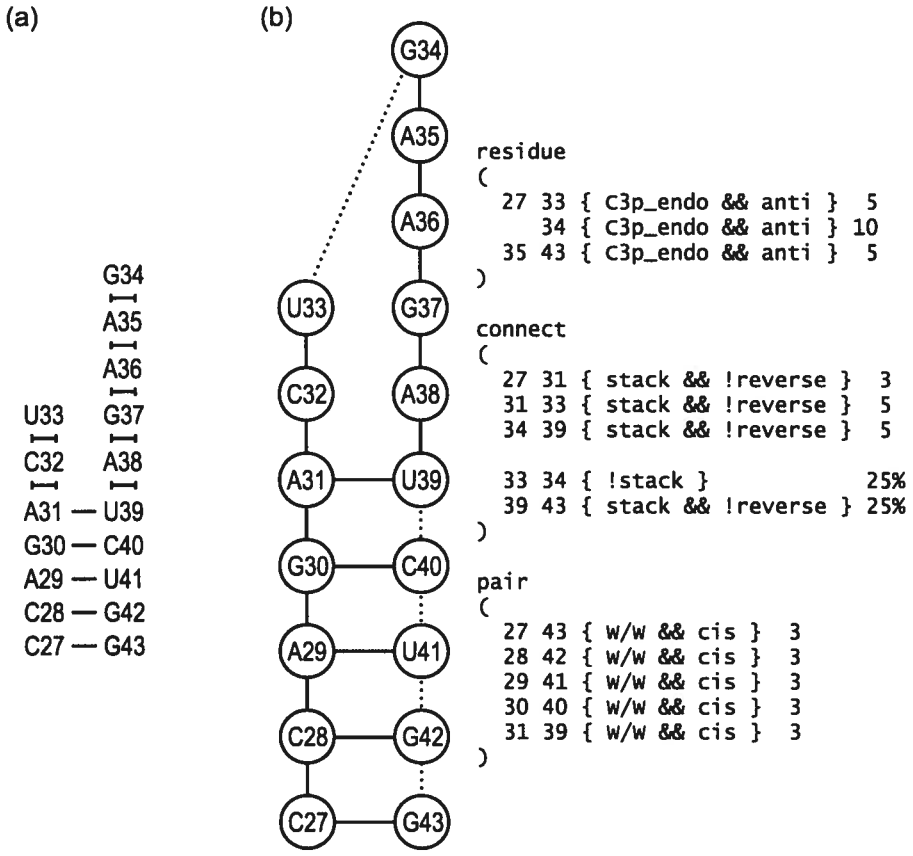


Figure 17 : Information structurale pour l'anticodon de ARNt^{Phe} donnée en entrée aux engins de modélisation. (a) Structure secondaire de l'anticodon selon la nomenclature Leontis & Westhof [30]. La numérotation suit celle de la structure secondaire de l'ARN de transfert au complet. (b) Graphe de relation annoté pour la modélisation. L'arbre de recouvrement spécifiant l'ordre de construction est représenté par des arêtes en trait plein, les arêtes en trait pointillé sont donc traitées par la sous-méthode de fermeture du nouvel engin. L'annotation des résidus et des relations est présentée selon la syntaxe d'entrée de *MC-Sym*.

L'information structurale présentée à la figure 17 a été donnée en entrée aux deux engins de modélisation. Le seuil de la contrainte de collision est fixé à 1Å et celui de la distance de fermeture d'adjacence à 4Å. L'algorithme de recherche utilisé dans les deux cas est le retour arrière classique, donc l'exploration de l'espace conformationnel est complète. Chaque modélisation utilise un filtre de similitude de 0.5Å sur les modèles trouvés. Ce filtre rejette tout modèle trouvé dont la déviation RMS avec un modèle déjà trouvé est inférieure à 0.5Å.

Les deux modélisations ont été réalisées sur un processeur *Intel Xeon* à 2.8 GHz. La recherche effectuée par le nouvel engin s'est complétée après 15 minutes et a trouvé 857 modèles différents. De son côté, la recherche effectuée par l'engin original n'a pas pu se compléter après 13 jours entiers d'exécution, et a été interrompue. Durant ces 13 jours, 701 modèles différents ont été trouvés par l'engin original. La figure 18a montre la progression de la découverte de modèles par les deux engins lors des 15 premières minutes de recherche, c'est-à-dire jusqu'à ce que le nouvel engin complète sa recherche. Pendant ces 15 minutes, l'engin original n'arrive à trouver que 144 modèles différents. Ainsi, du point de vue de l'efficacité de la recherche, le nouvel engin se distingue nettement de l'engin original, probablement dû au coût combinatoire associé à l'exploration du sous-espace des conformations. De plus, il est intéressant de considérer la représentativité des modèles trouvés, c'est-à-dire la précision avec laquelle la structure a été modélisée par rapport à un modèle de référence. Pour ce faire, la déviation RMS de chaque modèle trouvé par chacun des deux engins a été calculée avec le modèle de référence pour $ARNt^{Phe}$, soit celui construit par cristallographie et déposé dans *PDB* (code 1EVV). La figure 18b montre la distribution empirique des mesures obtenues. Puisque l'aire sous la courbe de densité entre un intervalle de déviation RMS donne la proportion des modèles dont leur déviation se trouve dans cet intervalle, la forme de la courbe de l'engin original relativement à celle du nouvel engin témoigne qu'une plus grande proportion des modèles trouvés par le nouvel engin sont plus précis que ceux trouvés par l'engin original. Bref, le nouvel engin de modélisation surpasse l'engin original sur tous les plans dans cette expérimentation de modélisation : il trouve plus de modèles qui sont généralement plus près du modèle cristallographique, et ce en complétant sa recherche beaucoup plus rapidement.

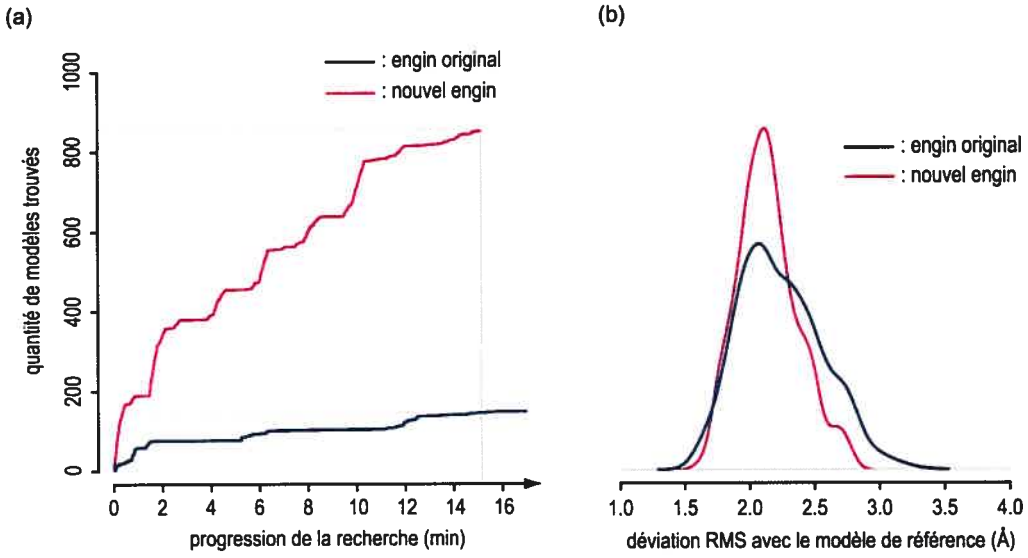


Figure 18 : Comparaison des performances de modélisation de l'anticodon de ARNt^{Phe} par l'engin original et le nouvel engin. (a) Progression de la découverte de modèles. Seulement les 15 minutes nécessaires à la complétion de la recherche et à la découverte de 857 modèles différents par le nouvel engin sont montrées. Durant cette période, l'engin original n'arrive à trouver que 144 modèles différents. (b) Estimation des courbes de densité empirique pour la déviation RMS des modèles trouvés avec le modèle de référence de ARNt^{Phe} (code *PDB* : 1EVV) par les deux engins. Ces courbes ont été évaluées à l'aide du logiciel de calcul statistique *R* [43].

2.4.3 Modélisation d'un motif de boucle interne

Dans l'expérimentation de modélisation présentée ci-haut, la structure modélisée était bien connue. En particulier, les empilements de bases dans la boucle terminale permettent une modélisation fine de la structure de la boucle. De plus, l'existence d'un modèle cristallographique permet la validation des modèles trouvés. Cependant, il est intéressant de comparer les performances des deux engins de modélisation sur une structure moins connue. Pour ce faire, nous avons modélisé la structure d'un motif hypothétique de boucle interne à l'aide des deux engins de modélisations. La figure 19a montre la structure secondaire de ce motif. Celui-ci est constitué de la réunion de deux tiges par une boucle interne de quatre nucléotides d'un côté et de deux de l'autre. Aussi, une des tiges contient un appariement G-U de type *wobble* (*Ww/Ws cis*) à l'extrémité en contact avec la boucle. La figure 19b illustre l'information structurale d'entrée aux engins de modélisation. Encore une fois, les spécifications de conformations de résidus ne concernent que l'engin original, tout comme les

spécifications des relations d'adjacence inutilisées par l'ordre de construction ne concernent que le nouvel engin. Ces relations de fermeture sont échantillonnées à 25%. De façon générale, un plus grand échantillonnage a été accordé aux éléments de la tige contenant l'appariement *wobble* (Ga8 avec Ub3) puisque cet appariement est moins bien représenté dans la base de relation que l'appariement Watson-Crick canonique.

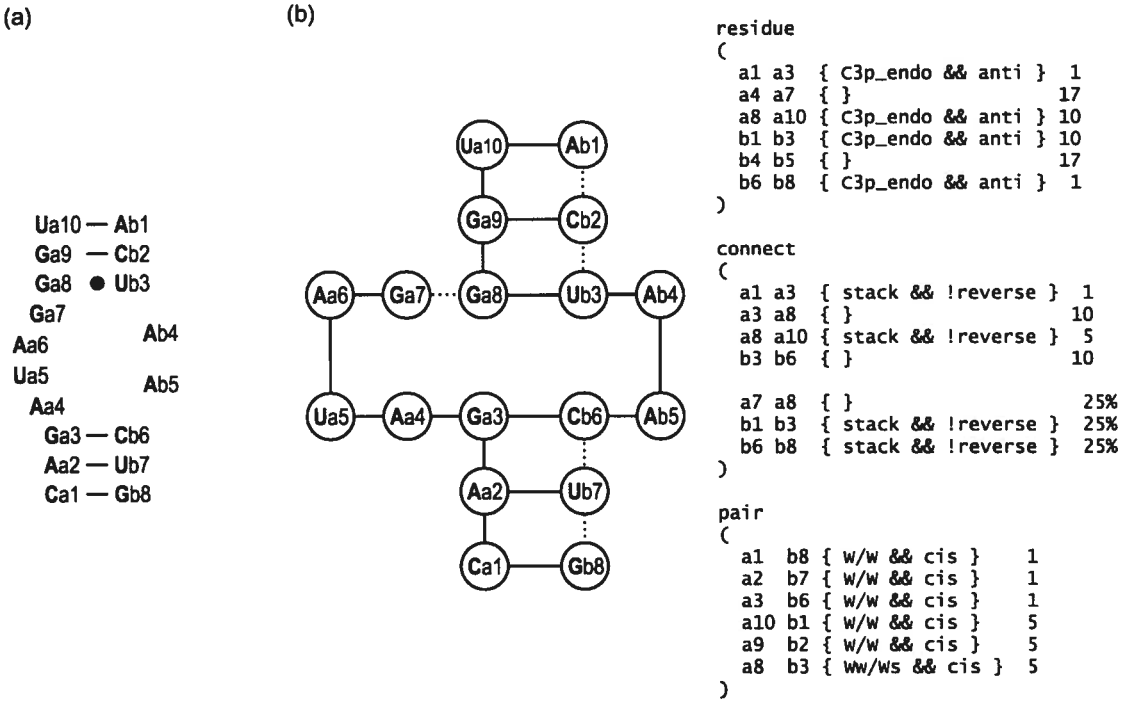


Figure 19 : Information structurale du motif de boucle interne donnée en entrée. (a) Structure secondaire de la boucle interne selon la nomenclature Leontis & Westhof [30]. Les deux chaînes sont numérotées 'a' et 'b'. (b) Graphe de relation annoté pour la modélisation. L'arbre de recouvrement spécifiant l'ordre de construction est représenté par des arêtes en trait plein, les arêtes en trait pointillé sont donc traitées par la sous-méthode de fermeture du nouvel engin. L'annotation des résidus et des relations est présentée selon la syntaxe d'entrée de *MC-Sym*. Ainsi, selon cette syntaxe, une spécification de relations (ou de conformations) par une requête vide, i.e. { }, considère toutes les relations (ou conformations) de la base de données.

L'information structurale présentée à la figure 19 a été donnée en entrée aux deux engins de modélisation. Le seuil de la contrainte de collision est fixé à 1Å et celui de la distance de fermeture d'adjacence à 4Å. Pour le nouvel engin de recherche, la contrainte de fermeture est utilisée, son degré de certitude est ajusté à 90% (voir la sous-section 2.3.3.3). L'algorithme de recherche utilisé dans les deux cas est le retour arrière classique, donc l'exploration de l'espace conformationnel est complète. Chaque modélisation utilise un filtre de similitude de 0.5Å sur les modèles trouvés. Les deux modélisations ont été réalisées sur un processeur *Intel*

Pentium 4 à 2.0 GHz. La recherche effectuée par le nouvel engin s'est complétée après 2.3 minutes et a trouvé 85 modèles différents. De son côté, la recherche effectuée par l'engin original s'est complétée après un total de 8 jours et 18 heures sans trouver aucun modèle. Puisque la spécification et l'échantillonnage des relations sont identiques dans les deux cas, il est justifié de croire que l'échantillonnage du sous-espace des conformations de l'engin original n'est pas suffisamment élevé pour permettre à la recherche de valider ses constructions. Or, déjà avec cet échantillonnage, la recherche a nécessité près de 9 jours. Puisque la taille du sous-espace des conformations a un impact exponentiel sur celle de l'espace conformationnel entier, il nous semble futile d'accroître davantage l'échantillonnage des conformations; le nouvel engin a déjà prouvé son efficacité supérieure. La figure 20 illustre 4 des 85 modèles dont les conformations tridimensionnelles sont très différentes les unes des autres, de sorte à souligner la grande diversité des modèles trouvés. Cette diversité se quantifie par la mesure de la déviation RMS maximale entre n'importe quelle paire de modèles de l'ensemble solution; elle est de 12.23Å.

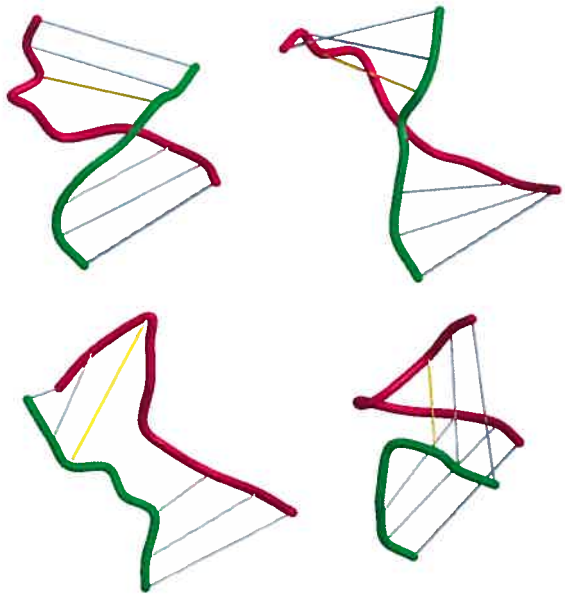


Figure 20 : Quelques modèles de boucle interne trouvés par le nouvel engin de modélisation en représentation simplifiée. La représentation simplifiée montre seulement la chaîne du squelette en passant par les groupements phosphates, la chaîne 'a' en rouge qui contient le côté à quatre nucléotides de la boucle et la chaîne 'b' en vert qui contient le côté à deux nucléotides. Les appariements de bases dans les tiges sont représentés par des traits gris. Le trait jaune montre l'appariement *wobble*. Images générées par *Molscript* [36] [44] et *Raster3D* [37] [45].

2.5 Modélisation complète de ARNt^{Phe}

Nous avons consacré la section précédente à l'étude comparative des performances du nouvel engin de modélisation par rapport à l'engin original au sein du système *MC-Sym* en réalisant de petites expérimentations de modélisation. Dans cette section, nous présentons la modélisation de la structure complète d'un ARN de transfert (ARNt^{Phe} de la levure) à l'aide du nouvel engin de modélisation.

La figure 4 montre la structure secondaire de ARNt^{Phe} qui représente l'information structurale utilisée dans cette modélisation. Cependant, dans cette expérimentation, la modélisation est fragmentée. La modélisation par fragment est un mode de modélisation offert par *MC-Sym*. Il consiste à intégrer dans la construction un fragment complet de structure qui est considéré comme un résidu rigide. S'il existe plusieurs modèles pour le fragment à insérer, leur réunion forme l'ensemble de conformations du résidu représenté par le fragment, et cet ensemble s'insère dans l'espace conformationnel de la modélisation. Ainsi, pour un problème de modélisation particulier, il est possible de le séparer en sous-problèmes plus simples et de les recombiner, à la manière d'un algorithme diviser pour régner (voir [31] pour une définition d'un algorithme diviser pour régner). Le résultat final est le même que lorsque toute la structure est considérée d'un seul coup, le gain en performance vient du fait que la région de l'espace conformationnel couverte par un fragment subdivisé n'est explorée qu'une seule fois. En effet, dans le contexte d'une modélisation sans subdivision, cette région de l'espace conformationnel sera explorée autant de fois qu'il y a de solution partielle à l'exploration des régions précédentes dans l'ordre de construction. Au chapitre 4, nous reviendrons sur le formalise de cette optimisation.

Évidemment, dans le processus de modélisation par fragments, puisque le fragment représente un résidu particulier, une méthode de construction particulière s'applique pour positionner le fragment. Le fragment est positionné sur une base azotée de référence, i.e. déjà positionnée, par relation d'appariement avec une base sélectionnée du fragment. Ainsi, tout le fragment est positionné de façon rigide selon une relation d'appariement dans le référentiel d'une base déjà positionnée, de la même façon que les méthodes de construction usuelles positionnent une seule base. En fait, la méthode de construction par fragment devient une généralisation des méthodes de construction d'une relation d'adjacence ou de non-adjacence. Dépendamment du type de la relation choisie pour placer le fragment, un des deux concepts

de méthode de construction s'applique de la même façon. Cependant, il faut considérer que plusieurs bases sont positionnées, donc le processus de vérification d'application des sous-méthodes de fermeture et de construction de ribose doit être mis en œuvre à plusieurs reprises. La figure 21 illustre les différentes étapes de fragmentation dans la modélisation de $ARNt^{Phe}$. Ici, le processus de modélisation commence par un petit fragment de la structure – la tige D avec la paire U8-A14 – et construit le reste de la structure en trois étapes sous-inclusives.

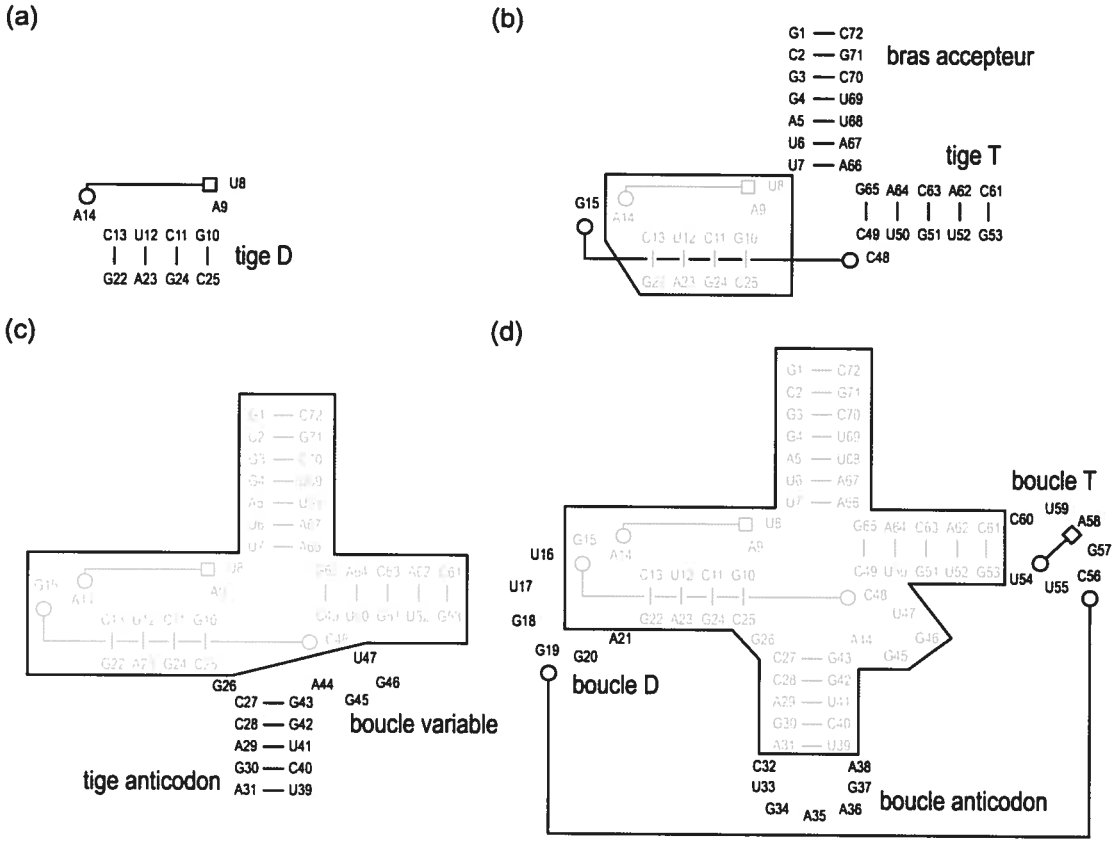


Figure 21 : Étapes de modélisation de $ARNt^{Phe}$ par fragments. (a) Fragment initial constitué de la tige D et de la paire U8-A14. (b) Fragment constitué du fragment précédent (en gris et encadré) et de l'ajout de la tige T du bras accepteur et de la paire G15-C48. (c) Fragment constitué du fragment précédent (en gris et encadré) et de l'ajout de la tige anticodon et de la boucle variable. (d) Étape finale de modélisation. Le reste de la structure est ajoutée au fragment précédent (en gris et encadré), soient les boucles D, T et anticodon.

Les contraintes sont ajustées de façon identique pour chaque étape de modélisation : seuil de 1Å pour les collisions et seuil de 3.5Å pour la distance de fermeture d'adjacence. La contrainte de fermeture est utilisée avec un degré de similitude de 95%. Un filtre de similitude de 1Å est aussi utilisé à toutes les étapes. Chaque étape de modélisation s'est

exécutée sur un processeur *Intel Pentium 4* à 2.0 GHz. La première étape de modélisation a utilisé l'algorithme de recherche par retour arrière classique tel que décrit à la sous-section 2.2.3, et la recherche exhaustive s'est complétée. Par contre, les étapes ultérieures n'utilisent pas cet algorithme exhaustif, mais plutôt une heuristique probabiliste proposée à la sous-section 2.2.3 et qui sera étudiée en détails au chapitre 4. Ainsi, les recherches à ces étapes ne sont pas complètes, elles ont été interrompues à un certain moment. Le tableau 3 détaille chaque étape de modélisation en montrant la quantité de modèles trouvés pour le fragment, le temps de recherche nécessaire et la diversité de l'ensemble des modèles trouvés, telle que mesurée par la déviation RMS maximale entre n'importe quelle paire de modèles. Au total, un peu plus de 6 jours de recherche ont été nécessaires à la construction de 11 modèles complets pour $ARNt^{Phe}$. L'étape la plus difficile a été la dernière où les boucles D et T sont reliées par l'appariement G19-C56, occupant la recherche pendant près de 5 jours.

étape	temps écoulé (heures)	modèles trouvés	diversité (Å)
1 : tige D	2	63	6.48
2 : tige T, bras accepteur	3	603	15.03
3 : tige anticodon, boucle variable	24	882	21.00
4 : boucles D, T et anticodon	118	11	18.03

Tableau 3 : Résultats des quatre étapes de modélisation de $ARNt^{Phe}$. Globalement, 147 heures (un peu plus de 6 jours) ont été nécessaires à la construction de 11 modèles complets. La diversité des modèles trouvés est mesurée par la déviation RMS maximale entre n'importe quelle paire de modèles.

Puisqu'il existe un modèle de référence pour la structure de $ARNt^{Phe}$ déterminé par cristallographie (code *PDB* 1EVV), il est intéressant de mesurer la déviation RMS de nos 11 modèles avec celui-ci. Malheureusement, ces mesures sont assez grandes, variant de 7.5Å à 13.7Å. Cependant, il faut remarquer que la variété de l'espace conformationnel des boucles terminales a un fort impact sur la déviation globale. Ainsi, en mesurant la déviation RMS obtenu seulement en considérant les quatre tiges – D, T, anticodon et accepteur –, l'écart diminue : de 5.6Å à 10.5Å. Cette mesure est théoriquement moins précise que la déviation globale, mais elle apporte une clarification de la similitude de la structure en tenant uniquement compte de l'orientation relative des éléments de structure secondaire et en retirant le bruit causé par la diversité artificielle des boucles terminales. La figure 22 illustre la comparaison de notre meilleur modèle selon la déviation des tiges avec le modèle de référence. À l'œil, la divergence semble être causée par la différence de la liaison des boucles D et T par l'appariement G19-C56. Cependant, cette différence n'a pas empêché notre

modèle de se replier dans la forme caractéristique en 'L' des ARN de transfert, puisque l'appariement U8-A14, nécessaire et suffisant à ce repliement [17], est quant à lui très bien représenté. En 1993, Major et ses collaborateurs ont modélisé cette même structure en utilisant le système *MC-Sym* de l'époque [17] et ont obtenu un modèle à 3.1Å, après un raffinement par minimisation d'énergie. À première vue, les résultats de notre modélisation font piètre figure face à cette précision. Cependant, mis à part le fait que nos modèles n'ont pas subi de processus de minimisation d'énergie, il est difficile de comparer notre nouvel engin de modélisation à celui utilisé en 1993 par Major et ses collaborateurs à cause de la différence entre la composition des bases de données de *MC-Sym* en 1993 et en 2004, qui ont grandement évolué.

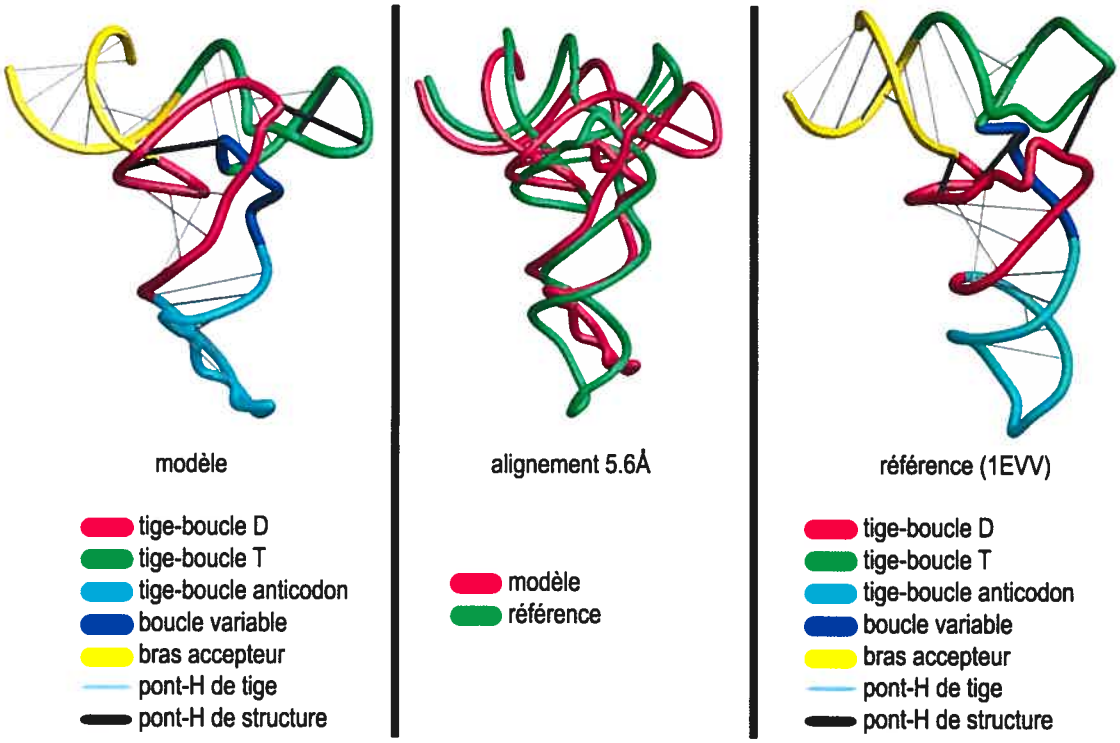


Figure 22 : Comparaison d'un modèle de ARNt^{Phe} construit par le nouvel engin de modélisation avec le modèle de référence obtenu par cristallographie. Les modèles sont illustrés en représentation simplifiée où seule la chaîne du squelette est visible. Le modèle et la référence sont illustrés séparément, respectivement à gauche et à droite de la figure, avec leurs éléments de structure secondaire identifiés par des couleurs différentes. De plus, les ponts-H des appariements de bases sont illustrés par des traits gris pour ceux des tiges et par des traits noirs pour ceux reliant les éléments de structure secondaire, c'est-à-dire pour les paires U8-A14, G15-C48 et G19-C56. Au centre, le modèle construit par l'engin de modélisation est superposé au modèle de référence. Images générées par *Molscript* [36] [44] et *Raster3D* [37] [45].

2.6 Conclusion

Nous avons redéfini l'engin de modélisation au cœur du système *MC-Sym* tant du point de vue de la définition de l'espace conformationnel que des méthodes de construction utilisées pour positionner les résidus dans l'espace tridimensionnel. Dans ce nouvel engin, les bases azotées et les phosphates sont positionnés indépendamment par relations d'appariement, les riboses s'ajoutent ensuite par construction artificielle. Déjà, cette approche améliore les performances combinatoires théoriques de l'engin de modélisation en supprimant le sous-espace conformationnel des conformations statiques de nucléotides. De plus, nous avons comparé la cohérence locale des méthodes de construction des deux engins en mesurant la distance de fermeture d'adjacence entre deux nucléotides adjacents positionnés par la méthode de construction. Dans cette expérimentation, le nouvel engin positionne la majorité des nucléotides adjacents avec une distance d'adjacence variant entre 1.0Å et 2.5Å, comparativement à plus de 4.5Å pour l'engin original.

Pour comparer les performances de modélisation entre l'engin original et le nouvel engin, nous avons réalisé deux expérimentations : la modélisation de l'anticodon de $ARNt^{Phe}$ et la modélisation d'un motif de boucle interne. La modélisation de l'anticodon par le nouvel engin surclasse celle réalisée par l'engin original tant au niveau de l'efficacité que de la précision. Côté efficacité, le nouvel engin complète la recherche en 15 minutes et trouve 857 modèles différents d'au moins 0.5 Å de déviation RMS, alors que l'engin original n'en trouve que 701 en 13 jours de recherche incomplète. Côté précision, les courbes de densité des mesures de déviation RMS avec le modèle cristallographié (figure 18b) montrent qu'une majorité des modèles construits par le nouvel engin sont plus près du modèle de référence que ceux construits par l'engin original. Pour sa part, l'expérimentation de la modélisation d'un motif de boucle interne démontre clairement l'efficacité supérieure du nouvel engin : ce dernier trouve 85 modèles et complète sa recherche en 2.3 minutes pendant que l'engin original n'en trouve aucun après plus de 8 jours de recherche complète. Bref, notre nouvel engin de modélisation démontre clairement des performances de modélisation supérieures à l'engin original. Dans ces deux expérimentations, les données structurales d'entrée étaient identiques pour les deux engins, à l'exception du sous-espace des conformations qui est unique à l'engin original. Donc, il semble que l'exploration du sous-espace des conformations par l'engin original réussit à elle seule à ralentir considérablement la

recherche, confirmant notre choix de l'éradiquer dans notre nouvelle définition d'engin de modélisation.

Finalement, nous avons modélisé la structure complète d'un ARN de transfert, $ARNt^{Phe}$, avec notre nouvel engin de modélisation. Pour ce faire, la structure a été brisée en fragments locaux plus simple à modéliser, la structure finale s'obtenant par recombinaisons des fragments à la manière d'un algorithme diviser pour régner. De cette façon, le nouvel engin a trouvé 11 modèles complets pour $ARNt^{Phe}$ en un peu plus de 6 jours de recherche probabiliste. La déviation RMS de ces modèles avec le modèle cristallographié (1EVV) varie entre 7.5Å et 13.7Å. L'inspection visuelle du meilleur modèle aligné sur le modèle de référence montre que la forme caractéristique des ARN de transfert en 'L' est conservée et que la divergence de structure provient principalement de la liaison des boucle D et T par l'appariement G19-C56.

Notre nouvelle approche de modélisation dans le contexte du système *MC-Sym* est donc justifiée. Certains aspects de ce nouvel engin de modélisation restent cependant ouverts. C'est le cas de la méthode de construction artificielle du ribose. C'est la méthode de construction la plus complexe utilisée par l'engin de modélisation puisqu'elle représente une construction artificielle atome par atome plutôt qu'un positionnement rigide et discret d'un résidu par transformation linéaire. Par conséquent, cette méthode de construction représente probablement le goulot d'étranglement des performances de modélisation, d'autant plus qu'elle est présente à chaque positionnement d'une base azotée. Dans le chapitre suivant, nous analyserons en détails cette méthode de construction en particulier. Par ailleurs, la tendance à vouloir construire le ribose hâtivement lors de la recherche implique qu'un même ribose va être reconstruit plusieurs fois lors de la recherche, chaque fois qu'un des trois résidus qu'il interconnecte sera repositionné. Dans une approche différente, nous proposons d'attendre d'avoir construit le modèle complet en ayant positionné ses bases et ses phosphates avant d'y construire les riboses. Le problème avec cette approche est que si un seul des riboses ne parvient pas à se construire de façon convenable, il est difficile de décider quel devrait être le comportement de l'engin de modélisation : tentative de correction locale ou élimination du modèle et recommencement de la recherche? Cependant, une observation des décomptes de collisions atomiques lors de la modélisation de l'anticodon montre que, proportionnellement, les collisions surviennent en majorité au groupement phosphate et de façon négligeable au ribose. Ainsi, la validité des riboses construits en fin de recherche est de

bon augure, puisqu'il semble que les collisions avec les groupements phosphates se chargent de ménager assez de place pour la construction ultérieure du ribose. Cependant, la construction hâtive du ribose sert aussi de contraintes de fermeture d'adjacence. Retirer la construction du ribose pendant la recherche, c'est retirer cette contrainte vitale à la validation du modèle. La contrainte de fermeture développée sur la distance entre deux bases adjacentes pourrait jouer le rôle de contrainte d'adjacence, cependant il faudrait prouver que sa validation permet d'accepter ou de rejeter univoquement une construction particulière, ce qui n'est pas le cas présentement. En effet, les résultats de paramétrage de cette contrainte, présentés au tableau 2, démontrent que son utilisation entraîne la perte de certaines solutions. Par conséquent, avant de retirer la construction des riboses de la recherche, il faut développer une contrainte de fermeture d'adjacence juste et précise.

Aussi, la vérification des contraintes de collision par la mesure d'un seuil de distance interatomique est plutôt naïve et génère un grand nombre de calculs de distance point à point. Nous avons proposé une amélioration dans ce nouvel engin de modélisation en considérant tout le groupement phosphate comme un seul atome, réduisant ainsi considérablement la quantité total de calculs de distance interatomique. Par ailleurs, une autre optimisation de la vérification des contraintes de collision pourrait être envisagée. Il s'agit de placer une couche de validation préalable au calcul complet de distance point à point. Cette couche utilise des boîtes englobantes (*bounding boxes*) placées autour des résidus dans le référentiel global. Si les boîtes englobantes de deux résidus ne se chevauchent pas, alors la contrainte de collision impliquant ces deux résidus est validée avant même d'avoir à mesurer toutes les distances interatomiques. Les coordonnées des coins d'une boîte englobante pour un résidu en particulier sont stockées dans ce résidu pour permettre de déplacer la boîte de façon rigide avec le résidu. L'optimisation serait d'autant plus marquée lors de la modélisation par fragment, où la boîte engloberait tout le fragment. Par contre, s'il y a chevauchement, il faut quand même passer par le calcul des distances interatomiques pour vérifier correctement la contrainte de collision.

Chapitre 3

Construction artificielle d'un squelette d'ARN

3.1 Introduction

Dans un contexte de modélisation de la structure tridimensionnelle d'un ARN, certains éléments géométriques sont identifiables. Par exemple, les appariements et les empilements de bases azotées, adjacentes ou non dans la chaîne polynucléotidique, sont des éléments de géométrie réutilisables d'autant plus que les bases sont planaires; leur géométrie est fixe et connue. Dans ce contexte, le système de modélisation *MC-Sym* [16] [40] utilise une base de données de relations d'appariement entre bases azotées qui ont été extraites sur des modèles de référence. Les éléments de cette base de relations sont utilisés pour positionner une base azotée par rapport à une autre. Pour compléter le modèle en ajoutant le squelette (*backbone*), l'engin original de modélisation au cœur du système *MC-Sym* se replie sur une base de données contenant des modèles complets de nucléotides figés dans différentes conformations. Cependant, bien que les bases azotées soient géométriquement fixes, ce n'est pas le cas des riboses qui composent le squelette. Plusieurs angles de torsions libres permettent à un ribose d'adopter maintes conformations différentes. Plutôt que de tenter de discrétiser ce vaste espace conformationnel par une base de nucléotides, nous avons décrit au chapitre précédent un nouvel engin de modélisation dont l'une des caractéristiques majeures est la construction artificielle des riboses, atome par atome. Dans ce chapitre, nous focaliserons sur cette méthode de construction particulière. Cette méthode se sépare en deux procédures distinctes : la construction paramétrée du ribose et la détermination des paramètres optimaux de construction en contexte. D'abord, nous décrirons la méthode de construction du ribose atome par atome depuis la base azotée. Cette méthode est paramétrée selon la torsion du lien glycosylique et la pseudorotation du furanose. Nous quantifierons sa précision et sa représentativité. Ensuite, trois procédures différentes de détermination des paramètres optimaux de la construction seront détaillées et comparées : deux méthodes d'optimisation linéaire numérique et une méthode d'estimation directe. Par des analyses comparatives, nous

démontrerons que la méthode par estimation réussit à construire un ribose de qualité appréciable en seulement deux itérations de construction.

3.2 Quelques définitions

Avant de décrire la méthode de construction artificielle du ribose, certaines conventions et définitions doivent être posées. Ainsi, cette section met en place les définitions nécessaires à la formalisation de la méthode de construction. D'abord, rappelons qu'un modèle est représenté par une suite de triplets, chaque triplet encodant la position tridimensionnelle d'un atome. Donc, dans un modèle, un atome est un point tridimensionnel.

3.2.1 Angle de torsion

Pour décrire la conformation tridimensionnel de la molécule de ribose, il est pertinent de considérer les différents angles de torsion mesurables à chaque lien covalent. Dans un modèle, un angle de torsion se mesure sur une suite de quatre atomes, i.e. de quatre points, de la façon formalisée à la figure 23.

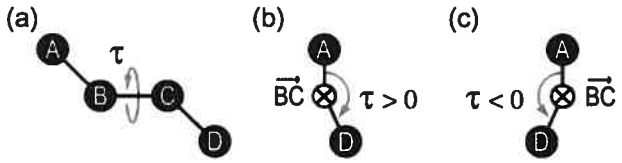
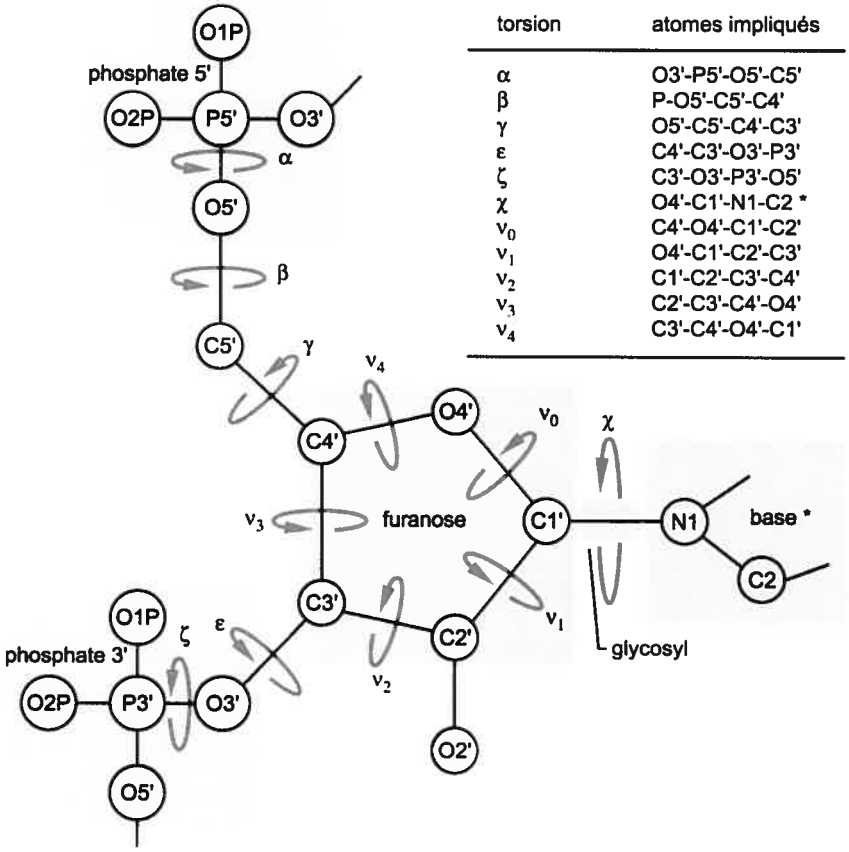


Figure 23 : Définition d'un angle de torsion. (a) L'angle de torsion τ est définit sur la suite des atomes ABCD et mesure l'angle entre le plan formé par les vecteurs $\vec{AB}$ et $\vec{BC}$ et celui formé par les vecteurs $\vec{BC}$ et $\vec{CD}$. Par convention, τ prend une valeur comprise entre -180° et 180°, le sens positif étant définit par le sens horaire de A vers D en regardant dans la direction du vecteur $\vec{BC}$. τ est donc positif en (b) et négatif en (c).

3.2.2 Nomenclature de la structure du ribose

En considérant le ribose complet, 13 angles de torsions sont mesurables. La figure 24 fixe la nomenclature de ces angles.



torsion	atomes impliqués
α	O3'-P5'-O5'-C5'
β	P-O5'-C5'-C4'
γ	O5'-C5'-C4'-C3'
ε	C4'-C3'-O3'-P3'
ζ	C3'-O3'-P3'-O5'
χ	O4'-C1'-N1-C2 *
ν_0	C4'-O4'-C1'-C2'
ν_1	O4'-C1'-C2'-C3'
ν_2	C1'-C2'-C3'-C4'
ν_3	C2'-C3'-C4'-O4'
ν_4	C3'-C4'-O4'-C1'

Figure 24 : Nomenclature des 13 angles de torsions mesurables dans le ribose complet [19]. Mis en évidence sont les groupements structuraux caractéristiques du ribose. La base présentée est une pyrimidine. Pour une purine, χ se mesure sur la suite O4'-C1'-N9-C4.

Puisque les atomes C1', C2', C3', C4' et O4' forment un cycle, les valeurs des cinq torsions mesurables sur ce cycle, soient ν_0 à ν_4 sont manifestement restreintes par la géométrie fixe des longueurs de lien covalent et angles de coude. Cette restriction ce quantifie par la relation suivante [20] :

$$\tan\rho = \frac{(\nu_4 + \nu_1) - (\nu_3 + \nu_0)}{2\nu_2(\sin 36° + \sin 72°)} \qquad (1)$$

En biologie moléculaire, un ribose s'identifie la mesure des angles ρ et χ, respectivement nommé pseudorotation et torsion du lien glycosyl. La figure 25 déclare les différents étiquettes attribuables à un intervalle de valeurs tant pour la pseudorotation que pour la torsion glycosylique. Une identification d'une conformation stéréochimique particulière pour

un modèle de ribose est appelée mode de *puckering* (froissement). Au total, 20 modes différents sont identifiables par différents intervalles de mesures pour ρ et χ.

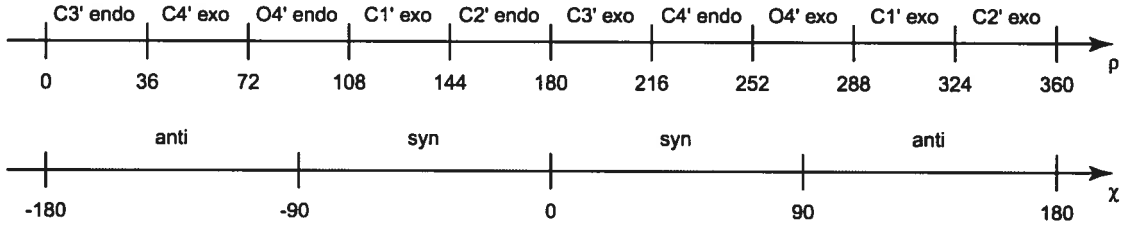


Figure 25 : Qualificatifs attribués à la pseudorotation ainsi qu'à la torsion du lien glycosyl [19]. La nomenclature de la pseudorotation réfère à l'atome du cycle qui pointe vers l'extérieur du plan formé par les autres atomes et à la direction relative à ce plan (endo ou exo). Celle du glycosyl réfère à l'orientation relative du ribose par rapport à la base azotée, face à face (syn) ou opposée (anti).

3.3 Construction artificielle du ribose

Dans cette section, nous présentons la description formelle et complète de notre méthode de construction artificielle d'un modèle de ribose isolé, c'est-à-dire hors du contexte de la modélisation. La méthode sera formellement présentée et évaluée selon des critères de précision, de représentativité et d'efficacité technique. La précision mesure l'aptitude à construire un ribose dont la géométrie respecte les mesures de référence. La représentativité vérifie que tout modèle de ribose construit par cette méthode ressemble à ceux des modèles de référence, et *vice-versa*. L'efficacité technique s'attardera à optimiser le temps de calcul requis par la construction.

3.3.1 Description de la méthode de construction

Dans l'ensemble de la structure tridimensionnelle du ribose, certains éléments géométriques sont ici considérés stéréochimiquement rigides. Ces éléments sont la longueur des liens covalents et l'angle entre deux liens covalents, i.e. l'angle de coude formé par une suite de trois atomes. Le tableau 4 établit la liste des mesures constantes dans le cadre de cette méthode de construction. Dans ce cadre de fixité, la mesure des différents angles de torsion catalogue directement les conformations adoptables par le ribose. Ainsi, une méthode de construction ayant comme paramètres χ, α, β, γ et ν_0 à ν_4 (voir figure 24) et comme

constantes l'ensemble des longueurs des liens covalents et des angles de coude définit complètement la position de chacun des atomes du ribose dans chacun des 20 modes de *puckering* différents, jusqu'aux groupements phosphates en direction 5' et 3'. Pour ce faire, chaque atome du ribose se positionne par rapport aux atomes précédemment positionnés et successivement liées entre eux. Conséquemment, dans ce positionnement successif, la flexibilité du ribose construit s'exprime par les angles de torsions précédemment mentionnés.

Liens covalents		Angles de coude	
Atomes	Longueur (Å)	Atomes	Angle (degrés)
C1' N	1.465	C2 N1 C1'	126.371
C1' C2'	1.529	N1 C1' C2'	112.028
C1' O4'	1.417	N1 C1' O4'	108.530
C2' O2'	1.414	C2' C1' O4'	107.424
C2' C3'	1.523	C1' C2' C3'	101.446
C3' O3'	1.431	C2' C3' C4'	102.211
C3' C4'	1.521	C3' C4' O4'	104.479
C4' C5'	1.510	C4' O4' C1'	109.700
C4' O4'	1.452	C1' C2' O2'	109.768
C5' O5'	1.440	C3' C2' O2'	112.956
O5' P5'	1.593	C2' C3' O3'	112.261
O3' P3'	1.593	C4' C3' O3'	111.820
		C3' C4' C5'	115.386
		O4' C4' C5'	109.184
		C4' C5' O5'	109.402
		C5' O5' P5'	120.934
		C3' O3' P3'	120.934

Tableau 4 : Éléments de géométrie rigides dans la construction du ribose. Ces mesures ont été prélevées sur un modèle de la structure du ribose déterminé par Parkinson [22] [42]. Les zones ombrées réfèrent à des éléments de la géométrie rigide qui ne sont pas exprimés explicitement dans la méthode de construction du ribose. Les mesures impliquant N1 ou C2 proviennent d'une base azotée associée au ribose de la famille des pyrimidines. Dans le cadre de cette méthode de construction, la famille des purines est considérée équivalente aux pyrimidines quant à la position de l'azote et du carbone impliqués dans le calcul de χ (voir figure 24). Le ribose construit est ici indépendant de la famille de la base azotée.

Plus précisément, Un atome se positionne dans le système de référence de l'atome précédemment positionné et lié à celui-ci. Dans ce système de référence, l'atome se place par une série de trois transformations linéaires qui expriment directement la géométrie de la position de l'atome, soient la distance du lien covalent avec l'atome précédent, l'angle de coude avec ce dernier et le deuxième précédent et enfin la torsion avec ces deux derniers et le troisième précédent (voir figure 26). Les deux premières transformations sont toujours

constantes car elles reflètent la géométrie rigide, alors que la troisième fait intervenir directement un paramètre de la construction : un angle de torsion. Le système de référence est aligné sur le lien en torsion, de sorte que la torsion par rapport à l'atome à positionner s'exprime directement par une rotation selon l'axe principal aligné.

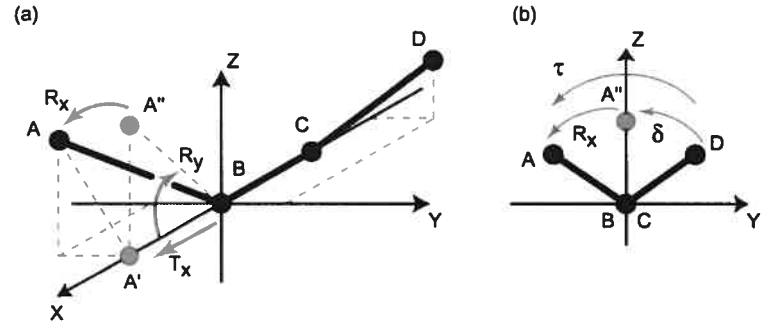


Figure 26 : Positionnement d'un atome dans son référentiel. (a) Soit l'atome A à positionner, considérant les trois atomes B, C et D positionnés précédemment à l'atome A et successivement liés. A doit être positionné selon la longueur du lien covalent AB, l'angle de coude ABC et l'angle de torsion ABCD. Le système de référence où B est à l'origine, C sur l'axe des X négatif d'une valeur absolue équivalente à la longueur du lien covalent BC, et D à une position quelconque s'obtient en réalisant d'abord la série complète de transformation linéaire qui a mené au positionnement de B. Dans ce système de référence, la distance du lien AB est obtenu simplement par une translation selon l'axe des X d'une valeur équivalente à la longueur du lien (T_x), ce qui porte A en A'. En A', une rotation selon l'axe des Y d'un angle R_y exprime directement l'angle de coude ABC, ce qui porte A' en A". Finalement, en A", une rotation selon l'axe des X d'un angle R_x exprime la torsion ABCD, et porte A dans sa position finale. (b) Cependant, l'angle R_x n'est pas nécessairement directement équivalent à l'angle de torsion ABCD voulu, tout dépendant de la position de D dans ce système de référence. Soit τ la torsion ABCD voulue et δ la torsion A"BCD. Sur le plan YZ, on voit que l'angle de rotation R_x est simplement $\tau - \delta$.

Cette méthode de construction atomique par transformation linéaire relative au référentiel de l'atome précédent implique un ordre dans la succession des atomes du ribose à construire. Si on reconnaît dans le modèle du ribose un graphe où les nœuds sont les atomes et les arêtes les liens covalents, un ordre de construction est équivalent à un arbre de recouvrement de ce graphe représentatif. Puisque le ribose contient un cycle (C1'-C2'-C3'-C4'-O4'), il existe plusieurs arbres de recouvrement qui utilisent différents sous-ensembles d'arêtes desquels aucun ne contiennent la totalité des arêtes du graphe. En particulier, la figure 27 illustre l'arbre choisi, ayant comme racine l'atome C1' et se présentant dans le référentiel de ce dernier. On peut voir que la construction positionne C3' par rapport à C4' (et *vice-versa*) de façon implicite, leur géométrie conjointe – longueur du lien covalent C3'-C4' et angles de coude C2'-C3'-C4' et C3'-C4'-O4' – n'étant pas exprimée par aucune transformation

linéaire. De plus, puisque le cycle contient cinq atomes et qu'une torsion se calcule selon quatre atomes successifs, seulement deux des cinq torsions ν_0 à ν_4 peuvent être explicitement exprimées. Ici ν_0 et ν_1 sont explicitement exprimées par le positionnement respectif des atomes C3' et C4', laissant ν_2, ν_3 et ν_4 s'exprimer de façon implicite. La conservation de toutes ces géométries implicites près des mesures de référence devient donc une mesure de précision de cette méthode de construction.

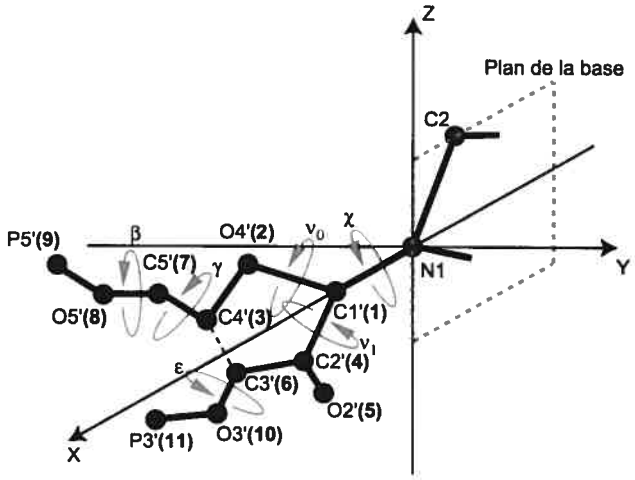


Figure 27 : Étapes de positionnement des atomes du ribose. La base azotée présentée ici est de la famille des pyrimidines. Pour une purine, interchanger les atomes N1 pour N9 et C2 pour C4 dans le reste de cette description. Le référentiel présentée ici est le référentiel initial du positionnement. Il est centré sur N1 et le plan XZ est aligné sur le plan de la base azotée. L'axe des X est orienté de manière à former l'angle de coude C2-N1-C1'. Ainsi, la construction débute en positionnant C1' par une simple translation. Les positionnements successifs se réalisent dans l'ordre numéroté sur la figure et de la manière illustrée à la figure 26, en impliquant les torsions présentées. Toutes les géométries rigides impliquant à la fois C3' et C4' ainsi que les torsions non présentées sont implicitement exprimées par cette méthode de construction. Leur conservation est une mesure de la précision de la construction du ribose.

Ainsi, cette méthode de construction utilise six paramètres de torsions : ν_0, ν_1, γ, β et ε. Cependant, la pseudorotation ρ mets en relation ν_0 à ν_4 selon l'équation 1 présentée précédemment. La série d'équations suivante est la source de cette relation [20] :

$$\theta_j = \theta_m \cos(\rho + j\delta), \delta = 144^\circ, j = 0\ldots4$$
$$\nu_j = \theta_{(j+2) \bmod 5} \quad (2)$$

Où θ_m est l'angle maximal de torsion. Pour extraire ν_0 à ν_4, il faut donc mesurer θ_m. La méthode la plus simple déduit θ_m directement de θ_0 (ν_2) et de ρ à partir de l'équation précédente :

$$\theta_m = \frac{\theta_0}{\cos\rho} \tag{3}$$

Les cinq angles de torsion ν_0 à ν_4 ont donc été mesurés sur le modèle du ribose qui a servi au prélèvement des géométries rigides (voir tableau 4) desquels ρ a pu être déduit, pour enfin obtenir θ_m :

$$\begin{aligned}
\theta_0 &= \nu_2 = 36.84^\circ \\
\theta_1 &= \nu_3 = -35.38^\circ \\
\theta_2 &= \nu_4 = 19.62^\circ \\
\theta_3 &= \nu_0 = 4.37^\circ \\
\theta_4 &= \nu_1 = -26.23^\circ \\
\rho &= \arctan\frac{(\theta_2+\theta_4)-(\theta_1+\theta_3)}{2\theta_0(\sin 36^\circ + \sin 72^\circ)} = 12.14^\circ \\
\theta_m &= \frac{\theta_0}{\cos\rho} = 37.68^\circ
\end{aligned} \tag{4}$$

Cet angle de torsion maximal θ_m est conservé pour pouvoir extraire ν_0 et ν_1 pour toutes valeurs de ρ, selon l'équation 2. La torsion θ_m devient donc un paramètre constant de la méthode de construction, tout comme la géométrie rigide. Ainsi, la pseudorotation ρ remplace les paramètres ν_0 et ν_1 dans la méthode, réduisant la quantité de paramètres à cinq.

3.3.2 Évaluation de la méthode de construction

L'évaluation de la méthode de construction décrite ci-haut passe par l'appréciation de trois critères : la précision, la représentativité et l'efficacité.

3.3.2.1 Évaluation de la précision

D'abord, pour évaluer la précision de cette méthode de construction du ribose, à l'intérieur du cadre géométrique présenté au tableau 4 ainsi qu'à l'endossement de la validité expérimentale de l'équation 2, il suffit de mesurer sur les modèles de ribose construits la déviation à ces mesures de référence. L'évaluation de la précision ne s'intéresse ici qu'à la géométrie du cycle furanosique. Donc, la méthode de construction évaluée prend comme paramètre seulement la pseudorotation et construit le cycle furanosique selon cette pseudorotation. Le critère de précision se subdivise en quatre sous-critères. Les deux premiers sous-critères mesurent les éléments de géométrie considérés rigides qui sont implicitement exprimés par la méthode de construction (voir les zones ombrées du tableau 4). Ainsi, le premier sous-critère mesure la longueur du lien covalent C3'-C4' dans les riboses construits et compare la mesure avec celle de référence. Cette mesure quantifie la fermeture du cycle furanosique par la méthode de construction. Quant au deuxième sous-critère, il mesure la différence entre tous les angles de coude implicitement exprimés et leur mesure de référence.

De plus, la différence entre les cinq angles de torsions mesurés sur le cycle furanosique construit et les mêmes angles de torsion calculés à partir de la pseudorotation soumise à la méthode par l'équation 2 représente le troisième critère de précision. Enfin, parallèlement au critère précédent, le quatrième sous-critère mesure la différence entre la pseudorotation soumise et la pseudorotation induite par la construction. À la figure 28, nous rapportons les mesures de ces quatre sous-critères de la précision du cycle furanosique construit sur une période complète de pseudorotation. Ces résultats montrent que la géométrie rigide du cycle furanosique est très bien reproduite par la méthode : des erreurs absolues moyennes de 0.058 Å pour la fermeture du cycle et 1.41° pour les angles de coude induits. Pour la représentation des cinq angles de torsion du cycle, une moyenne de 0.68° de variation s'observe avec les valeurs calculées à partir de la pseudorotation initiale. Finalement, la pseudorotation elle-même varie en moyenne de 1.60° lors de la construction. De façon générale, certains intervalles de pseudorotation, correspondant à certains modes de *puckering*, sont représentés de façon moins précise que d'autres selon tous les sous-critères : par exemple, les modes C1' endo, C1' exo, O4' endo et O4' exo semblent toujours présentés les plus importantes déviations.

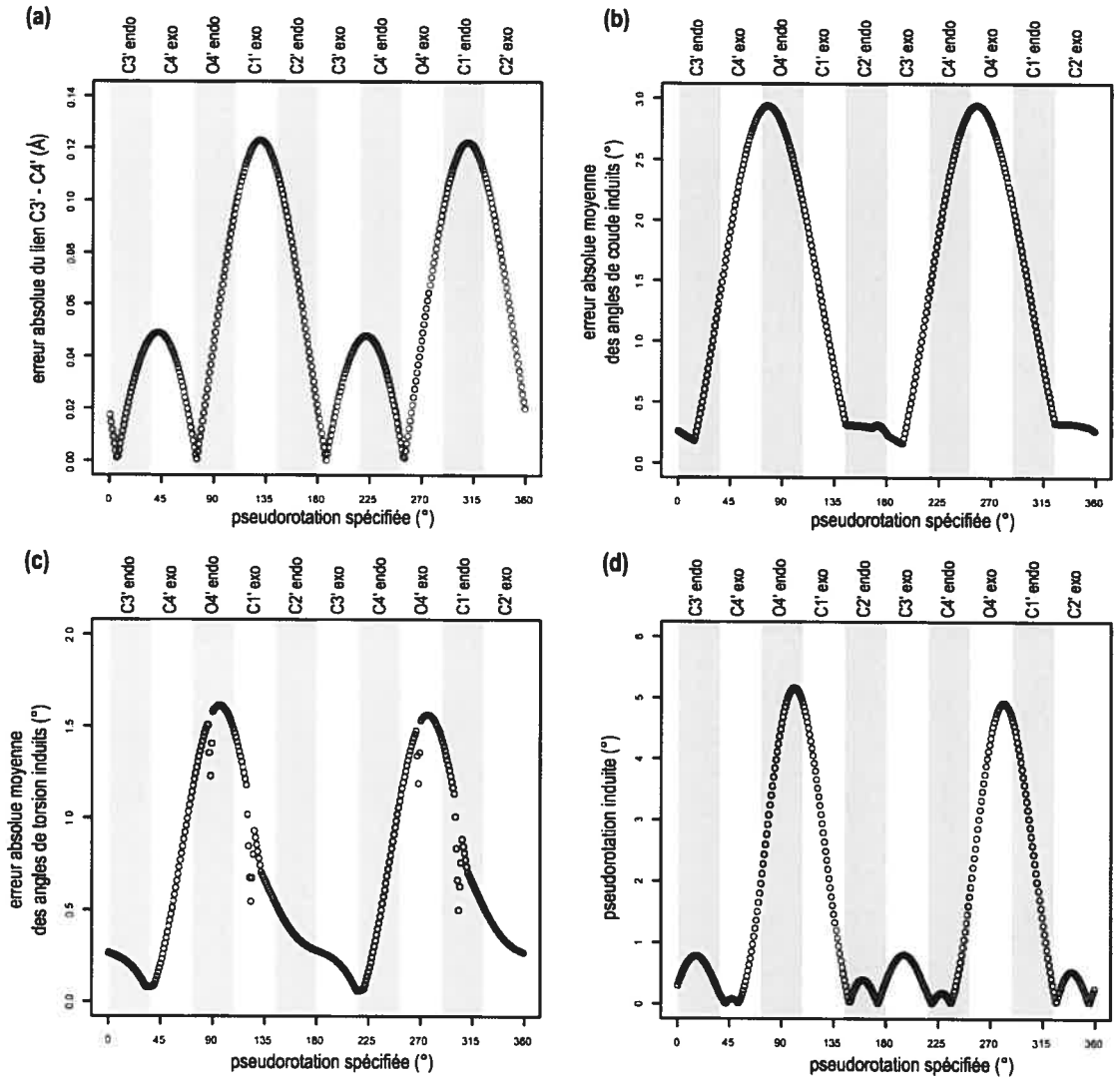


Figure 28 : Évaluation de la précision de la construction du cycle furanosique. La méthode de construction prend comme parametre l'angle de pseudorotation et construit le cycle furanosique. Quatre critères sont ici évalués sur une période complète de pseudorotation. (a) Le premier critère mesure l'erreur absolue sur le lien covalent implicite C3'-C4' du furanose construit, c'est-à-dire la fermeture du cycle. (b) Le second critère mesure les angles de coude non explicitement exprimés par la méthode de construction et rapporte la déviation moyenne avec les angles de référence. (c) Le troisième critère s'intéresse aux angles de torsion ν_0 à ν_4 construits et rapporte la déviation moyenne avec les torsions induites par la pseudorotation soumise à la méthode selon l'équation 2. (d) Enfin, le quatrième critère rapporte la différence entre la pseudorotation soumise à la méthode et celle mesurée après la construction.

3.3.2.2 Évaluation de la représentativité

Le critère d'évaluation précédent s'est intéressé à la précision du ribose construit selon le modèle de référence. Bien que ce critère valide la méthode de façon nette par rapport à lui-même et assure techniquement une réponse précise de la méthode pour tout paramètre de construction, il reste qu'il se réfère à l'unique modèle du ribose dont les mesures sont détaillées au tableau 4. Pour élargir la portée de l'évaluation, il faut évaluer la représentativité de cette méthode de construction du ribose, c'est-à-dire qu'il faut comparer les modèles de ribose construits avec un ensemble de différents modèles de référence. Pour ce faire, nous avons créé un ensemble de modèles de référence par l'extraction des riboses des modèles d'ARN provenant des bases de données publiques *PDB* [46] et *NDB* [47] d'une résolution supérieure à 3Å (selon l'état des bases de données en juillet 2003). Ainsi, soit **R** cet ensemble référence. Notre extraction a isolé 79 231 modèles de ribose desquels seulement 1457 subsistent après un passage dans un filtre de similitude par déviation RMS à 0.5 Å. Pour se comparer à cet ensemble de référence, nous avons créé un second ensemble de modèles de ribose cette fois-ci construits par la méthode et couvrants l'ensemble de l'espace conformationnel atteignable par la méthode. La méthode de construction prend cinq paramètres d'entrée : la pseudorotation du cycle furanosique ρ et les torsions χ, γ, β et ε. Ainsi, nous avons créé l'ensemble **T** des riboses construits en faisant varier ces cinq angles sur une période complète de façon discrète. Des incréments de 15° pour ρ et χ et de 30° pour γ, β et ε ont généré 995 328 modèles, dont 32 137 sont conservés après un filtre de déviation RMS à 0.5 Å.

Finalement, la comparaison des ensembles R et T consiste en la mesure de représentativité de la méthode de construction. Cette comparaison se subdivise en deux sous-critères : la couverture de la référence (R→ T) et l'homologie de la construction (T→ R). La couverture de la référence trouve pour chaque modèle de R le modèle de T le plus près par déviation RMS et rapporte la mesure. Ce sous-critère de représentativité assure que l'espace des modèles de référence est couvert maximalement par la méthode de construction. Pour ce qui est de l'homologie de la construction, ce sous-critère s'intéresse inversement à trouver pour chaque modèle de T le modèle de R le plus près. Ce sous-critère assure plutôt que l'espace des modèles construits (T) reste inclus dans l'espace de référence (R). À la figure 29, nous présentons les résultats de comparaison de R et T selon le critère de représentativité, à la fois

pour le ribose complet ainsi que pour le cycle furanosique uniquement. Ces résultats démontrent que, avec un seuil de 1Å, on s'attend à retrouver près de 100% des modèles de ribose de R dans T, et environ 75% des modèles de T dans R. Par conséquent, la méthode de construction couvre l'entièreté de l'espace conformationnel de référence, mais dépassent un peu les bornes de cet espace de 25%. Cependant, en ne considérant que le cycle furanosique dans la mesure de la représentativité, le seuil à 1Å est atteint à près de 100% dans les deux directions. La représentativité de la méthode de construction est donc excellente, compte tenu que les modèles de référence sont à une résolution maximale de 3Å.

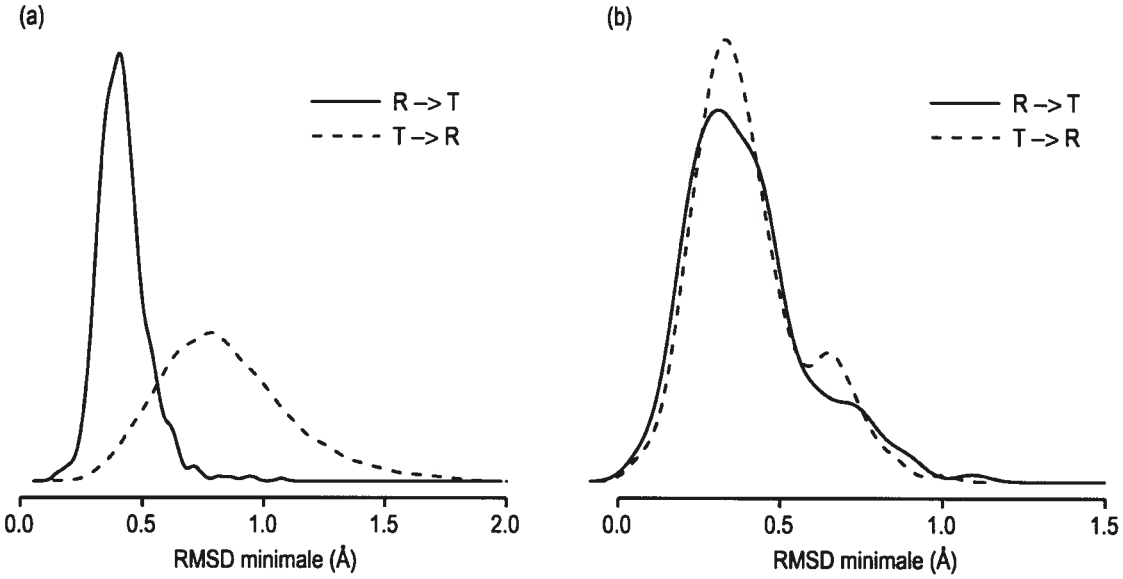


Figure 29 : Résultats de l'évaluation de la représentativité de la méthode de construction. Les courbes représentent les fonctions de densité estimées empiriquement sur les ensembles de mesures de déviation RMS. R réfère à l'ensemble de référence des modèles de ribose, alors que T réfère à l'ensemble théorique des modèles construit par la méthode. (a) Évaluation du critère lorsque le ribose complet est considéré. La courbe en trait plein représente la distribution des mesures minimales de déviation RMS pour chaque modèles de R comparés à ceux de T, soit la tendance du sous-critère de couverture de la référence. La courbe en trait pointillé représente la distribution des mesures inverses de T vers R, soit la tendance du sous-critère d'homologie de la construction. (b) Ici seulement le cycle furanosique est considéré lors de la comparaison. Ces courbes ont été évaluées à l'aide du logiciel de calcul statistique *R* [43].

3.3.2.3 Évaluation de l'efficacité

Par ce dernier critère d'évaluation, nous voulons démontrer l'efficacité technique de notre méthode de construction, c'est-à-dire ses performances calculatoires. Théoriquement, chaque atome se positionne depuis l'origine par une succession de transformations linéaires, i.e. par

la multiplication d'une série de matrices. Selon la procédure illustrée à la figure 26, chaque atome impliqué dans la chaîne de transformations ajoute trois matrices de transformation linéaire, translation pour le lien covalent, rotation pour l'angle de coude et rotation pour la torsion. De ces trois transformations, seulement la dernière est variable, les deux premières sont toujours constantes. Ainsi, techniquement, pour chaque atome à positionner, 2/3 des matrices présentes dans la série de transformations peuvent être pré-multipliées et stockées de façon statique. Mise à l'épreuve, cette approche accélère la méthode de construction d'un facteur de 237% par rapport à l'approche où toutes les matrices de transformations sont explicitement calculées. Le tableau 5 présente des résultats comparatifs plus détaillés sur cette accélération technique. Évidemment, les performances de cette optimisation sont complètement dépendantes du compilateur et de l'architecture machine.

	+	*	Δ	√	=	t (μs)
originale	1160	1491	42	8	1374	16.8
pré-multiplicative	531	692	18	4	428	7.2
accélération	218%	215%	233%	200%	321%	233%
					accélération moyenne	237%

Tableau 5 : Résultats comparatifs de l'approche de construction par matrices pré-multipliées. L'approche originale calcule explicitement toutes la matrices de transformations linéaires nécessaire au positionnement des atomes du ribose. L'approche pré-multiplicative veut profiter du constat que 2/3 des matrices de transformation nécessaire sont constantes, dû à la rigidité des liens covalents et angles de coude. Ainsi, pour chaque atome, la série de matrices menant à sa position globale est pré-multipliée en une seule matrice résultante, où chaque élément de la matrice est une équation linéaire dont les variables sont les paramètres libres, c,est-à-dire les angles de torsions. Les critères de comparaison sont la quantité d'opérations additives '+' (addition, soustraction) et multiplicatives '*' (multiplication, division), ainsi que d'opérations de triginométrie 'Δ' de racine carré '√' et d'assignation de variable '='. Tous ces critères sont relatifs à la construction d'un seul ribose. Le temps de calcul 't' est aussi comparé, représentant le temps moyen de construction d'un ribose calculé sur un ensemble de construction de 10^7 riboses sur un processeur *AMD Athlon* à 1.2 GHz.

3.4 Détermination des paramètres de la construction

À la section précédente, nous nous sommes intéressé à la description formelle de notre méthode de construction artificielle d'un ribose hors du contexte de la modélisation. Dans cette section, la méthode est mise à l'épreuve en contexte de modélisation pour résoudre le problème suivant : soient les positions tridimensionnelles d'une base azotée et des deux

groupements phosphates adjacents de part et d'autre à celle-ci. Trouver une affectation au quintuplet de paramètres $< \rho, \chi, \gamma, \beta, \varepsilon >$ tel que la méthode construise, selon ce quintuplet, un modèle de ribose depuis la base azotée de telle façon que les groupements phosphates construits par la méthode et ceux déjà positionnés par la modélisation soient confondus. Avant de s'attaquer au problème de la détermination des paramètres optimaux, il faut d'abord définir une métrique de qualité du ribose construit.

3.4.1 Mesure de qualité de la construction

Pour évaluer si un quintuplet de paramètres construit un meilleur ribose qu'un autre étant donné un contexte de modélisation, nous avons développé une mesure de la qualité de la construction. Un contexte de modélisation positionne dans l'espace tridimensionnel une base azotée et ses deux phosphates adjacents de part et d'autre. La méthode de construction s'aligne sur la base azotée et positionne chaque atome du ribose jusqu'aux atomes de phosphore des phosphates. Ainsi, pour évaluer la qualité de cette construction, nous mesurons la déviation entre ces atomes de phosphore construits et les groupements phosphates positionnés par la modélisation. Une façon simple de quantifier cette déviation est de mesurer la distance entre les atomes construits O5', P5', O3' et P3' et leur homologue dans les phosphates déjà positionnés (voir figure 30a). La somme de ces quatre mesures de distance constitue une mesure de la qualité du ribose construit en contexte. En utilisant cette mesure de qualité, la liaison finale du modèle de ribose construit avec les phosphates donnés se réalise en éliminant les atomes O5', P5', O3' et P3' du ribose construit et en reliant directement ses atomes C5' et C3' respectivement aux phosphates en direction 5' et 3' par les atomes O5' et O3'.

À la figure 30b, nous illustrons une seconde mesure de qualité basée uniquement sur la longueur du lien covalent implicite entre l'atome C5' du ribose construit et l'atome O5' du phosphate positionné en direction 5' ainsi que celle du lien covalent entre l'atome C3' du ribose construit et l'atome O3' du phosphate positionné en direction 3'. L'idée est de ne considérer que l'atteinte d'une extrémité du phosphate par le ribose construit, c'est-à-dire de négliger la position relative des atomes de phosphore construits et positionnés. En utilisant cette seconde mesure de qualité, la liaison finale du modèle du ribose construit avec les phosphates donnés se réalise de la même façon qu'en utilisant la première. Cependant,

l'omission des atomes construits O5', P5', O3' et P3' dans la mesure de qualité permet de les négliger lors de la construction du ribose, bref de ne construire que le cycle furanosique. Par conséquent les paramètres γ, β et ε deviennent obsolètes, réduisant conséquemment le nombre de paramètres à deux, soient ρ et χ.

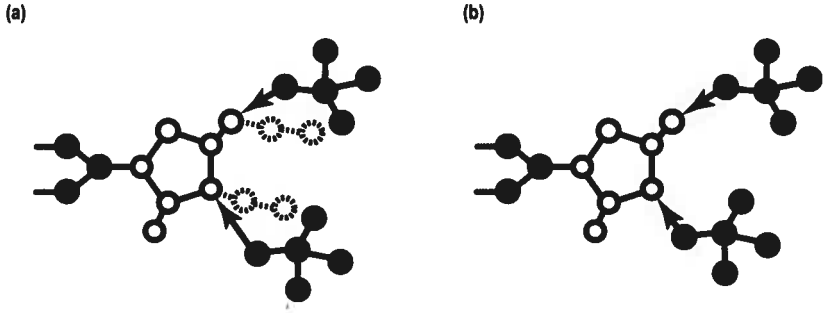


Figure 30 : Deux mesures de la qualité du ribose construit dans un contexte de modélisation. Les cercles vides sont les atomes du ribose construit par la méthode et les cercles pleins sont les atomes des éléments du contexte de la modélisation : la base azotée et les deux phosphates adjacents. (a) Une fois le ribose construit depuis la base, la déviation des atomes construits dans les branches de phosphate (trait pointillé) avec leur homologue dans les phosphates déjà positionnés constitue une première mesure de qualité. Pour réunir le ribose construit aux phosphates, les atomes construits dans les branches de phosphate sont éliminés. (b) Par ailleurs, une seconde mesure de qualité s'intéresse uniquement à la longueur des liens covalents implicitement créés par l'élimination des branches de phosphate construites et la jonction avec les phosphates positionnés. Ici, les branches de phosphate deviennent inutiles à construire.

3.4.2 Optimisation

Soit $f(\tau_1,\tau_2,\tau_3,\tau_4,\tau_5)$ la fonction objectif qui construit le ribose selon les torsions τ_1 à τ_5, respectivement ρ, χ, γ, β, ε, et qui retourne la mesure de la qualité du ribose construit selon ces paramètres et le contexte de modélisation. La mesure de qualité se fait en $Å^2$ pour des raisons d'efficacité technique. La minimisation de f selon ses cinq variables indépendantes détermine le quintuplet $< \rho, \chi, \gamma, \beta, \varepsilon >$ qui construit le ribose de la meilleure qualité. Pour éviter les dérivations partielles de f, une méthode d'approximation numérique s'impose. La méthode choisie est une simplification de *cyclic coordinate method* présentée par Bazaraa et Shetty [23]. Cette méthode numérique consiste à minimiser f dans chaque dimension prise indépendamment des autres et de façon itérative jusqu'à ce que la minimisation des cinq dimensions isolées n'ait amélioré f que d'une façon négligeable. Puisque f est une observation, c'est-à-dire qu'elle est inconnue analytiquement, sa minimisation unidimensionnelle se fait à l'aveuglette. Cette simplification se borne à un pas constant dans

chaque dimension, où il suffit alors de tenter le pas dans chaque direction de chaque dimension pour déterminer si *f* diminue dans cette dimension ou si la position courante est optimale compte tenu du pas courant. Le dernier cas implique une diminution du pas pour raffiner la recherche du point optimal. Cette méthode ainsi que d'autres variantes plus ou moins exotiques ont été testées par Lemieux [24] qui en est arrivé à la conclusion que, dans un contexte de modélisation discrète où le temps utilisé pour construire un ribose est critique, cette simplification du *cyclic coordinate method* suffit à approximer un ribose de qualité appréciable et dans le délai le plus bref. Le pseudocode de l'algorithme est présenté à la figure 31. CCM-5D est lancé avec un incrément initial de 180°. Deux paramètres sont à considérer : ϕ et λ, respectivement l'incrément minimal et le seuil de déplacement. L'incrément minimal est simplement le critère d'arrêt de la minimisation et permet d'ajuster la finesse de l'approximation numérique. Le seuil de déplacement paramètre le degré de courbure nécessaire à la portion de la fonction couverte par le pas de recherche pour permettre un déplacement. Ces paramètres ont été ajustés par essais et erreurs à $\phi = 0.1^{\circ}$ et $\lambda = 1 \times 10^{-5}$ Å^2.

Par ailleurs, à la sous-section précédente, nous avons présenté une mesure de qualité qui réduit les paramètres nécessaires à la méthode de construction du ribose à seulement ρ et χ. Ainsi, une seconde fonction objectif g (τ_1, τ_2) construit le cycle furanosique selon les torsions τ_1 et τ_2, respectivement ρ et χ, et retourne la qualité du ribose construit en mesurant la somme des longueurs des liens covalents C5'-O5' et C3'-O3'. Cependant, cette mesure de qualité ne tient pas compte des angles de coude C5'-O5'-P5' et C3'-O3'-P3'. Le même algorithme de minimisation par approximation numérique est utilisé, mais en deux dimensions plutôt qu'en cinq. Ainsi, CCM-2D détermine le couple $< \rho, \chi >$ qui minimise g.

```
procédure CCM-5D
entrée : φ, δ, τ1,...,τ5
1.   incrément ← 180°
2.   tant que incrément > φ faire
3.   début
4.     immobile ← VRAI
5.     pour chaque τi ← τ1...τ5 faire
6.     début
7.       τi' ← τi + incrément                    (recherche à droite)
8.       si f(τ1,...,τi',...,τ5) < f(τ1,...,τi,...,τ5) − δ alors
9.         τi ← τi'                              (exécution du pas)
10.        immobile ← FAUX                       (confirmation du mouvement)
11.      sinon
12.        τi' ← τi − incrément                  (recherche à gauche)
13.        si f(τ1,...,τi',...,τ5) < f(τ1,...,τi,...,τ5) − δ alors
14.          τi ← τi'                            (exécution du pas)
15.          immobile ← FAUX                     (confirmation du mouvement)
16.    fin
17.    si immobile = VRAI alors
18.      incrément ← incrément/2                 (raffinement du pas)
19.  fin
20. retourner f(τ1,...,τ5)
21. fin de la procédure
```

Figure 31 : Pseudocode pour la procédure CCM-5D qui implante un algorithme numérique linéaire de minimisation sans dérivation. L'algorithme est une simplification de *cyclic coordinate method* [23]. La procédure prend en entrée ϕ et δ, respectivement l'incrément minimal et le seuil de déplacement, et retourne l'évaluation de la fonction objectif *f* sur les cinq torsions optimisées.

3.4.3 Estimation

Jusqu'à maintenant, nous avons présenté deux méthodes de détermination des paramètres libres de la construction du ribose par optimisation numérique linéaire. Cependant, il s'avère que l'espace des torsions mesurées sur des modèles de ribose de référence est plutôt éparse. La figure 32 illustre l'espace 2D formé par la mesure des couples $< \rho, \chi >$ sur les riboses extraits des modèles d'ARN de référence provenant de *PDB* [46] et de *NDB* [47]. En arrondissant chaque mesure à 5° près, l'espace bidimensionnel $< \rho, \chi >$ ne reste couvert qu'à 30%. Pire, en faisant le même exercice pour l'espace 5D formé par les quintuplets $< \rho, \chi, \gamma, \beta, \varepsilon >$, la couverture devient presque négligeable (0.002%). La figure 32 confirme

que, même lorsqu'ils sont pris séparément, les paramètres $< \rho, \chi, \gamma, \beta, \varepsilon >$ ne sont pas très bien dispersés sur l'espace de référence, spécialement pour ρ.

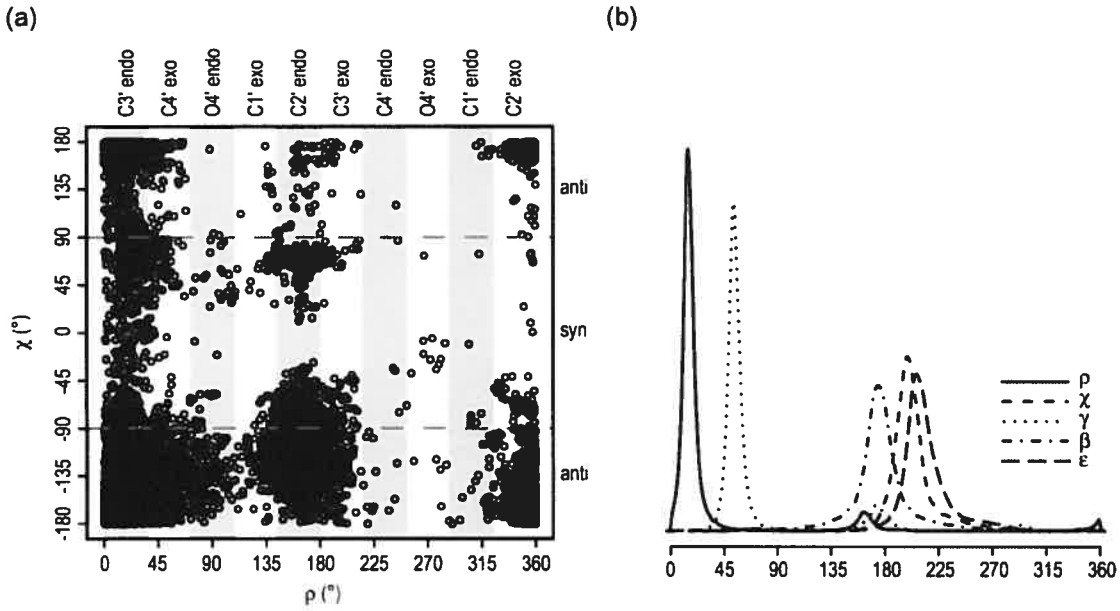


Figure 32 : Dispersion des torsions caractéristiques mesurées sur les riboses des modèles d'ARN de référence. (a) Mesure des couples $< \rho, \chi >$. (b) Estimation des courbes de densité empirique [43] pour chacun des cinq paramètres de la méthode de construction.

Ainsi, la détermination des paramètres $< \rho, \chi, \gamma, \beta, \varepsilon >$ par la méthode d'optimisation CCM-5D a vraisemblablement une plus grande propension à éviter les zones de quintuplets valides de l'espace 5D qu'autrement. La réduction à deux dimensions sur l'espace formé par les couples $< \rho, \chi >$ augmente certes cette propension, cependant la couverture reste mince. Ce que nous avons conclu à partir de cette observation est qu'une méthode d'optimisation numérique linéaire, telle CCM-5D, n'est pas souhaitable, d'autant plus que chaque pas de la recherche doit être validé par une construction complète du ribose. Par conséquent, nous avons étudié une méthode d'estimation directe des paramètres, en temps constant. Celle-ci s'appuie sur la méthode de construction du cycle furanosique seulement, de la même façon que la méthode CCM-2D. Deux paramètres sont donc à estimer, ρ et χ. La construction du cycle furanosique positionne l'atome O3', puisqu'il est relié directement au furanose par le lien covalent C3'-O3'. Or, l'atome O3' se retrouve aussi dans le contexte de modélisation, plus précisément dans le phosphate en direction 3'. Donc, une mise en relation de ρ et χ avec la position résultante de O3' implique que la position de l'atome O3' du phosphate du contexte pourrait déterminer analytiquement les valeurs respectives de ρ et χ. Pour découvrir

cette relation, il faut examiner les positions possibles de l'atome O3' dans les différentes constructions du ribose. Dans le référentiel initial de construction aligné sur le lien glycosyl (voir figure 27), χ s'exprime directement par une rotation selon l'axe des X. Donc, la coordonnée X de l'atome O3' ne varie pas en fonction de χ. À la figure 33 , nous avons tracé la projection des positions de O3' dans le plan YZ pour une période complète du couple $< \rho, \chi >$. L'idée est d'isoler ρ et χ dans deux relations indépendantes. La figure 33a décrit le déplacement particulier de O3' en fonction de ρ en fixant χ à une valeur constante. Dans le plan YZ, la forme décrite ressemble à un pétale de fleur, ce qui est loin d'être une conique. De plus, l'observation de plusieurs intersections entre ces pétales, chaque pétale correspondant à une valeur particulière de χ, combiné au fait que O3' varie en X selon ρ font qu'une relation où χ est indépendant semble plutôt difficile à déterminer. Cependant, si ρ est fixe et χ varie, O3' décrit un cercle centré à l'origine dans le plan YZ, comme l'illustre la figure 33b, dont le rayon varie en fonction de la valeur fixée pour ρ. De surcroît, le référentiel choisit assure que ce cercle est parfaitement aligné sur le plan YZ, c'est-à-dire que la coordonnée X n'a pas d'impact sur la relation entre ρ et le rayon du cercle décrit.

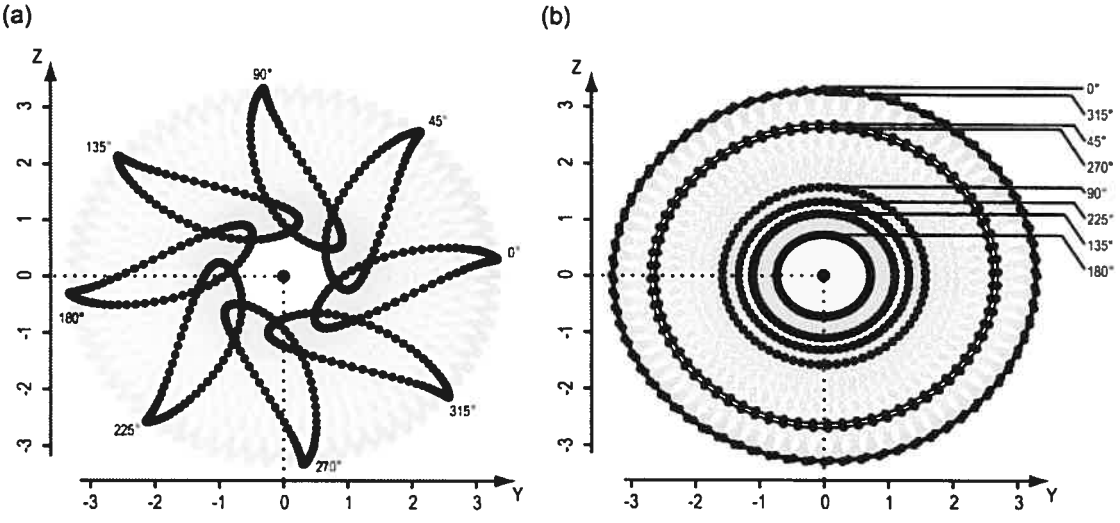


Figure 33 : Projection des positions de l'atome O3' dans le plan YZ du référentiel initial de construction du ribose (en gris) pour une période complète du couple $< \rho, \chi >$. (a) Déplacement selon la valeur de ρ pour différentes valeurs constantes de χ (en noir). (b) Déplacement selon la valeur de χ pour différentes valeurs de ρ (en noir).

Il y a donc de l'espoir dans la détermination d'une relation entre la position de l'atome O3' et ρ. En effet, lorsque le rayon du cercle décrit en YZ par O3' est rapporté en fonction de la valeur fixe ρ, la courbe observée s'approche d'une fonction cosinus, tel que présenté à la

figure 34. La forme générale d'une fonction cosinus $f(\theta) = A\cos(\theta + \lambda) + D$ possède trois paramètres libres : l'amplitude (A), la phase (λ) et le déplacement vertical (D). Il suffit alors de trouver le triplet < A, λ, D > qui minimise la somme des distances au carré entre les points tracés par $f(\theta)$ et les points observés sur la figure 34. Nous avons déterminé de façon itérative par essais et erreurs que le triplet < 1.3305, 17.4236°, 2.0778 > est optimal, à une précision de 0.01%. La figure 34 illustre cette courbe théorique en superposition à la courbe observée. Puisque le cercle décrit par O3' est toujours centré à l'origine, le rayon du cercle n'est d'autre que la distance avec l'origine de O3' projeté sur le plan YZ. Donc, en inversant la fonction cosinus préalablement déterminée, nous obtenons une relation d'estimation de ρ en fonction de la position YZ de O3' :

$$\rho = \arccos(\Omega) - 17.4236^{\circ} \quad où \ \Omega = \frac{\sqrt{y^2 + z^2} - 2.0778}{1.3305} \tag{5}$$

Où y et z sont respectivement les coordonnées Y et Z de O3'. Seul point litigieux, il faut que $-1 \le \Omega \le 1$ pour que $\arccos(\Omega)$ soit définie. En se basant sur les modèles de référence, nous avons observé que 7.5% des riboses sur lesquels Ω a été mesuré ne respectent pas cette condition. En permettant un dépassement de 0.2 des limites du domaine de $\arccos(\Omega)$ (si $-1.2 \le \Omega < -1$ ou $1 < \Omega \le 1.2$ alors respectivement $\Omega = -1$ ou $\Omega = 1$), ce taux de rejet est réduit à 0.4%. De plus, l'erreur sur l'estimation de la pseudorotation par l'équation 5 sur les riboses construits est de seulement 6.96° en moyenne. En faisant le même exercice sur les modèles de référence, l'erreur maximale sur l'estimation de la pseudorotation passe à 9.57° en moyenne.

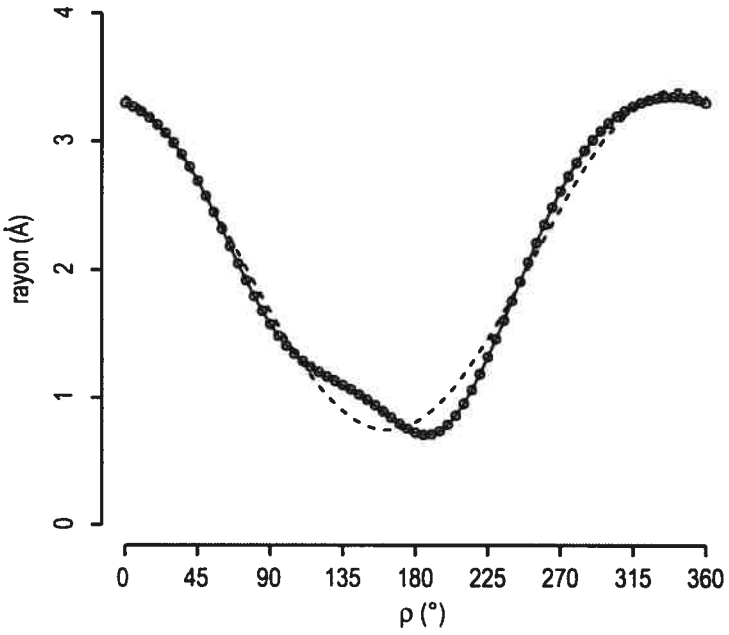


Figure 34 : Courbe d'estimation de ρ en fonction de la distance à l'origine de la projection de l'atome O3' dans le plan YZ du référentiel initial de construction. La courbe en trait plein relie les mesures observées à la figure 33. La courbe en trait pointillé représente une estimation par une fonction cosinus qui minimise la somme des distances au carré avec les points d'observation (voir équation 5).

Grâce à l'équation 5 qui met en relation par une fonction cosinus la pseudorotation avec la position de l'atome O3' du contexte de modélisation, il nous est dorénavant possible d'estimer les paramètres ρ et χ qui construisent le cycle furanosique dans un contexte de modélisation. D'abord, ρ s'obtient en alignant la base azotée et les phosphates du contexte sur le référentiel initial de construction du ribose et en mesurant la distance à l'origine de la projection de l'atome O3' sur le plan YZ. Ensuite, l'équation 5 donne un estimé de ρ selon cette distance mesurée. Cependant, l'inverse d'une fonction cosinus résulte toujours en deux valeurs : l'angle entre 0° et 180° ainsi que son homologue dans la demi-période adjacente, c'est-à-dire son complément à 360°. L'estimation de ρ résulte donc en deux mesures, ρ_1 et ρ_2. Pour l'estimation de χ, aucune relation n'a été trouvée avec les éléments du contexte de modélisation. Par contre, dans le référentiel initial, χ s'exprime directement par une rotation selon l'axe des X. Donc, une fois la valeur de ρ fixée, il suffit de construire le ribose avec $\chi = 0$ et de mesurer l'angle de torsion selon l'axe des X entre l'atome O3' construit selon ces paramètres et l'atome O3' du contexte. Cette torsion est équivalente à la valeur de χ nécessaire pour déplacer O3' de sa position à $\chi = 0$ jusqu'à sa position en contexte, selon le plan YZ. Donc, deux couples de paramètres sont estimés : $< \rho_1, \chi_1 >$ et $< \rho_2, \chi_2 >$. Pour

choisir, il faut construire les deux modèles de ribose et conserver celui qui minimise la mesure de la qualité (voir sous-section 3.4.1). Ainsi, notre méthode d'estimation des paramètres ρ et χ implique un total de deux applications de la méthode de construction. D'ailleurs, c'est l'avantage majeur de cette méthode d'estimation des paramètres sur celles par optimisation, ces dernières n'assurant aucune constance quant à la quantité requise d'applications de la méthode de construction.

3.4.4 Comparaison des méthodes de détermination

Dans la sous-section précédente, nous avons décrit trois différentes méthodes de détermination des paramètres de la construction artificielle du ribose : optimisation à cinq dimensions, optimisation à deux dimensions et estimation. Dans cette sous-section, nous mettons à l'épreuve nos trois méthodes de détermination et nous évaluons leurs performances respectives. L'épreuve consiste en l'application de la méthode de construction selon les paramètres mesurés sur tous les riboses des modèles de référence provenant de *PDB* et de *NDB*, c'est-à-dire sur tous les 1457 riboses de l'ensemble R (voir sous-section 3.3.2.2). Chaque ribose est reconstruit par chacune des trois méthodes : CCM-5D, CCM-2D et l'estimateur. Pour chaque ribose, la reconstruction artificielle est comparée avec chaque modèle original de l'ensemble R et la déviation RMS minimale est rapportée. La figure 35a montre la distribution de ces mesures pour les trois méthodes. Selon ces résultats, la méthode CCM-5D est la plus précise : 98.6% des reconstructions ont une déviation RMS minimale inférieure à 1Å, comparativement à 95.6% pour la méthode CCM-2D et 93.6% pour la méthode par estimation. Cependant, le degré de précision atteint par les méthodes CCM-5D et CCM-2D coûte une certaine quantité variable d'itérations de construction, comme en témoigne la figure 35b. En effet, pour CCM-5D, la quantité d'itérations de construction nécessaire se situe majoritairement entre 65 et 110 itérations, alors qu'elle oscille plutôt entre 25 et 45 pour CCM-2D. L'estimateur, quant à lui, est le grand gagnant à ce niveau, puisqu'il nécessite une quantité toujours constante de deux itérations.

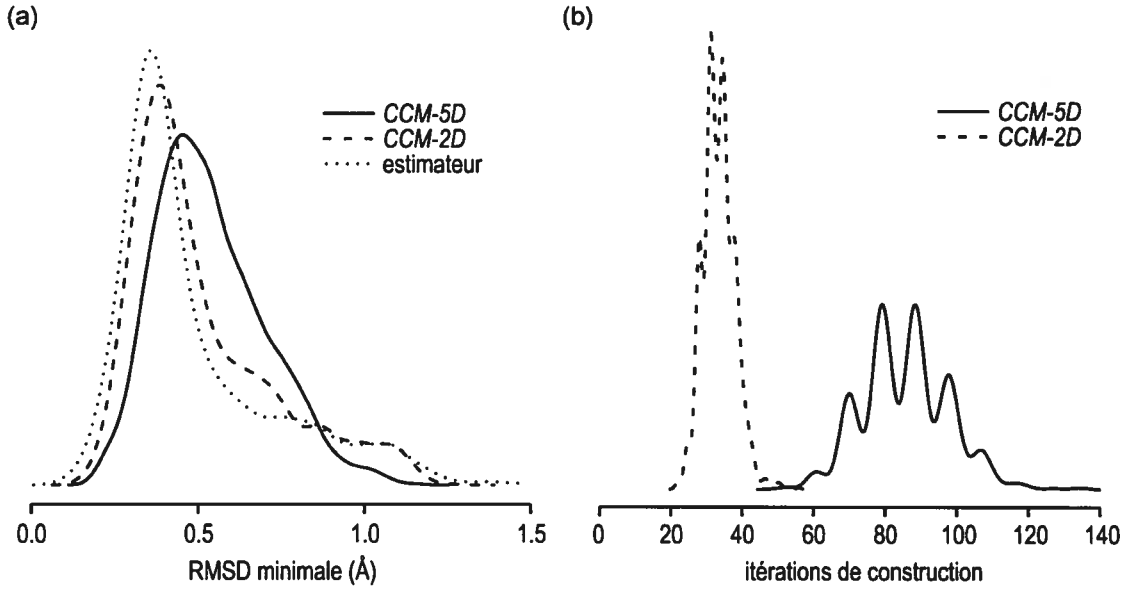


Figure 35 : Comparaison des trois différentes méthodes de détermination des paramètres de la construction du ribose sur l'épreuve de reconstruction des riboses des modèles de référence. (a) Courbes de densité empirique pour la précision de la construction évaluée selon la déviation RMS minimale entre chaque ribose construit et l'ensemble des modèles originaux. (b) Courbes de densité empirique pour la quantité d'itérations de construction nécessaire à la détermination des paramètres de construction pour chaque modèle.

3.5 Conclusion

Dans la méthode de construction artificielle d'un modèle de ribose d'ARN que nous avons présenté dans ce chapitre, deux sous-procédures se distinguent : la construction paramétrée hors-contexte et la détermination des paramètres de construction en contexte. À la section 3.3, nous avons décrit en détails notre procédure de construction atomique paramétrée en $< \rho, \chi, \gamma, \beta, \varepsilon >$ et nous avons évalué quantitativement ses performances hors du contexte de la modélisation : précision géométrique, représentativité expérimentale et efficacité calculatoire. Ainsi, nous avons démontré que les riboses construits sont représentatifs à 100%, à un seuil de 1Å, des modèles de référence. D'un autre côté, 25% des riboses construits sont mal représentés (plus de 1Å), cependant cette divergence est entièrement due aux branches libres vers les phosphates, puisque la représentativité bidirectionnelle est à 100%, à 1Å près, lorsque seul le cycle furanosique est considéré. Il est donc raisonnable de

postuler que la liberté géométrique des branches vers les phosphates, soient les atomes O5', P5' et P3', paramétrée par $< \gamma, \beta, \varepsilon >$, est plutôt restreinte dans les modèles de référence.

Ainsi, notre procédure de construction hors-contexte du ribose est validée. Ensuite vient le problème de la détermination des paramètres de la construction qui mènent à un modèle précis. À la section 3.4, nous avons décrit trois méthodes différentes de détermination des paramètres : CCM-5D, CCM-2D et estimation, respectivement une optimisation numérique linéaire des paramètres $< \rho, \chi, \gamma, \beta, \varepsilon >$, une même optimisation réduite aux couple $< \rho, \chi >$ et une estimation directe du couple $< \rho, \chi >$. Côté précision, CCM-5D construit le moins de modèles divergents de l'ensemble de référence, comme en témoigne la figure 35a. Cependant, l'écart de précision avec les autres méthodes est faible, d'autant plus si la valeur moyenne de déviation RMS est considérée : 0.54Å pour CCM-5D, 0.51Å pour CCM-2D et 0.50Å pour l'estimateur. Le problème est que, pour arriver à cette précision, CCM-5D et CCM-2D nécessitent l'application de plusieurs itérations de construction, respectivement en moyenne 86 et 34 itérations, alors que l'estimateur utilise toujours une quantité constante de deux itérations. Bref, si la quantité d'itérations nécessaires à la détermination des paramètres de la construction a autant de poids que la précision, il est évident que la méthode par estimation est préférable aux méthodes d'optimisation numérique linéaire.

Par ailleurs, il est intéressant de comparer l'erreur de détermination des paramètres lors de la reconstruction des modèles de référence par les trois méthodes. Pour ce faire, nous avons rapporté à la figure 36 la différence entre les paramètres ρ et χ mesurés avant et après la reconstruction de chaque ribose. Dans le cas des deux méthodes d'optimisation, i.e. CCM-5D et CCM-2D, les courbes d'erreur sont parfaitement confondues puisque le cycle furanosique, paramétré par ρ et χ, est complètement déterminé de la même façon par ces deux méthodes. À la figure 36, nos résultats mettent en évidence que l'estimation reproduit mieux les mesures des modèles de référence que ses homologues par optimisation, surtout pour la détermination de ρ où l'optimisation distribue l'erreur entre 0° et 180° de façon quasi uniforme. De plus, la méthode par estimation présente une erreur sur la détermination de ρ inférieure à 36° – la longueur d'un intervalle de ρ correspondant à un mode de *puckering* particulier (voir figure 25) – dans 77% des cas. Bien qu'aucune des trois méthodes ne permettent explicitement de restreindre la détermination des paramètres à un ou plusieurs modes de *puckering* en particulier, il est intéressant de voir que notre méthode par estimation assure à un certain niveau que si les groupements phosphates sont positionnés relativement à la base de façon à

permettre un mode de *puckering* en particulier, du moins tel qu'observé dans l'ensemble de référence, alors l'estimation déterminera les paramètres qui mènent à ce mode de *puckering*.

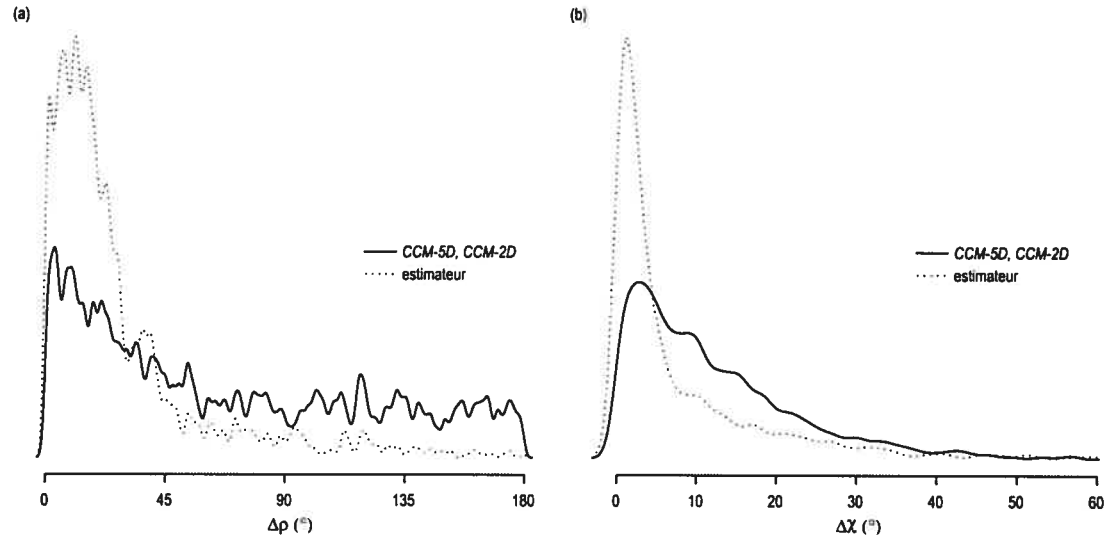


Figure 36 : Erreur sur la détermination de ρ (a) et de χ (b). Les courbes représentent les estimations de densité empirique. Les courbes pour les méthodes CCM-5D et CCM-2D sont parfaitement confondues.

Finalement, nous avons considéré notre méthode de construction artificielle d'un ribose dans un contexte de modélisation où les bases azotées et les phosphates sont positionnés dans l'espace tridimensionnel de façon individuelle et relative. Dans ce contexte, les relations d'appariement entre les bases et les phosphates sont connues et le problème est de créer un ribose entre une base et ses deux phosphates directement adjacents, problème évidemment résolu par les méthodes décrites ici. Cependant, il faut se pencher sur une évaluation quantitative du modèle du ribose construit dans ce contexte de modélisation où il n'y a pas de référence. Théoriquement, nos méthodes de détermination des paramètres de construction trouvent toujours une solution, il faut donc décider de la validité du ribose construit. Pour ce faire, nous avons choisi la mesure de qualité du ribose construit, telle que décrite à la sous-section 3.4.1, c'est-à-dire la fonction objectif de la méthode de détermination des paramètres de la construction. Cette mesure de qualité se calcule par la somme de la mesure des liens covalents implicites entre le cycle furanosique construit et les phosphates du contexte. Ainsi, un seuil sur cette mesure de qualité permet d'accepter ou de rejeter une construction de ribose en contexte. À la figure 37, nous avons rapporté les mesures de qualité lors de la reconstruction de l'ensemble de référence. Encore ici, les courbes pour CCM-5D et CCM-2D sont confondues puisque la mesure de qualité ne dépend que de ρ et χ. Un seuil de 4Å accepte près de 100% des riboses reconstruits par optimisation et 96% par estimation. Par ces

résultats, nous proposons donc un seuil minimal de 4Å sur la qualité du ribose construit dans le contexte d'une modélisation de la structure complète d'un ARN.

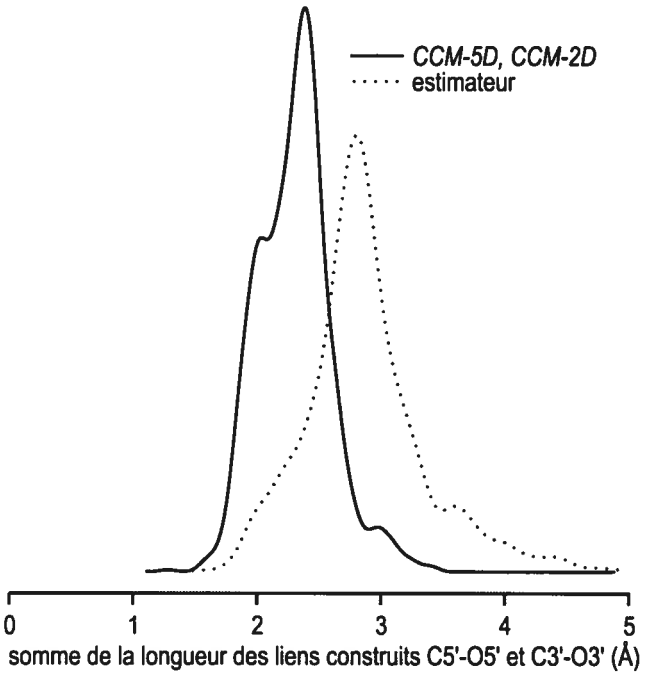


Figure 37 : Mesure de la qualité des modèles de ribose reconstruits depuis l'ensemble de référence. La mesure est l'erreur normalisée sur les liens covalents implicite entre les atomes C5' et O5' ainsi que C3' et O3'. Les courbes sont les évaluations de densité empirique.

Chapitre 4

Modélisation des ARN par satisfaction de contraintes

4.1 Introduction

Au cours des chapitres précédents, nous avons mis l'accent sur la méthode de construction d'un modèle d'ARN dans le cadre du système de modélisation *MC-Sym*. Nous avons montré comment positionner les bases azotées dans l'espace tridimensionnel par relations d'appariement et compléter le modèle en y ajoutant artificiellement le squelette. Plus précisément, au chapitre 2, nous avons mentionné l'utilisation d'un algorithme pour explorer systématiquement l'espace conformationnel formé par l'union de toutes les relations d'appariement connues qui s'appliquent à une modélisation en particulier. Ce processus d'exploration de l'espace conformationnel se formalise dans le cadre théorique d'un problème de satisfaction de contraintes [27]. Ce formalisme permet une définition claire du problème de modélisation et un développement efficace des algorithmes de résolution. C'est exactement ce que nous étudions dans ce chapitre. Dans les prochaines sections, le propos se divise comme suit : D'abord, nous proposons une définition théorique du problème de satisfaction de contraintes, avec son application au problème particulier de la modélisation des ARN. Ensuite, nous présentons l'algorithme classique de résolution par retour arrière (*backtrack*), en critiquant son efficacité et en discutant de ses possibles optimisations, notamment en utilisant l'algorithmique probabiliste. De cette section, nous sortirons des variantes d'algorithmes probabilistes de résolution qui seront prototypés en utilisant le problème des n reines. Enfin, nous mettrons à l'épreuve nos algorithmes de résolution sur la modélisation de l'anticodon d'un ARN de transfert d'abord, puis d'une partie du domaine I d'un intron du groupe II. Nous terminerons par une discussion des résultats obtenus.

4.2 Problème de satisfaction de contraintes

Pour décrire le problème de modélisation des ARN à l'intérieur du cadre général d'un problème de satisfaction de contraintes, il faut d'abord décrire le cadre théorique lui-même. C'est le rôle de cette section, dont le formalisme est inspiré de [25].

4.2.1 Définition d'un problème de satisfaction de contraintes

Un *problème de satisfaction de contraintes* (l'abréviation CSP sera utilisée dans le reste du texte, se référant à la nomination anglaise *constraint satisfaction problem*) se définit d'abord avec un ensemble de **variables** $V = \{v_1, v_2, \ldots, v_n\}$ où chaque variable v_i possède un **domaine** fini de valeurs d_i dans $D = \{d_1, d_2, \ldots, d_n\}$. Une assignation à la variable v_i d'une valeur de son domaine d_i est une **instanciation** de cette variable. Une instanciation de toutes les variables de V est une **solution** du CSP. La conjugaison de l'ensemble V avec l'ensemble D établit l'**espace conformationnel** du CSP. L'exploration de cet espace par l'instanciation progressive des variables de V génère un ensemble de solutions.

Évidemment, comme l'appellation du problème l'indique clairement, il faut satisfaire aux contraintes imposées par le problème pour qu'une solution explorée soit valide. Soit $C = \{c_1, c_2, \ldots, c_n\}$ l'ensemble des **contraintes.** Une contrainte c_i démontre une **portée**, c'est-à-dire qu'elle s'adresse à un sous-ensemble des variables dans V. L'**arité** d'une contrainte est la quantité de variables adressées par celle-ci. Ainsi, une contrainte k-aire est un sous-ensemble du produit cartésien des k domaines associés aux k variables de la portée de la contrainte. Chaque k-tuplet du sous-ensemble désigne une instanciation valide pour toutes les k variables de la portée. Chaque contrainte est donc une relation k-aire, au sens propre du terme [26]. Par exemple, une contrainte unaire spécifie les valeurs valides pour le domaine de la variable concernée, alors qu'une contrainte binaire spécifie les couples de valeurs valides provenant des deux variables concernées, *etc.* Avec les ensembles V, D et C, le CSP est complètement définit.

Finalement, résoudre un CSP, c'est trouver une (ou plusieurs) instanciation complète des n variables de sorte que toutes les contraintes soient satisfaites, i.e. validées. La méthode classique utilise un algorithme de retour arrière (*backtrack*) pour explorer de façon

exhaustive l'espace conformationnel, une instanciation à la fois. Même si la validation systématique des solutions partielles par la vérification des contraintes dès que leur portée est instanciée émonde une grande partie de l'espace de recherche, il n'en reste pas moins que la complexité de la méthode est exponentielle en terme de quantité de variables (n).

4.2.2 Application à la modélisation des ARN

Maintenant, il s'agit de définir le problème de la modélisation discrète des ARN comme un CSP. Cet exercice a déjà été réalisé pour *MC-Sym* [27] [28], nous voulons apporter ici une nouvelle définition qui utilise le formalisme décrit ci-haut. Pour ce faire, il suffit de définir les ensembles V (variables), D (domaines) et C (contraintes) dans le contexte d'application.

4.2.2.1 Graphe de relation et ensemble des domaines

Grâce à la représentation de la structure de l'ARN à modéliser par graphe de relation (voir chapitre 1), les ensembles représentatifs d'un CSP se définissent directement. L'ensemble des arêtes de l'arbre de recouvrement qui spécifie l'ordre de construction définit l'ensemble V des variables du CSP. Le domaine d_i de la variable v_i est l'ensemble des matrices de transformation décrivant la relation d'appariement représentée par v_i. Ainsi, lorsque le graphe de relation est annoté selon la syntaxe des requêtes à la base de relations de MC-Sym, les ensembles résultats qui sont associées aux arêtes du graphe définissent l'ensemble D des domaines.

4.2.2.2 Contraintes

Les variables du CSP et leurs domaines étant définis, il ne reste que l'ensemble C des contraintes. Les contraintes de validation émergentes de la méthode de construction par relation d'appariement ont déjà été décrites au chapitre 2, cependant il faut les intégrer au formalisme du CSP. Ici, l'ensemble des contraintes se veut une méthode de validation d'une instanciation complète donnée. Une instanciation des n variables est un modèle moléculaire tridimensionnel de l'ARN en question où toutes les bases azotées ont été placées les unes par rapport aux autres. Quant au squelette (*backbone*) de l'ARN, il est construit progressivement

et artificiellement en ajoutant des riboses aux bases positionnées d'une manière décrite en détails aux chapitres 2 et 3. Bref, lorsqu'une variable du CSP est instanciée, une nouvelle base est ajoutée au modèle en construction, et aussi certains riboses. Les bases et les riboses sont ici les éléments constitutifs du modèle.

Pour être valide, ce modèle doit, dans un premier lieu, être exempt de toute collision stérique au niveau atomique. Cette contrainte de collision est aisément vérifiée sur le modèle, ou sur une partie du modèle, en fixant un seuil minimal sur la distance entre chaque paire d'atomes placés. Plus précisément, il existe une contrainte de collision pour chaque instanciation entre la base nouvellement placée par cette instanciation, ainsi que ces groupements ribose associés, et tous les autres éléments – bases et riboses – déjà placés. Chaque élément nouvellement placé implique donc autant de contraintes de collision qu'il y a d'éléments dans le modèle partiellement instancié. Par exemple, soit l'instanciation de la $k^{\text{ième}}$ variable, dans l'ordre du parcours préfixe de l'arbre de recouvrement. Supposons que l'instanciation de cette variable ajoute, en plus d'une nouvelle base, r_k nouveaux éléments ribose. Le modèle partiel contient donc $\sum_{i=1}^{k-1}(1+r_i)$ éléments, où $r_1 = 0$. Donc, $(1+r_k)\sum_{i=1}^{k-1}(1+r_i)$ contraintes de collisions sont nécessaires à l'instanciation de la k^e variable, chacune étant k-aire. En effet, puisque les bases sont positionnées les unes par rapport aux autres, la position globale d'une base ne dépend pas uniquement de l'instanciation de sa variable associée, i.e. de la matrice de transformation qu'il l'a placée dans le référentiel local de l'autre base de la relation, mais bien de l'instanciation de toutes les autres variables précédentes dans le parcours préfixe.

De par l'approximation de la structure du squelette d'après le positionnement des bases azotées, un deuxième type de contrainte est nécessaire à la validation du modèle atomique : la qualité du ribose construit. Cette métrique quantifie la précision et la représentativité avec lesquelles le ribose a été ajouté lors de la construction d'une partie du modèle (voir chapitre 3). À la k^e instanciation, r_k contraintes de qualité du ribose sont nécessaires, chacune étant k-aire pour la même raison que précédemment. La quantité totale de contraintes nécessaires au CSP est donc déduite de la façon suivante :

$$|C| = \sum_{k=1}^{n}\left(r_k + (1+r_k)\sum_{i=1}^{k-1}(1+r_i)\right)$$

La taille de l'ensemble C est en grande partie dépendante de r_k, la quantité d'éléments riboses ajoutée à la $k^{\text{ième}}$ instanciation, quantité pratiquement imprévisible puisqu'elle est

totalement dépendante à la fois de la structure elle-même de l'ARN à modéliser et du choix de l'arbre de recouvrement de son graphe de relation. Par contre, cette quantité est bornée à 3 riboses (voir chapitre 2), ce qui conserve la complexité quadratique de $|C|$ en fonction de n. D'autres contraintes plus spécifiques peuvent être utiles à la résolution d'un tel CSP, cependant les contraintes de collision et de qualité du ribose sont à elles seules suffisantes à la validation du modèle construit.

Ici, il faut remarquer que dans l'application à la modélisation des ARN, une contrainte n'est pas réellement une relation mathématique définit sur les domaines des variables concernées par la portée, mais plutôt un processus d'observation *a posteriori* de l'instanciation de ces variables.

4.3 Algorithmes de résolution

D'après la définition théorique déclarée à la sous-section précédente, un algorithme de résolution d'un CSP est une méthode systématique de génération d'instanciations complètes et valides des n variables. Cette méthode de résolution passe par l'exploration de l'espace conformationnel formé par la conjugaison des ensembles des domaines des n variables. Du point de vue de l'application à la modélisation des ARN, une instanciation complète et valide est un modèle complet de l'ARN. La conjugaison des domaines des variables donne une série de modèles intermédiaires qui doivent être validés par les contraintes de collision et de qualité de ribose. De cette façon, l'exploration de l'espace conformationnel devient réellement une exploration discrète dans l'espace tridimensionnel des différentes conformations adoptables par le modèle moléculaire selon les diverses assignations de relations d'appariement entre les bases azotées. Dans cette section, nous nous penchons sur l'algorithmique nécessaire à ce processus de résolution.

4.3.1 Retour arrière et arbre implicite d'exploration

Comme il a été mentionné précédemment, l'algorithme de retour arrière classique est la méthode simple et directe de résolution de ce problème (voir [31] pour une définition complète d'un algorithme de retour arrière). En effet, l'idée est d'instancier les variables, i.e.

les relations d'appariement, de façon successive dans l'ordre de l'arbre de recouvrement - dont la racine est connue - du graphe de relation de la structure de l'ARN, l'épuisement d'un domaine entraînant un retour arrière à la variable précédente. Virtuellement, cette méthode crée un arbre implicite de recherche où chaque niveau représente l'instanciation d'une variable (voir figure 38). Un chemin depuis la racine jusqu'à une feuille représente donc une instanciation complète. L'algorithme de retour arrière parcours tous les nœuds de cet arbre d'une manière dite en profondeur d'abord. Cette représentation graphique permet de bien saisir à la fois la taille exponentielle de l'espace conformationnel et l'optimisation apportée par la validation hâtive des contraintes, c'est-à-dire dès que la portée de la contrainte est instanciée. En effet, la taille maximale de l'arbre implicite d'exploration est $1+\sum_{i=1}^{n}\prod_{j=1}^{i}\left|d_j\right|$, cependant l'émondage résultant de la validation hâtive des contraintes réduit la taille effective. Cette taille effective est difficile à calculer *a priori*, puisqu'elle dépend de la satisfaction de contraintes spatiales observées. La satisfaction de ces contraintes ne peut donc être prédite qu'au moment de l'instanciation elle-même.

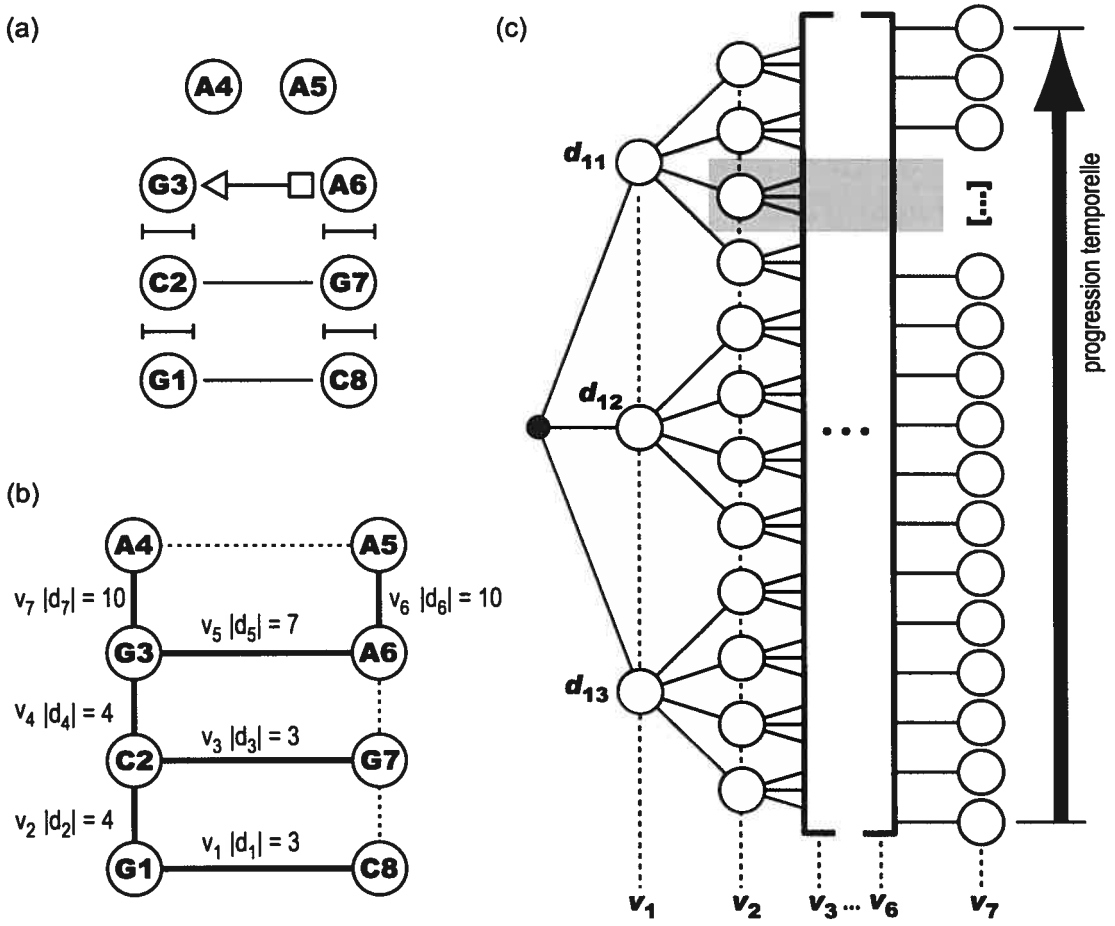


Figure 38 : Arbre implicite de recherche décrivant l'application de l'algorithme de retour arrière pour la modélisation d'une structure tige-boucle. (a) Structure secondaire de la tige-boucle selon la nomenclature de Leontis & Westhof [30]. Une tige formée de deux paires de bases est fermée par une boucle GAAAÀ qui contient un appariement non-canonique. (b) Graphe de relation de la structure. L'arbre de recouvrement spécifiant l'ordre de construction est représenté en trait plein, les arêtes en trait pointillé sont donc implicitement représentées. Les arêtes sont numérotées selon leur variable associée, et la taille de chaque ensemble de domaine est indiquée. (c) L'arbre implicite de recherche. Chaque niveau de l'arbre représente l'application d'une relation d'appariement, dans l'ordre du parcours préfixe de l'arbre de recouvrement. Les feuilles de l'arbre sont l'aboutissement d'une instanciation complète. La validation hâtive des contraintes permet de réduire la taille effective de cet arbre. Par exemple, si le chemin d_{11}-d_{23} invalide une contrainte de collision entre les variables v_1 et v_2, tout le sous-arbre dont la racine se situe à la fin de ce chemin, illustré par la zone ombrée, ne sera pas parcouru; tout cet espace sera émondé. La flèche de progression indique l'ordre de sortie des solutions du problème.

L'utilisation du retour arrière comme d'une méthode de résolution d'un CSP présente l'avantage marqué d'assurer une exploration exhaustive de l'espace conformationnel : toutes les solutions possibles selon les domaines choisis et les contraintes apposées sont atteintes. La figure 39 montre le pseudocode de l'algorithme de retour arrière utilisé ici.

```
    procédure RETOURARRIÈRE
    entrée : V = {v_1 ... v_n}, D = {d_1 ... d_n}, C
1.      i ← 1                                   (niveau courant)
2.      J = {j_1 ... j_n} ← {1,...,1}           (itérateurs de domaine)
3.      tant que i > 0 faire
4.      début
5.        si j_i > |d_i| alors                  (fin du domaine atteinte)
6.          j_i ← 1
7.          i ← i − 1                           (retour arrière)
8.        sinon
9.          v_i ← d_{i,j_i}                     ( j_i^e élément de d_i )
10.         j_i ← j_i + 1
11.         si VALIDE({v_1 ... v_i}, C) alors
12.           si i = n alors
13.             afficher la solution {v_1 ... v_n}
14.           sinon
15.             i ← i + 1                       (progression)
16.     fin
    fin de la procédure
```

Figure 39 : Pseudocode pour RetourArrière, l'algorithme de résolution classique d'un CSP. Les ensembles *V*, *D*, et *C* sont les ensembles caractéristiques du CSP, soient respectivement les variables, domaines et contraintes. La sous-procédure VALIDE(*X*,*C*) retourne vrai si et seulement si l'instanciation partielle $X = \{ v_1 ... v_i \}$ satisfait à toutes les contraintes dans *C*.

4.3.2 Failles dans l'algorithme de retour arrière

Bien que le retour arrière soit une méthode simple de résoudre un CSP de façon exhaustive, la complexité exponentielle de la taille de l'espace conformationnel, et parallèlement de celle de l'arbre implicite de recherche, rend la tâche difficilement praticable et très avare en ressources matérielles. La taille de l'espace conformationnel, $\prod_{k=1}^{n} |d_k|$, s'accroît dans deux dimensions : la taille des domaines et la quantité de variables. La taille des domaines représente les différentes façons d'appliquer une relation d'appariement localement, alors que la quantité de variables représente directement la longueur de la séquence d'ARN à

modéliser. Sont donc restreints à la fois la taille de l'ARN modélisable et l'espace tridimensionnel couvert par les différentes conformations adoptables selon la stéréochimie de la structure. Le premier point est certes majeur, puisqu'il réduit directement le champ d'application du système de modélisation, mais le second point a quant à lui un impact plus grave au point de vue de la complétude. En effet, la réduction de la taille des domaines peut mener à des culs-de-sac conformationnels où certaines solutions deviendraient inatteignables de par le retrait de certaines représentations de relations d'appariement.

D'un point de vue pratique, nous observons que le retour arrière manifeste certains comportements problématiques. D'abord, l'unidirectionnalité du parcours en fait une méthode d'exploration de l'espace conformationnel un peu naïve. Bien que tout à fait exhaustive, l'exploration a tendance à parcourir souvent les mêmes régions [25], valides ou non. Le second cas échéant, le temps de recherche est évidemment gaspillé. La figure 40 illustre deux conséquences hasardeuses découlant directement de cet aspect. La première suppose l'hypothèse suivante : l'instanciation $v_j = x$, au niveau *j*, invalide une contrainte dont la portée contient aussi la variable v_k, $k > j$. Cette contrainte sera donc évaluée seulement à l'instanciation de v_k, $k - j$ niveaux plus bas. Dans le pire des cas, l'instanciation partielle $\{v_1 \ldots v_j \ldots v_{k-1}\}$ avec $v_j = x$ sera toujours cohérente avec les contraintes, par conséquent un sous-arbre de $k - j$ niveaux devra être vainement parcouru par l'algorithme de retour arrière avant de pouvoir modifier v_j et régler le conflit avec v_k. Dans le cas de l'application à la modélisation des ARN, cette hypothèse est très souvent vérifiée : il suffit que la position d'une base azotée entre en collision avec une base positionnée à quelques niveaux précédents pour que toutes les positions de toutes les bases des niveaux intermédiaires soient vainement essayées avant de pouvoir modifier la position de la base qui cause réellement la collision. Pire, la taille d'un tel sous-arbre vainement parcouru croît de façon exponentielle avec la taille des domaines des variables de l'instanciation partielle. L'autre conséquence fâcheuse illustrée à la figure 40 tient plutôt de l'orientation de la recherche en profondeur d'abord. Lors de l'exécution du retour arrière, la génération temporelle de solutions au CSP s'effectue par blocs où les solutions y sont très similaires entre elles, ayant comme distinctions seulement les quelques derniers niveaux instanciés. En fait, pour atteindre une solution dont *k* variables diffèrent entre elles depuis une solution atteinte donnée, un sous-arbre de *k* niveaux doit être complètement parcouru. Dans le contexte de la modélisation des ARN, deux modèles dont seulement une petite proportion

relative des bases azotées ont été positionnées différemment restent plutôt similaires de façon globale. Une grande portion relative de la progression de la recherche devra donc s'écouler avant de s'éloigner d'une région de l'espace conformationnel vers une autre plus distancée, et obtenir des modèles distinctifs. Ainsi, la diversité du groupe de solutions générées progresse lentement.

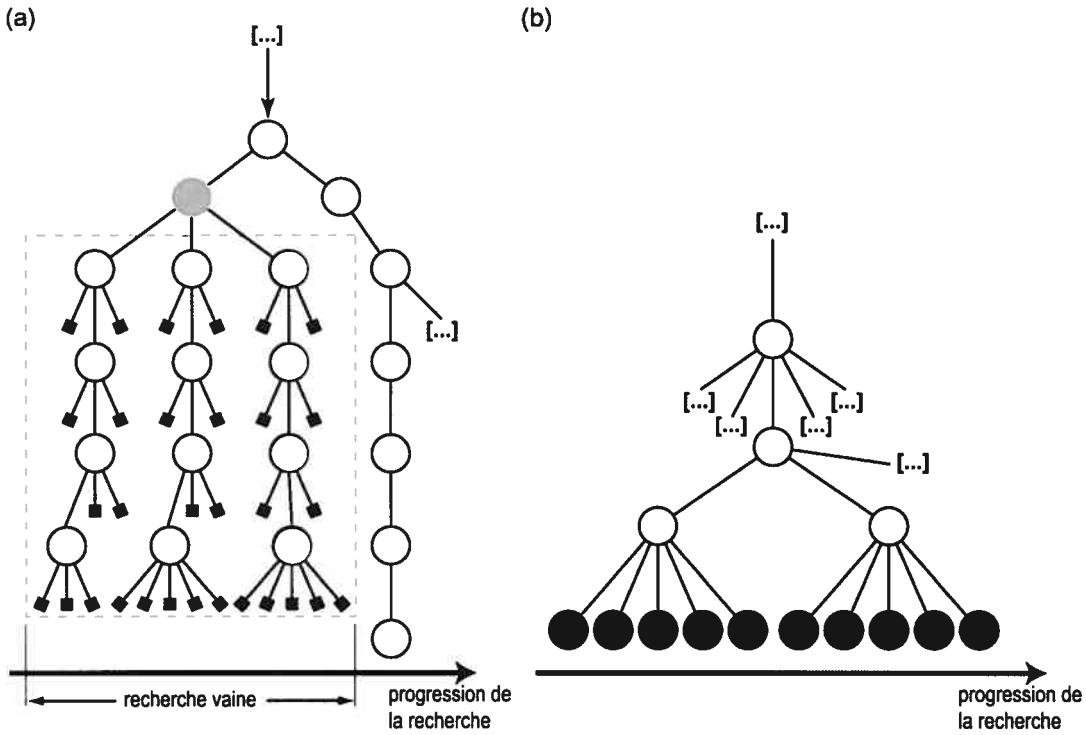


Figure 40 : Conséquences hasardeuses découlant directement de l'utilisation de la méthode de résolution par retour arrière. (a) Illustration d'un sous-arbre parcouru vainement. Ici est supposé que l'instanciation ombrée fait en sorte que la variable située quatre niveaux plus bas ne puisse s'instancier de façon valide. Le résultat est qu'un sous-arbre de quatre niveaux est parcouru vainement avant de pouvoir modifier l'instanciation réellement coupable; tout le temps de recherche nécessaire à ce parcours étant gaspillé. (b) Illustration du temps de recherche nécessaire à la génération de solutions différentes. Les dix solutions successivement générées diffèrent entre elles seulement selon l'instanciation des deux dernières variables. De plus, pour passer d'une solution à une autre dont les deux dernières variables sont toutes deux différentes entre elles, tout le premier bloc de cinq solutions doit être généré.

4.3.3 Optimisation de l'algorithme de retour arrière classique

Pour s'attaquer au problème des parcours vains de sous-arbres, la solution la plus directe est de permettre à l'algorithme de résolution d'exécuter des sauts arrières de plusieurs niveaux,

pour permettre d'éviter le parcours fastidieux de régions invalides en remontant directement au niveau conflictuel, en supposant que ce dernier ait été identifié hors de tout doute. Cette technique de saut arrière (*backjump*) entre dans le cadre plus général de l'amélioration dynamique de la puissance d'émondage du retour arrière. L'idée derrière la technique de saut arrière est d'analyser la cause de l'invalidation d'un niveau lors du parcours de l'arbre de recherche et de sauter directement à ce niveau, au lieu de retourner au niveau directement précédent, pour le corriger immédiatement. Il faut cependant s'assurer que le sous-arbre dynamiquement émondé ne contenait pas de solutions, i.e. de feuilles de l'arbre complet. L'enjeu ici est la conservation de l'exhaustivité de l'algorithme de résolution, de sa capacité à trouver toutes les solutions logiquement possibles du CSP tel que défini. Cette notion de saut arrière sécuritaire est donc primordiale. Pour y arriver, il faut vérifier que l'instanciation partielle $\{v_1 \ldots v_j\}$, où v_j est le niveau cible du saut arrière, ne fasse partie d'aucune solution. En effet, si $\{v_1 \ldots v_j\}$ ne peut s'étendre à aucune solution, la modification hâtive de v_j, sans parcourir l'entièreté du sous-arbre dont la racine se situe à l'instanciation v_j, n'encourra la perte d'aucune solution. Malheureusement, dans le cas de l'application à la modélisation des ARN tel que défini ici, il n'existe pas de moyen de vérifier si $\{v_1 \ldots v_j\}$ peut ou non s'étendre à une solution autrement qu'en parcourant le sous-arbre sous-jacent. Une telle vérification possède donc la même complexité que le retour arrière pur et simple; elle est donc obsolète. Même en supposant l'existence d'une méthode optimisée pour vérifier qu'un saut arrière est sécuritaire, il reste que l'identification du niveau coupable, autre que le niveau directement précédent à celui qui s'est épuisé, n'est pas possible dans ce contexte. En effet, pour déterminer le niveau coupable d'un épuisement au niveau *k*, il faut trouver l'instanciation partielle de taille minimale telle que celle-ci invalide toute instanciation pour v_k [25]. Or, pour ce faire, il faut arriver à valider une instanciation partielle $\{v_1 \ldots v_j\} \cup \{v_k\}$, où $j < k$, ce qui n'est pas possible ici puisque les variables, c'est-à-dire les relations d'appariement, sont totalement interdépendantes par rapport à la satisfaction des contraintes. L'ordre de l'instanciation des variables doit être respecté. Bref, puisqu'il est à la fois impossible de déterminer si un niveau en particulier est coupable d'une invalidation et si un saut arrière à un niveau quelconque est sécuritaire, la technique de saut arrière ne s'applique pas dans le cadre d'une recherche exhaustive comme le retour arrière.

C'est pourquoi nous nous sommes tourné vers les heuristiques probabilistes. Sachant qu'une solution se retrace par un chemin dans l'arbre implicite de recherche depuis la racine jusqu'à

la feuille, l'idée est de générer des chemins de façon aléatoire en validant chaque nœud jusqu'à la feuille. Pour ce faire, nous utilisons un algorithme probabiliste de type Las Vegas [31]. Chaque variable est instanciée dans le même ordre qu'avec le retour arrière, seulement chaque instanciation est aléatoire. À chaque instanciation, les contraintes applicables sont vérifiées : un succès fait progresser l'instanciation à la prochaine variable, alors qu'un échec est irréversible et redémarre la recherche à la première variable. Cette approche dite purement probabiliste, un peu naïve, a piètrement performé lors d'expériences préliminaires qui ne sont pas publiées ici. Nous avons donc immédiatement rejeté cette première approche. Par contre, d'une approche inspirée de Brassard et Bratley [31], nous avons développé une variante mi-probabiliste mi-déterministe du retour arrière. L'idée est de ne pas capituler trop vite lors d'un échec et de tenter de corriger la situation en démarrant un retour arrière de taille fixe à partir du niveau en conflit avec les contraintes. À la figure 41, nous dévoilons le pseudocode de cet algorithme probabiliste de résolution d'un CSP. La taille fixe de la phase de correction par retour arrière est paramétrée en deux dimensions : s, la quantité de réessais d'instanciation aléatoire d'un niveau en échec avant d'exécuter un retour arrière, et t, le nombre maximal de ces retours arrière qu'il est possible d'exécuter avant de capituler et de redémarrer la recherche au tout début.

```
    procédure RETOURARRIÈRELV
    entrée : V = {v_1 ... v_n}, D = {d_1 ... d_n}, C, s, t
1.      D' ← D                                  (copie des domaines)
2.      i ← 1                                   (niveau courant)
3.      s' ← 0                                  (compteur réessai)
4.      t' ← 0                                  (compteur retour arrière)
5.      tant que vrai faire
6.      début
7.        si s' ≥ s ou d'_i = { } alors         (épuisement du niveau)
8.          s' ← 0
9.          t' ← t' + 1
10.         si t' ≥ t alors                     (échec total)
11.           REDÉMARRE()
12.         sinon
13.           d'_i ← d_i
14.           si i > 1 alors
15.             i ← i − 1                       (retour arrière)
16.       sinon
17.         choisir un élément aléatoire x ∈ d'_i
18.         v_i ← x
19.         d'_i ← d'_i \ {x}
20.         si VALIDE({v_1 ... v_i}, C) alors
21.           s' ← 0
22.           si i = n alors
23.             afficher la solution {v_1 ... v_n}
24.             REDÉMARRE()
25.           sinon
26.             i ← i + 1                       (progression)
27.         sinon
28.           s' ← s' + 1                       (réessai du niveau)
29.     fin
    fin de la procédure

    sous-procédure REDÉMARRE
30.     tant que i > 0 faire
31.       d'_i ← d_i
32.       i ← i − 1
33.     i ← 1
34.     t' ← 0
    fin de la sous-procédure
```

Figure 41 : Pseudocode pour RETOURARRIÈRELV, l'algorithme de résolution probabiliste d'un CSP avec correction par retour arrière de taille fixe. Les paramètres s et t forment la taille fixe de la correction par retour arrière, soient respectivement la quantité de réessais permis à chaque niveau et le nombre maximal de niveaux atteignables par retour arrière.

Puisque cet algorithme est une heuristique probabiliste, l'exhaustivité de l'exploration de l'espace conformationnel est sacrifiée. De plus, puisque le parcours de l'arbre implicite de recherche est aléatoire, rien n'empêche le même chemin d'apparaître plus d'une fois. Un mécanisme de filtrage de solutions identiques doit donc être implanté pour éviter la redondance dans l'ensemble des solutions. Heureusement, la probabilité qu'exactement le même chemin soit parcouru plus d'une fois s'amenuise à mesure que grandit l'espace conformationnel. D'ailleurs, cette descente aléatoire dans l'arbre de recherche est manifestement plus apte à générer des solutions plutôt différentes qu'identiques en se fiant au chaos aléatoire, d'autant plus que lorsqu'une solution est atteinte, la recherche est redémarrée au tout début (ligne 26). Voilà qui règle la deuxième conséquence hasardeuse à l'utilisation du retour arrière classique (voir figure 40) et qui devrait assurer un bon développement de la diversité dans la génération temporelle des solutions. Pour ce qui est du problème du parcours vain de sous-arbres localement cohérents mais tout de même absents de toute solution, nous avons expliqué qu'il n'était pas possible d'identifier sécuritairement le niveau coupable d'une invalidation dans le contexte d'une recherche exhaustive. Cependant, dans le cadre d'une heuristique probabiliste non exhaustive, la technique du saut arrière peut être appliquée sans dénaturer l'algorithme. Par exemple, lorsqu'une instanciation particulière échoue à la validation des contraintes (ligne 29), le niveau en conflit est conservé dans une liste, cette dernière s'allongeant au fur et à mesure que le domaine de la variable courante s'épuise. Lorsque le domaine est finalement épuisé et que la recherche doit retourner en arrière (ligne 17), un saut arrière est plutôt exécuté jusqu'à un niveau sélectionné dans la liste susmentionnée. Tout réside dans la sélection du niveau coupable, menant à différentes variantes de saut arrière pour l'algorithme. À la figure 42, nous détaillons le pseucocode de notre version avec saut arrière, où la méthode de sélection du coupable est laissée libre.

Cependant, plus d'une contrainte peut invalider une seule instanciation, conséquemment la liste des coupables peuvent devenir rapidement très longue et rendre la sélection du coupable difficile. Nous avons utilisé un exemple de CSP plus simple que la modélisation tridimensionnelle des ARN comme prototype pour comparer ces différents algorithmes de résolution : le problème des n reines. À la section suivante, nous décrivons ce prototype et présentons les résultats de la comparaison des algorithmes de résolution. L'algorithme qui se démarquera de cette comparaison sera transposé dans le système *MC-Sym*.

```
     procédure RETOURARRIÈRELV_SA
     entrée V = {v_1 ... v_n}, D = {d_1 ... d_n}, C, s, t
1.     D' ← D                                   (copie des domaines)
2.     i ← 1                                    (niveau courant)
3.     s' ← 0                                   (compteur réessai)
4.     t' ← 0                                   (compteur retour arrière)
5.     P ← { }                                  (liste des coupables)
6.     tant que vrai faire
7.     début
8.       si s' ≥ s ou d'_i = { } alors          (épuisement du niveau)
9.         s' ← 0
10.        t' ← t' + 1
11.        si t' ≥ t alors                      (échec total)
12.          REDÉMARRE()
13.        sinon
14.          si i > 1 alors
15.            i' ← SÉLECTIONCOUPABLE(P)
16.            tant que i > i' faire            (saut arrière à i')
17.              d'_i ← d_i
18.              i ← i − 1
19.          sinon
20.            d'_i ← d_i
21.      sinon
22.        choisir un élément aléatoire x ∈ d'_i
23.        v_i ← x
24.        d'_i ← d'_i \ {x}
25.        si VALIDE({v_1 ... v_i}, C, p) alors
26.          s' ← 0
27.          P ← { }
28.          si i = n alors
29.            afficher la solution {v_1 ... v_n}
30.            REDÉMARRE()
31.          sinon
32.            i ← i + 1                        (progression)
33.        sinon
34.          P ← P ∪ {p}
35.          s' ← s' + 1                        (réessai du niveau)
36.    fin
     fin de la procédure
```

Figure 42 : Pseudocode pour RETOURARRIÈRELV_SA, une version généralisée de RETOURARRIÈRELV avec sauts arrières. La sous-procédure VALIDE(*X*,*C*,*p*) retourne toujours vrai si et seulement si l'instanciation partielle $X = \{v_1 \ldots v_i\}$ satisfait à toutes les contraintes dans *C*. Si ce n'est pas le cas, cette sous-procédure conserve le niveau responsable de l'échec et le retourne dans *p*.

4.4 Le problème des *n* reines

Le problème des n reines consiste à placer n reines sur un échiquier $n \times n$ d'une telle façon qu'aucune d'entre elles ne mette en échec une autre. Le lecteur néophyte apprendra qu'une reine est en échec si elle se trouve sur la même ligne, la même colonne ou la même diagonale (45° ou 135°) qu'une autre reine.

4.4.1 Encadrement dans un CSP

Ce problème se définit aisément sous la forme d'un CSP : les variables sont les n reines, les domaines sont l'ensemble des couples $(x,y) \in \{1 \ldots n\} \times \{1 \ldots n\}$, c'est-à-dire les n^2 cases de l'échiquier, et les contraintes sont les suivantes :

1. $\{(x_1,y_1),(x_2,y_2) \in \{1 \ldots n\} \times \{1 \ldots n\} \mid x_1 \neq x_2\}$ (ligne)
2. $\{(x_1,y_1),(x_2,y_2) \in \{1 \ldots n\} \times \{1 \ldots n\} \mid y_1 \neq y_2\}$ (colonne)
3. $\{(x_1,y_1),(x_2,y_2) \in \{1 \ldots n\} \times \{1 \ldots n\} \mid x_1 + y_1 \neq x_2 + y_2\}$ (diagonale 135°)
4. $\{(x_1,y_1),(x_2,y_2) \in \{1 \ldots n\} \times \{1 \ldots n\} \mid x_1 - y_1 \neq x_2 - y_2\}$ (diagonale 45°)

Ces contraintes spatiales font en sorte que ce CSP ressemble beaucoup à celui appliqué à la modélisation des ARN, plus spécifiquement par rapport aux contraintes de collision. En effet, le choix d'une position particulière pour une reine affecte directement les positions possibles pour les reines subséquentes. La taille de l'espace conformationnel de ce CSP est de n^{2n}. Cette taille peut être aisément réduite en remarquant que toute solution place immanquablement chacune des n reines sur une ligne différente. Ainsi, chaque variable v_i est un couple (i,y) où $y \in \{1 \ldots n\}$. La contrainte de ligne devient redondante. De plus, en conservant une trace des colonnes utilisées dans l'algorithme de résolution, la contrainte de colonne tombe également. La méthode de résolution choisit donc une colonne libre pour la reine i sur la ligne i, puis vérifie les diagonales. La taille de l'espace conformationnel est ainsi réduite à $n!$. D'autres simplifications sont encore possibles à cette définition : une seule variable dont le domaine est l'ensemble des permutations de la suite de 1 à n, par exemple Cependant toute simplification supplémentaire rend le problème des n reines trop dissimilaire à celui de la modélisation des ARN pour être utilisé comme un prototype.

4.4.2 Évaluation des algorithmes de résolution

Dans cette sous-section, nous comparons quatre algorithmes de résolution appliqués à la définition du problème ci-haut. Ces algorithmes sont le retour arrière classique (RETOURARRIÈRE), le retour arrière probabiliste (RETOURARRIÈRELV), et deux versions du retour arrière probabiliste avec saut arrière (RETOURARRIÈRELV_SA) : saut minimal et saut maximal. Un saut arrière minimal choisit toujours le niveau dans la liste des coupables qui se retrouve le plus près du niveau courant, alors qu'un saut maximal choisit le niveau le plus éloigné. Ces deux versions différentes correspondent directement à deux implantations particulières de la sous-procédure SÉLECTIONCOUPABLE, du pseudocode présenté à la figure 42, une choisissant le niveau maximal de la liste, l'autre le niveau minimal. Bref, des quatre algorithmes évalués, trois sont différentes versions probabilistes qui sont comparées au retour arrière classique, faisant office de référence. Tous les résultats ci-bas concernant ces algorithmes ont été réalisés sur des stations de travail *Intel Pentium 4* à 2.2 GHz.

4.4.2.1 Validation des paramètres

Chacune de nos trois versions du retour arrière probabiliste possède deux degrés de liberté : les paramètres s et t qui forment ensemble la taille fixe de l'étape de correction par retour arrière local. Pour se comparer au retour arrière de référence, il faut donc d'abord valider ces paramètres. Ces deux paramètres ont été optimisés sur nos trois algorithmes probabilistes pour la résolution du problème à $n = 200$. Les quantités de solutions uniques trouvées ont été rapportées après 30 minutes d'exécution pour différentes combinaisons de valeurs pour s et t. À la figure 43, nous détaillons les résultats pour la version sans saut arrière. L'analyse de ces résultats nous donne une estimation de la taille optimale, avec $s = 75$ et $t = 100$. Le rapport de taille avec celle de l'espace complet (200!) est tout à fait négligeable (de l'ordre de 10^{-188}), donc l'espace couvert par la correction peut être qualifié de local.

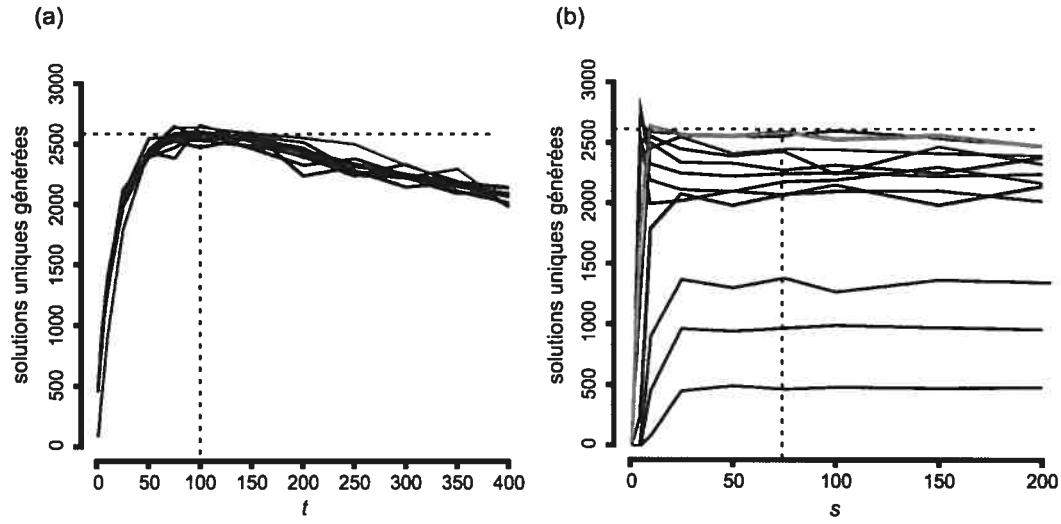


Figure 43 : Quantité de solutions uniques générées par la résolution du problème des 200 reines après 30 min. d'exécution de l'algorithme RETOURARRIÈRELV pour différentes combinaisons de valeurs pour les paramètres *s* et *t*. (a) Ici, les quantités de solutions trouvées sont rapportées en fonction de *t*, chaque courbe avec une valeur différente pour *s*. Les courbes sont pratiquement confondues, témoignant de la relative indépendance de *s* vis-à-vis de *t*. Le maximum approximatif global se trouve près de $t = 100$. (b) Cette fois les quantités de solutions trouvées sont rapportées en fonction de *s*, chaque courbe avec une valeur différente pour *t*. La courbe en gris correspond à $t = 100$, son maximum approximatif se trouvant près de $s = 75$.

Nous avons réalisé le même exercice pour nos deux versions de RETOURARRIÈRELV_SA. Pour la version à saut minimal, l'estimation de la taille optimale de correction par retour arrière est de $s = 150$ et $t = 50$, alors que celle pour la version à saut maximal est de $s = 200$ et $t = 15$. Ces paramètres confèrent des tailles d'espace couvert lors de la correction qui sont toutes deux inférieures à celle de la version sans saut arrière : de l'ordre de 10^{187} pour cette dernière, alors que les versions avec petits et grands sauts sont respectivement de l'ordre de 10^{108} et 10^{34}.

4.4.2.2 Mise à l'épreuve et interprétation des résultats

Une fois ces paramètres validés, nos trois versions probabilistes peuvent se mesurer entre elles face au retour arrière classique. L'épreuve consiste à obtenir le plus de solutions uniques possibles au problème des n reines pour différentes valeurs de n, après trois heures d'exécution. Les performances du retour arrière classique sont très limitées : aucune solution

trouvée après trois heures lorsque $n > 32$. C'est pourquoi ces résultats ne sont pas détaillés ici. Par contre, il réussit à trouver toutes les solutions du problème en moins de trois heures lorsque $n < 16$. Pour ce qui est des trois algorithmes probabilistes, leur performance relative est plutôt médiocre lorsque $n < 32$, à cause de la redondance fréquente des solutions trouvées. Par contre, ces performances deviennent tout à fait excellentes lorsque $n > 32$, vis-à-vis de celles du retour arrière classique, puisque ce dernier reste bredouille. À la figure 44, nous montrons les performances des trois algorithmes probabilistes : sans saut arrière, avec saut arrière minimal et avec saut arrière maximal.

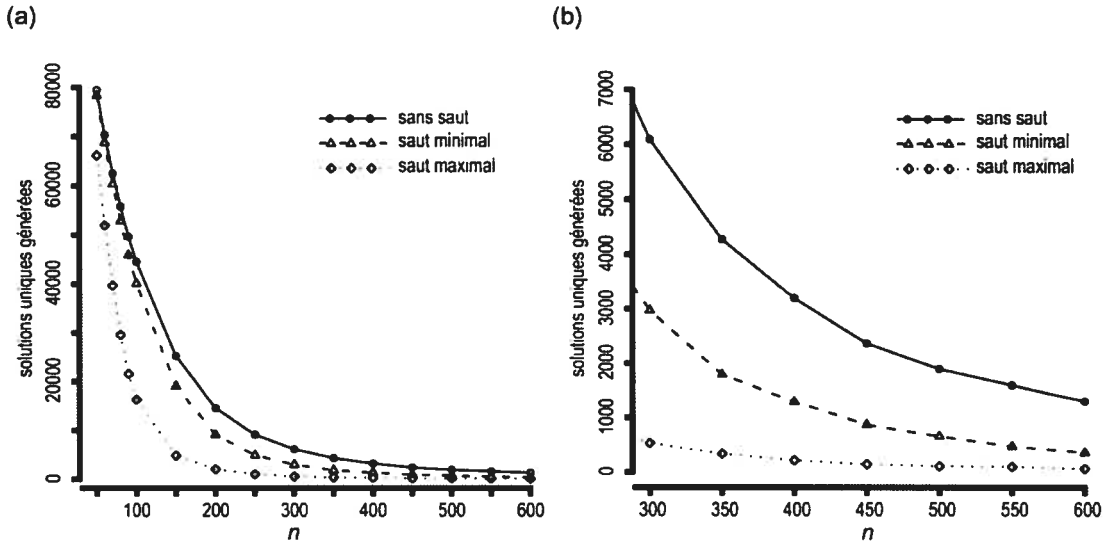


Figure 44 : Quantité de solutions uniques générées par la résolution du problème des n reines après trois heures d'exécution, pour différentes valeurs de n. (a) Seulement les performances des trois versions probabilistes sont présentées ici, car le retour arrière classique ne trouve aucune solution après trois heures lorsque $n > 32$. La version sans saut arrière surpasse les deux autres pour toutes valeurs de n testées. (b) Agrandissement du graphique à $300 \leq n \leq 600$. Clairement, les versions avec sauts arrières n'apportent aucune amélioration.

Ainsi, nous avons comparé nos trois algorithmes probabilistes entre eux face au retour arrière classique pour résoudre le problème des n reines. Tous trois le surpassent lorsque $n > 32$.

Par ailleurs, nous avons montré que les versions avec sauts arrières, minimaux ou maximaux, n'apportent pas d'amélioration concrète à la version sans saut arrière. La version avec saut maximal navigue à grands pas dans l'espace conformationnel, sans se soucier des solutions possiblement enjambées, alors que la version avec saut minimal réduit ce risque au coût de probables parcours vain de sous-arbres dus à un autre coupable situé moins profondément dans l'arbre de recherche (voir figure 40). Cependant, toutes deux présentent la possibilité d'émonder certaines solutions lors des sauts arrières, puisque le coupable est sélectionné de

façon non-sécuritaire. Puisque ni l'une ni l'autre ne surpasse la version originale sans saut arrière, l'algorithme de résolution qui sort gagnant de ce prototypage par le problème des n reines est la version originale de RETOURARRIÈRELV, sans saut arrière. Pour surpasser l'algorithme de retour arrière classique, il faut cependant considérer un espace conformationnel suffisamment grand ($n > 32$). Heureusement, dans le cadre de l'application à la modélisation des ARN, l'espace conformationnel est habituellement assez grand, que ce soit dû à la longueur de la séquence où à la variété des relations d'appariement utilisées. Il existe d'autres algorithmes de résolution du problème des n reines que ceux que nous avons utilisés dans ce prototype (voir [32] et [33]), plus efficaces que le retour arrière probabiliste en exploitant notamment le fait que les contraintes sont toutes binaires et que les instanciations ne sont pas interdépendantes. Ces deux critères sont totalement à l'opposé de la définition du problème de modélisation des ARN par satisfaction de contraintes, rendant de telles optimisations inapplicables. C'est pourquoi nous avons choisi une méthode de résolution simple du problème des n reines; pour permettre un prototypage adéquat.

4.5 Résultats de modélisation

Dans cette section, notre méthode de résolution du problème de modélisation des ARN par satisfaction de contraintes, tel que défini dans les sections précédentes, est mise à l'épreuve sur des exemples concrets. D'abord, la modélisation de la structure atomique de la tige et la boucle anticodon d'ARN de transfert sera utilisée pour comparer les performances de notre algorithme de résolution RETOURARRIÈRELV sur des espaces conformationnels de différentes tailles. Ensuite, la modélisation d'une partie de la structure d'un intron du groupe II sera présentée comme un exemple concret d'accomplissement.

4.5.1 Modélisation de l'anticodon et taille de l'espace conformationnel

Maintenant que notre méthode de modélisation est dûment décrite et prototypée, il reste à expérimenter son comportement sur un banc d'essai : la modélisation de la structure moléculaire de l'anticodon de $ARNt^{Phe}$. Cette structure en tige et boucle terminale est simple et bien connue : cinq paires de bases en double hélice dans la tige suivit d'une boucle 2-5 (les sept bases de la boucle sont empilés à l'exception de la 2^e et 3^e, où le squelette se replie en

'U' pour relier l'autre côté de la double hélice). La figure 45 montre la structure secondaire de l'anticodon avec l'arbre de recouvrement de son graphe de relation utilisé pour la modélisation.

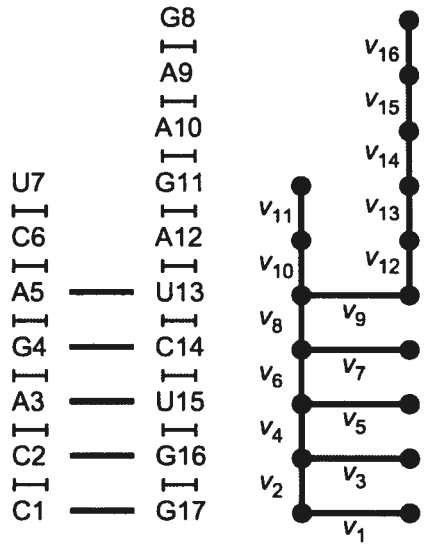


Figure 45 : Structure secondaire de l'anticodon de ARNt^{Phe} (à gauche) et arbre de recouvrement du graphe de relation utilisé pour le banc d'essai de modélisation (à droite). La structure secondaire suit la nomenclature de Leontis & Westhof [30]. L'ordre de construction est indiqué par la numérotation des variables du CSP.

Pour le banc d'essai, nous avons défini trois espaces conformationnels de différentes tailles, selon les variétés de relations d'appariement utilisées. Le premier espace est qualifié d'espace de base; il servira de point de référence. Le second espace double la taille des domaines du premier. Pour le troisième, celui-ci considère la taille maximale possible, en fonction de la base de données de relations d'appariement utilisée par *MC-Sym*. Dans la modélisation de l'anticodon, seulement deux types de relations d'appariement sont nécessaires : empilements de bases adjacentes et appariements Watson-Crick. Cependant, une distinction est réalisée entre les empilements de bases au long de la double hélice et à l'intérieur de la boucle : une plus grande liberté est permise dans ce dernier cas en augmentant la variété de ces empilements de base relativement aux autres. Le tableau 6 détaille les différentes variétés de relations d'appariement qui caractérisent chacun des trois différents espaces conformationnels utilisés pour le banc d'essai.

	empilements		Watson-Crick	ordre de la taille totale
	hélice	boucle		
base	3	7	5	10^{11}
double	6	14	10	10^{16}
maximale	388	388	141	10^{39}

Tableau 6 : Définition des trois différents espaces conformationnels utilisés pour la modélisation de l'anticodon. Les empilements de bases adjacentes qui se retrouvent à l'intérieur de la double hélice concernent les domaines des variables v_2, v_4, v_6 et v_8, alors que ceux à l'intérieur de la boucle concernent v_{10} à v_{16}. Quant aux appariements Watson-Crick, ils concernent les variables v_1, v_3, v_5, v_7 et v_9.

Ainsi, le banc d'essai contient trois espaces conformationnels différents pour la modélisation de l'anticodon. Pour chaque espace, l'algorithme probabiliste de résolution qui est ressorti gagnant du prototypage avec le problème des *n* reines, RETOURARRIÈRELV, est comparé à son homologue déterministe, RETOURARRIÈRE. Donc, le banc d'essai contient six instances du problème : deux méthodes de résolution sur trois espaces conformationnels différents. La figure 46 montre le premier critère d'analyse des résultats : le taux de génération de solutions à la modélisation. Nous y rapportons la progression de 360 minutes de recherche pour les six instances, à l'exception de l'instance déterministe avec l'espace maximal qui n'a trouvé aucune solution durant ce délai. L'instance probabiliste avec l'espace double génère le plus de solutions, suivit de celle avec l'espace maximale. À long terme, l'instance probabiliste avec l'espace de base et les deux instances déterministes ont toutes trois un comportement similaire, lequel est éclipsé par les autres instances probabilistes. Cependant, à court terme, c'est-à-dire sur les dix premières minutes de recherche, la distinction est plus claire pour ces trois instances : l'instance probabiliste avec l'espace de base progresse plus linéairement que les deux instances déterministes qui démontrent des plateaux de plusieurs minutes durant lesquelles aucune nouvelle solution n'est trouvée. De plus, toujours à court terme, le fait de doubler la taille des domaines a un effet inverse sur le taux de génération de solutions chez les instances déterministes et probabilistes : ralentissement chez les premiers et accélération chez les seconds. Ce ralentissement chez les instances déterministes s'estompe à long terme, l'espace double réussissant finalement à trouver plus de solutions que l'espace de base. Cependant, l'effet d'accélération par l'accroissement de la taille des domaines est beaucoup plus marqué sur les instances probabilistes, bien que l'espace maximal apporte un léger ralentissement relativement à l'espace double.

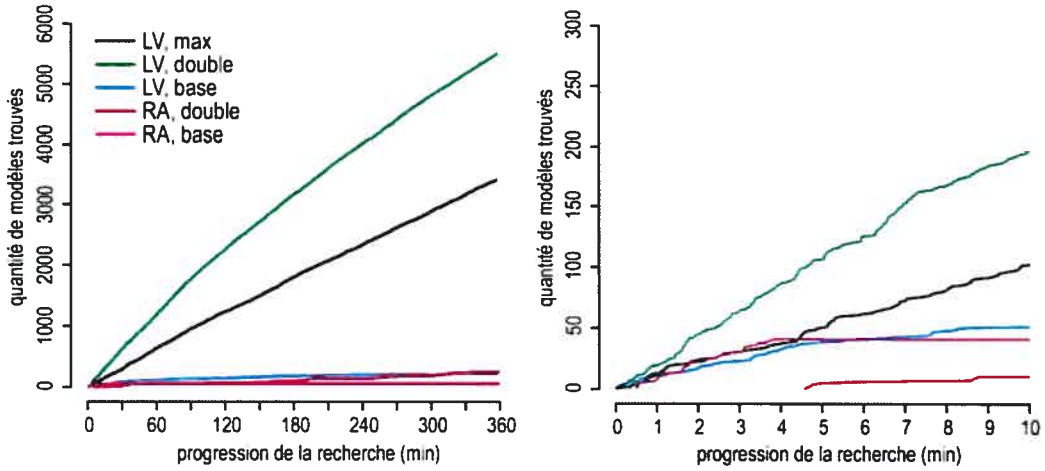


Figure 46 : Taux de génération de solutions pour les six instances de résolution du problème de la modélisation de l'anticodon. Les instances de l'algorithme RETOURARRIÈRE sont dénotées par l'étiquette RA, et celles de l'algorithme RETOURARRIÈRELV le sont par l'étiquette LV. Les 360 minutes de recherche ont été réalisées sur des stations de travail *Intel Pentium 3* à 500 MHz. Le graphique de droite concerne seulement les dix premières minutes de recherche. L'instance de RETOURARRIÈRE avec l'espace maximale n'est pas rapportée puisqu'elle n'a trouvé aucune solution durant le délai alloué.

Le second critère d'analyse mesure la progression de la diversité des solutions trouvées durant la recherche. Un taux de diversification élevé témoigne d'une exploration large de l'espace conformationnel, d'une navigation qui se déplace rapidement sans rester au même endroit trop longtemps. La diversité d'un ensemble de solutions se quantifie par la déviation RMS maximale entre toute paire de solutions de l'ensemble. Ainsi, à chaque nouvelle solution trouvée, la diversité de l'ensemble des solutions est calculée, donnant une progression du taux de diversification tout au long de la génération de solutions. À la figure 47, nous montrons cette progression pour les mêmes instances du problème précédemment utilisées. Ici, les trois instances probabilistes montre un taux de diversification plus élevé que leurs homologues déterministes, ce que nous avions prévu lors de l'observation des failles de l'algorithme de retour arrière classique à la sous-section 4.3.2. Par ailleurs, l'accroissement de la taille de l'espace conformationnel réduit le taux de diversification dans le cas des instances déterministes, alors qu'il l'augmente sensiblement dans le cas des instances probabilistes, où l'espace maximal démontre la plus grande diversité.

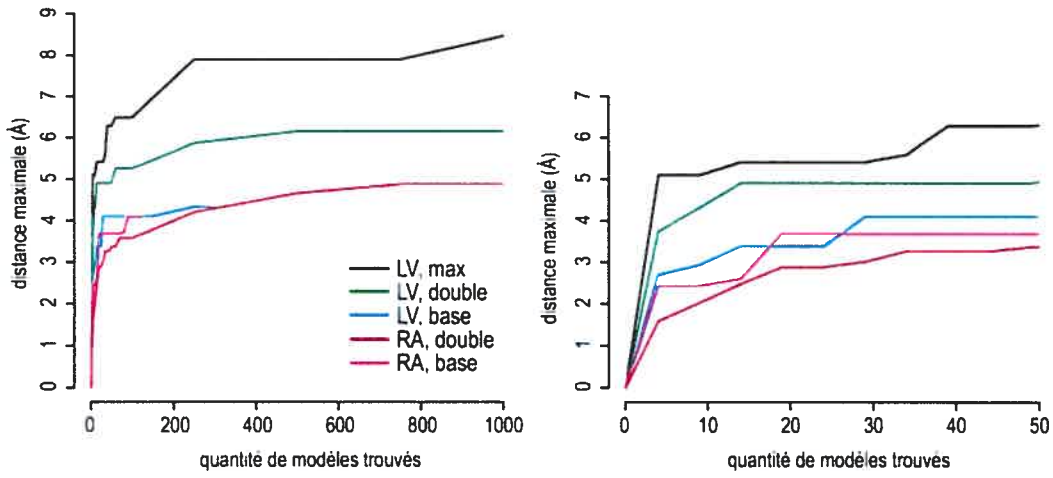


Figure 47 : Taux de diversification des solutions au cours de la génération des 1000 premières solutions pour les instances du banc d'essai de modélisation de l'anticodon. La diversité est mesurée par la déviation RMS maximale entre toute paire de solutions. Le graphique de droite ne concerne que les 50 premières solutions trouvées.

Bref, l'interprétation de ces résultats du banc d'essai de la modélisation de l'anticodon fait ressortir notre algorithme de résolution RETOURARRIÈRELV comme la méthode la plus efficace, si l'espace conformationnel est suffisamment grand, ce qui est en accord parfait avec les résultats obtenus lors du prototypage avec le problème des *n* reines. De plus, lors de l'utilisation de RETOURARRIÈRELV, l'accroissement de la taille des domaines des variables du CSP est bénéfique tant au niveau du taux de génération de solutions que du taux de diversification. Ceci renforce l'hypothèse que la restriction de l'espace conformationnel pour un problème de modélisation donné entraîne la perte de solutions. Avec RETOURARRIÈRELV, il nous est possible de conserver l'entièreté de l'espace conformationnel, de bénéficier d'un taux de diversification maximal tout en conservant un taux de génération de solutions plus qu'enviable.

4.5.2 Modélisation d'une partie d'un intron du groupe II

Comme exemple d'utilisation du système complet de modélisation des ARN dont le cadre algorithmique a été décrit jusqu'ici, nous présentons dans cette sous-section une modélisation de la structure tertiaire d'un intron du groupe II. La structure secondaire de la séquence de l'intron de *Lactococcus lactis* est illustrée à la figure 48. Cette information structurelle provient d'une collaboration avec le laboratoire de Dr. Steve Zimmerly à University of

Calgary. Sur la structure, les régions annotées correspondent aux interactions tertiaires déterminées en laboratoire par Dr. Zimmerly et son équipe. Une interaction tertiaire est un (ou plusieurs) appariement de bases qui relie ensemble deux éléments de la structure secondaire. Dans le cadre de cet exemple de modélisation, seulement une partie du domaine I est considérée, avec les interactions α-α', β-β', ISB1-ESB1 et ISB2-ESB2.

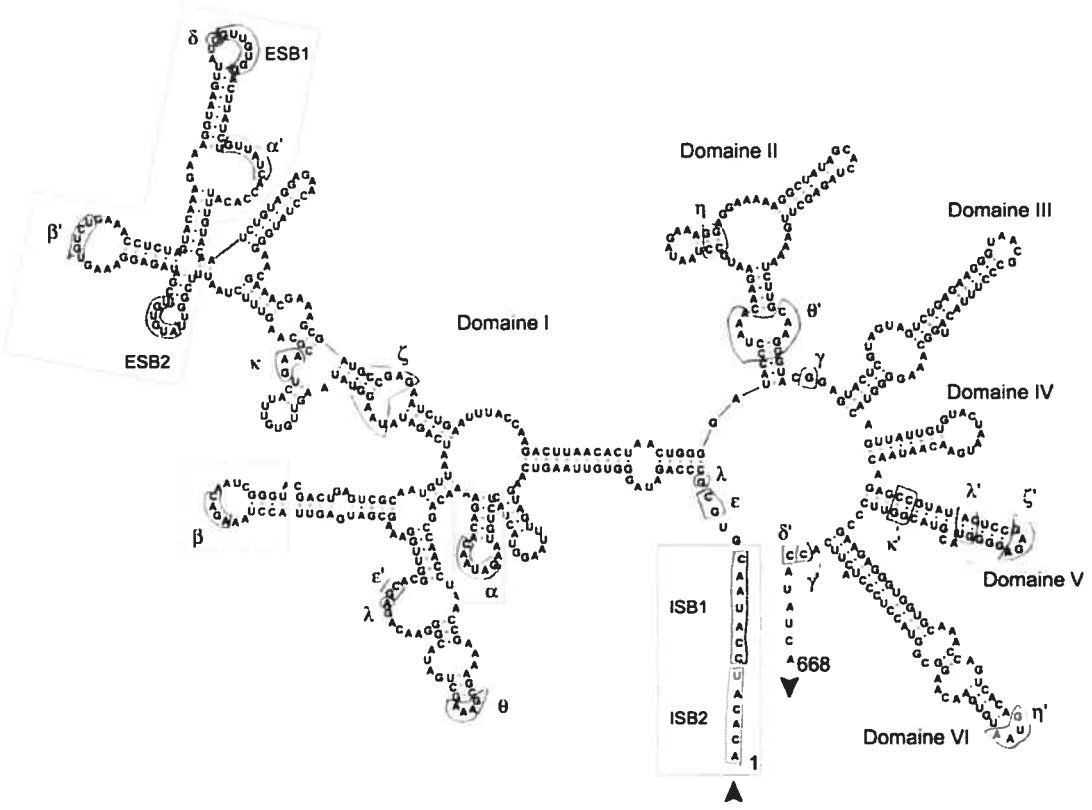


Figure 48 : Structure secondaire d'un intron du groupe II provenant de *Lactococcus lactis*, gracieuseté de Dr. Steve Zimmerly. Les annotations concernent les interactions tertiaires déterminées en laboratoire. Les flèches indiquent le début et la fin de la séquence de 668 nucléotides. Les zones encadrées forment le fragment considéré pour la modélisation.

Les interactions tertiaires considérées dans ce fragment de structure font toutes intervenir des appariements de type Watson-Crick. Par conséquent, de nouvelles tiges se forment et sont représentables dans la structure secondaire du fragment, que nous illustrons à la figure 49. Notre fragment contient quatre chaînes distinctes qui se replient sur elle-mêmes selon la structure secondaire globale ci-haut, et aussi entre elles selon les interactions Watson-Crick présentes dans ce fragment. Pour faciliter la modélisation, nous avons divisé ce fragment de nouveau en sous-fragments, tel qu'indiqué à la figure 49. Chaque sous-fragment est modélisé localement, de façon indépendante du reste de la structure globale. Nous avons décrit ce

processus de séparation d'un problème de modélisation en fragments locaux au chapitre 2, cependant il est intéressant de voir l'implication de cette méthode au niveau de la représentation en CSP. Clairement, un fragment se représente par des sous-ensembles aux ensembles de variables, domaines et contraintes du CSP global, c'est-à-dire qu'il constitue un sous-problème du CSP global. La résolution de ce sous-problème apporte des solutions partielles au problème global. En fait, les variables du sous-problème résolu sont retirées du problème global et remplacées par une seule variable dont le domaine est l'ensemble des solutions partielles au sous-problème. L'avantage est que le sous-arbre implicite de recherche formé par la résolution du sous-problème n'est parcouru qu'une seule fois, au contraire de la résolution du problème global dans son entièreté. En fait, isoler un fragment de k variables est un cas particulier de renforcement de k-consistence [25]. La modélisation de ce fragment filtre d'une certaine façon le domaine conjugué des k variables de sorte que toute instanciation de ces k variables est valide par rapport aux contraintes dont la portée est constituée de ces k variables. Bref, la modélisation par fragment permet la propagation des contraintes.

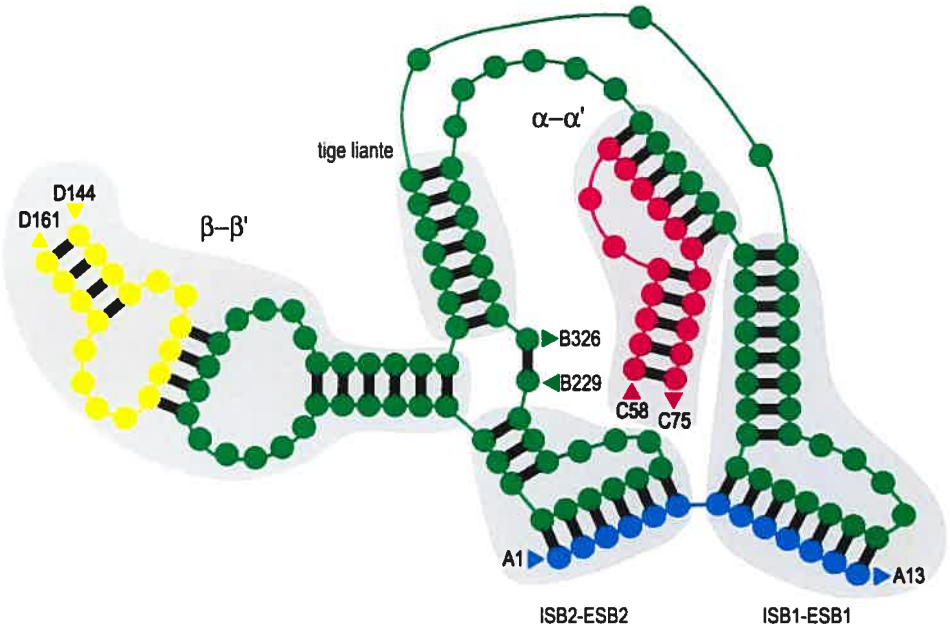
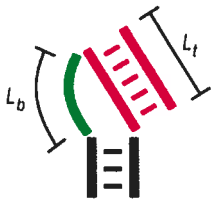


Figure 49 : Structure secondaire du fragment de l'intron du groupe II considéré pour la modélisation, incluant les interactions tertiaires. La dénotation des bases azotées a été omise. Chaque couleur correspond à une chaîne distincte, dont le sens est indiqué par des flèches aux extrémités, et dont la numérotation suit la structure secondaire globale. Les tiges composées de deux chaînes différentes sont le résultat de la considération des interactions Watson-Crick. Les zones ombrées montrent comment la modélisation est fragmentée.

La sous-fragmentation présentée ci-haut nous permet d'en apprendre un peu plus sur la structure en isolant un motif structural commun pour les fragments α-α', ISB1-ESB1 et ISB2-ESB2 : une chaîne repliée en tige terminée par une boucle qui est elle-même appariée à une autre chaîne. Ainsi, le motif est généralement composé de deux tiges dont les extrémités sont reliées directement d'une part et par une boucle d'autre part. L'aspect structural caractéristique est la séparation de la boucle terminale en une partie complètement appariée avec une autre chaîne, et une autre non-appariée qui relie la tige appariée à la tige principale. Soit L_t la longueur de la tige appariée et L_b la longueur de la boucle liante. Le fragments α-α' se caractérise par $(L_t, L_b) = (6,2)$, ISB1-ESB1 par $(L_t, L_b) = (6,3)$ et ISB2-ESB2 par $(L_t, L_b) = (7,4)$. Ces trois motifs structuraux particuliers ont été recherchés dans des modèles d'ARN de référence provenant des bases de données publiques *PDB* [46] et *NDB* [47], grâce à un outil de recherche mis au point par Patrick Gendron (Laboratoire de Biologie Informatique et Théorique, Université de Montréal). Un seul exemplaire du motif a été retrouvé pour α-α' et ISB2-ESB2, et seulement neuf exemplaires pour ISB1-ESB1, lesquelles sont à moins de 0.4 Å de déviation RMS. À la figure 50, nous détaillons ces résultats et nous illustrons les modèles tridimensionnels des trois motifs trouvés. Le motif structural général possède donc une certaine singularité. Par ailleurs, dans tout les cas, la jonction entre la tige principale et la tige appariée démontre un empilement des bases terminales, tel que mesuré par annotation automatique [29]. Cette information supplémentaire vient rigidifier les contraintes structurales du motif, et rend sa modélisation plus aisée.

(a)



	L_t	L_b	quantité
α-α'	6	2	1
ISB2-ESB2	6	3	9
ISB1-ESB1	7	4	1

(b)

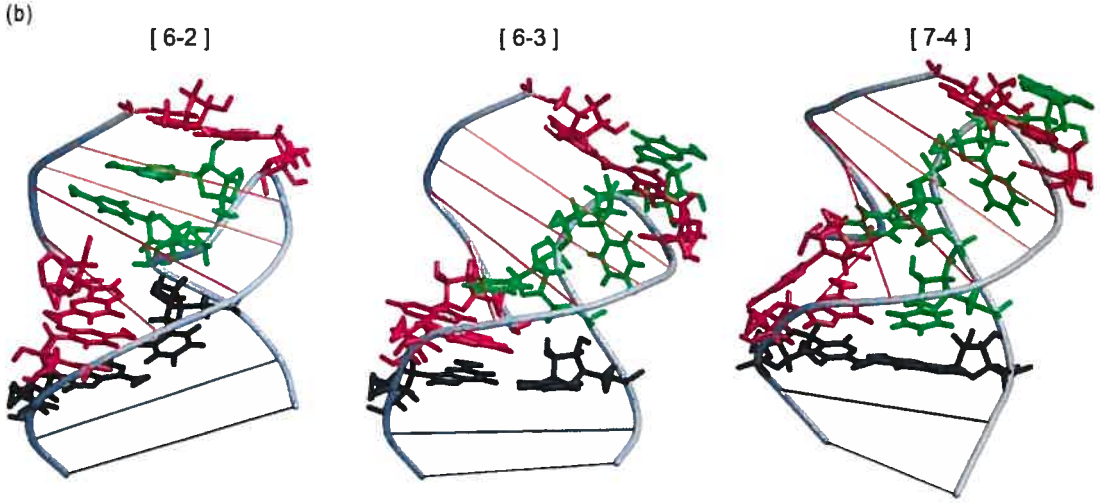


Figure 50 : Motif structural des fragments α-α', ISB1-ESB1 et ISB2-ESB2. (a) Schéma du motif avec ces paramètres L_t et L_b, respectivement la longueur de la tige appariée (rouge) et la longueur de la boucle (vert), et résultats de la recherche des trois variétés du motif dans des modèles de référence. (b) Modèles tridimensionnels des motifs extraits. Le code de couleur suit le schéma en (a). Les lignes représentent les appariements Watson-Crick des tiges, seules les bases azotées aux extrémités des tiges et à l'intérieur de la boucle sont illustrées. Le squelette des deux chaînes est mis en évidence par un cordon gris. L'empilement des bases à la jonction de la tige principale (noir) et de la tige appariée (rouge) est bien visible sur ces modèles. Images générées par *Molscript* [36] [44] et *Raster3D* [37] [45].

Grâce à cette identification de motif commun, les fragments α-α', ISB1-ESB1 et ISB2-ESB2 sont modélisés conformément à l'information structurale que nous avons apprise. Pour sa part, le fragment β-β' est modélisé avec un peu plus de liberté, où l'interaction entre les boucles terminales est considérée comme une tige. Au tableau 7, nous détaillons les résultats de modélisation pour ces fragments. Nous avons réalisé toutes les modélisations en utilisant notre algorithme probabiliste de résolution (RETOURARRIÈRELV) sur des stations de travail *Intel Pentium 4* à 2.2 GHz. La modélisation finale regroupe ensemble les fragments dans une seconde étape d'exécution pour former la structure complète telle qu'illustrée à la figure 49. Cette dernière étape a duré pendant environ 49 heures pour produire 28 modèles dont la déviation RMS par paire de modèles varie entre 8.86Å et 31.86Å. À la figure 51, nous

illustrons un de ces modèles ainsi qu'un agrandissement du motif de la jonction à quatre directions formé par la tige liante, la tige du fragment β-β', la tige du fragment ISB2-ESB2 et la paire de bases initiale. Fait intéressant : nous avons tenté la même modélisation en utilisant l'algorithme classique de retour arrière, qui n'a trouvé aucune solution après une semaine d'exécution.

	quantité	temps de recherche (min)	déviation RMS (Å)	
			minimale	maximale
α-α'	312	72	1.72	8.20
β-β'	17	7	1.12	8.26
ISB1-ESB1	122	1/3	1.19	7.93
ISB2-ESB2	53	1/12	1.13	5.28

Tableau 7 : Résultats de la modélisation des fragments α-α', β-β', ISB1-ESB1 et ISB2-ESB2. La diversité minimale est la plus petite déviation RMS entre n'importe quelle paire de modèles de l'ensemble de solutions, alors que la diversité maximale est la plus grande déviation.

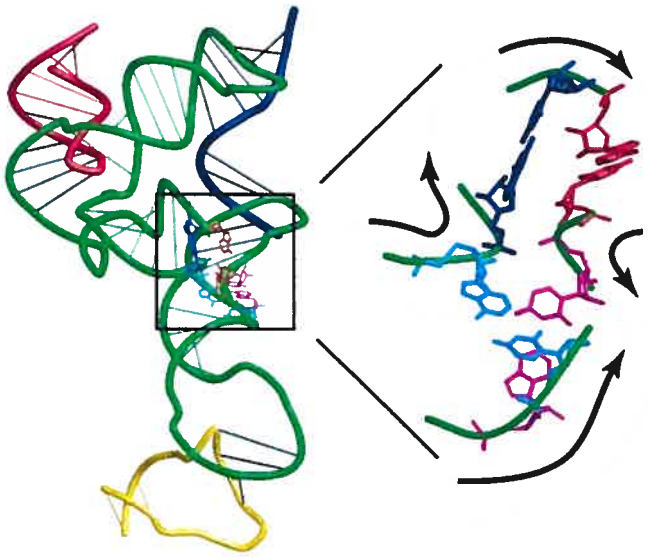


Figure 51 : Un des 28 modèles tridimensionnels trouvés par la fusion des fragments préalablement modélisés. Les couleurs des chaînes suivent celles de la structure secondaire. Seul le squelette et les appariements des tiges sont illustrés, pour plus de clarté. À droite est l'agrandissement du motif de la jonction à quatre directions, où les paires de base terminales sont montrées : en rouge la paire initiale, en bleu le début de la tige liante, en cyan le début de la tige du fragment β-β' et en magenta le début de la tige du fragment ISB2-ESB2. Les directions des chaînes sont indiquées. Images générées par *Molscript* [36] [44] et *Raster3D* [37] [45].

4.6 Conclusion

Au cours de ce chapitre, nous avons formellement défini le problème de la modélisation des ARN comme un problème de satisfaction de contraintes. Si le contexte s'applique parfaitement à la résolution classique par retour arrière, il restreint par ailleurs toute utilisation de mécanismes d'optimisation dynamique qui assume l'indépendance relative des variables instanciées vis-à-vis de la validation des contraintes. En effet, nous avons expliqué que l'application particulière de la méthode de construction par relations d'appariement implique que l'arité de toutes les contraintes applicables est toujours maximale, i.e. qu'elles ont comme portée toutes les variables instanciées à tout moment donné, et qu'il n'est pas possible de transgresser l'ordonnancement de l'instanciation des variables de quelque façon que ce soit. C'est pourquoi nous avons développé une heuristique probabiliste, où l'exploration de l'espace conformationnel instancie les variables de façon aléatoire en ne se permettant qu'une quantité fixe de corrections et de retours arrière de un ou plusieurs niveaux à la fois (sauts arrières). Par la suite, nous avons prototypé trois versions de cet algorithme probabiliste de résolution sur le problème des *n*-reines et nous les avons comparées face au retour arrière déterministe. Nos trois versions probabilistes surpassent l'algorithme déterministe lorsque l'espace conformationnel est suffisamment grand, la version sans saut arrière se classant en première position devant ceux avec saut arrière. La technique de saut arrière est donc écartée de notre algorithme probabiliste de résolution. La question est de savoir si tout saut arrière entraîne l'émondage de solutions, dans le cadre d'une heuristique probabiliste non-exhaustive, ou plutôt si c'est seulement l'identification du niveau cible du saut arrière qui est déficiente. Assumant la deuxième hypothèse, nos résultats montrent clairement que les méthodes de sélection du coupable tant au niveau le plus près qu'au niveau le plus éloigné du niveau invalidé sont inefficaces. Cependant d'autres méthodes de sélection pourrait s'avérer plus profitable, par exemple choisir à partir du niveau invalidé celui qui l'a fait échoué le plus fréquemment. Mieux encore : établir pour chaque niveau de la recherche une liste de priorité sur tous les autres niveaux précédents pour la sélection du candidat cible d'un saut arrière, advenant l'échec de ce niveau. Il serait difficile d'automatiser l'apprentissage de telles listes, cependant l'usager pourrait les établir selon ses connaissances du problème de modélisation en particulier.

L'utilisation de notre algorithme probabiliste sans saut arrière pour la résolution du problème de la modélisation de l'anticodon de $ARNt^{Phe}$ a démontré son efficacité sur un espace

conformationnel de taille maximale. En effet, pour arriver à résoudre le même problème en utilisant le retour arrière classique, l'espace conformationnel doit être tronqué. Si la conservation du déterminisme de la méthode de recherche doit passer par la troncature de l'espace conformationnel pour s'effectuer concrètement, il y a paradoxe puisque cette troncature compromet le déterminisme global du problème de modélisation. Comment peut-on réduire la taille de l'espace conformationnel d'un problème particulier de modélisation tout en garantissant la conservation de toutes les instanciations complètes et valides? L'utilisation de notre algorithme probabiliste écarte ce problème ouvert en permettant une recherche sur un espace maximal avec des performances supérieures.

Chapitre 5

Conclusion

Les études et les efforts de développement que nous avons présentés dans ce mémoire résultent finalement en une nouvelle version complète du système de modélisation *MC-Sym*. Nous y avons redéfini tous les éléments conceptuels reliés à l'exploration de l'espace conformationnel de la structure tertiaire d'un ARN, tant au niveau de la méthode de construction du modèle qu'à celui de l'algorithmique nécessaire à la résolution du problème de modélisation par satisfaction de contraintes. Les éléments du système *MC-Sym* original que nous avons conservés et qui constituent le cadre de conception sont la base de relations d'appariement avec ses interfaces et l'analyseur syntaxique du format des données d'entrée. De plus, les structures de données utilisées pour manipuler des modèles d'ARN proviennent de la librairie *MC-Core* [41].

5.1 Retour sur les chapitres et sur leurs résultats

Au cours du chapitre 2, nous avons établit notre approche globale de modélisation en comparaison avec celle de l'engin de modélisation original de *MC-Sym*. D'abord, nous avons mis en évidence que l'engin original n'arrive pas à construire adéquatement le squelette du modèle par exploration de l'espace des conformations statiques des nucléotides. En effet, l'engin souffre d'incohérence locale dans 83.5% des constructions de relation d'adjacence. Pour palier à cette incohérence, nous avons développé un engin de modélisation qui positionne d'abord les bases azotées et les groupements phosphates par relations d'appariement et de façon indépendante, puis qui les interconnecte par la construction artificielle des riboses. L'incohérence locale manifestée par notre nouvel engin est réduite à la proportion quasi négligeable de 1.5%. Sur une expérimentation comparative de la modélisation de l'anticodon d'un ARN de transfert, notre nouvel engin complète sa recherche exhaustive par retour arrière en 15 minutes, alors que l'engin original n'arrive pas à compléter la sienne après 13 jours entiers, sur le même espace conformationnel de relations d'appariement. Le succès de notre approche est tout aussi retentissant lors de

l'expérimentation comparative de la modélisation d'un motif de boucle interne 4-2 : notre nouvel engin complète sa recherche en moins de 3 minutes alors que l'engin original nécessite près de 9 jours. De plus, lors de cette dernière expérimentation, notre engin a trouvé 85 modèles différents alors que l'engin original n'en a trouvé aucun. Dans la conception de ces deux expérimentations comparative de modélisation, la seule différence vient de la spécification de l'espace des conformations statiques de nucléotides qui est nécessaire seulement pour l'engin original. Par conséquent, nous concluons que l'exploration de cette espace par l'engin original est la source principale de l'inefficacité de la recherche. Finalement, nous avons mis à l'épreuve notre nouvel engin sur la modélisation de la structure complète d'un ARN de transfert.

Par la suite, nous avons consacré le chapitre 3 à la définition et à l'analyse formelle de notre méthode de construction artificielle d'un modèle de ribose. Cette sous-procédure de construction est utilisée par l'engin de modélisation pour interconnecter une base azotée et deux phosphates par un ribose. La méthode de construction est paramétrée par les torsions caractéristiques de la molécule de ribose : ρ, χ, γ, β et ε. Une procédure d'optimisation numérique trouve le quintuplet de ces torsions qui construit un ribose s'interconnectant depuis la base jusqu'aux phosphates de la façon la plus précise. Pour ce faire, nous avons développé et comparé deux approches différentes : une méthode d'optimisation linéaire sans dérivés et une méthode d'estimation directe inédite. Des deux approches de détermination des paramètres optimaux de construction, aucune ne se démarquent substantiellement l'une de l'autre lors de l'évaluation de la précision de la reconstruction des riboses de modèles de référence. Cependant, notre méthode d'estimation présente l'avantage clé de présenter une complexité constante quant au appels à la méthode de construction : exactement deux itérations de construction. De plus, la majorité des erreurs sur l'estimation de la pseudorotation sont inférieures à la longueur de l'intervalle d'identification d'un mode de *puckering*. Par conséquent, sans toutefois permettre de fixer un mode de *puckering* en particulier pour un nucléotide dans le contexte de la modélisation, nous pouvons conclure que le mode de *puckering* adéquat aux positions relatives de la base azotée et des phosphates, étant donné celles observées dans les modèles de référence, sera bien représenté par la construction artificielle.

Enfin, dans le chapitre 4, nous nous sommes intéressé à l'encadrement formel du problème de modélisation discrète des ARN dans un CSP (*constraint satisfaction program*). Par cet

encadrement, l'algorithme de résolution classique pour ce problème est un retour arrière simple et direct. Puisque cet algorithme a une complexité exponentielle au double degré de la taille de l'ARN à modéliser et de la taille de l'espace conformationnel considéré, nous avons développé une heuristique de recherche probabiliste non-exhaustive. Nous avons comparés les performances de trois versions de cet algorithme probabiliste, dont deux utilisent la technique du saut arrière, avec l'algorithme classique de retour arrière d'abord sur un prototype, le problème des n reines, où toutes nos versions probabilistes ce sont montrées d'une efficacité supérieure lorsque $n > 32$. De plus, ce prototypage a permis de démontrer que la technique du saut arrière, telle qu'utilisée dans ces algorithmes, n'apporte pas d'amélioration concrète. Par la suite, nous avons modélisé l'anticodon d'un ARN de transfert avec notre algorithme probabiliste et avec l'algorithme par retour arrière en utilisant plusieurs espaces conformationnels de taille différente. En plus de démontrer un taux de génération de solution beaucoup plus élevé, tel que prédit par le prototypage, notre algorithme probabiliste permet de considérer l'espace conformationnel de taille maximale, ce qui est pratiquement impossible dans le cas du retour arrière. En effet, ce dernier doit tronquer son espace conformationnel pour permettre un taux de génération de solution minimal, rejetant potentiellement certaines solutions par cette troncature. Finalement, nous avons mis à l'épreuve notre algorithme probabiliste sur la modélisation d'une partie du domaine I d'un intron du groupe II.

5.2 Développements futurs

Bref, notre nouveau système de modélisation a fait ses preuves. Dans l'engin de modélisation au cœur de *MC-Sym*, l'approche de modélisation discrète où les bases azotées et les phosphates sont positionnés par relations d'appariement et interconnectés par la construction artificielle des riboses remplace donc celle où tout le nucléotide entier est positionné par relation d'appariement et où le mode de *puckering* du ribose est discrétisé dans un sous-espace conformationnel. Toutefois, certaines voies de développement que nous avons empruntées dans la conception de ce système peuvent être explorées davantage.

C'est le cas de la méthode de positionnement du phosphate. En fait, celle que nous avons implantée est plutôt simple : elle reflète directement la position originale du phosphate dans la relation d'adjacence extraite du modèle de référence analysé lors de la création de la base

de relations. Cette méthode est adéquate lors de la construction d'une relation d'adjacence, cependant nous savons que certaines relations d'adjacence ne sont pas explicitement construites lors de la modélisation selon la spécification de l'ordre de construction. C'est pourquoi nous avons dû recourir à une sous-procédure de construction pour positionner les phosphates manquants : la fermeture. La fermeture cherche dans l'ensemble de relations d'adjacence inutilisé celle qui représente le plus précisément la relation d'adjacence implicite, et l'utilise pour positionner le phosphate manquant. Cependant, advenant le cas où la relation extraite la plus précise présente quand même une différence non-négligeable avec la relation implicite exprimée dans la construction, cas assez fréquent d'ailleurs, la position résultante du phosphate sera tout a fait différente dépendamment de la base azotée choisi comme référentiel pour l'application de la transformation. La figure 52 illustre cette divergence, et montre un compromis simple où la position finale choisie est à mi-chemin entre les deux options. Cependant, il n'y a aucune garantie que ce compromis reflète le positionnement adéquat du phosphate.

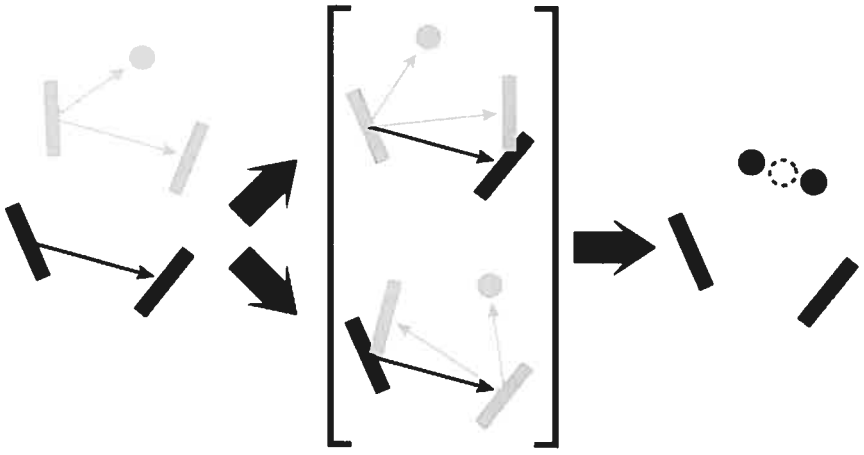


Figure 52 : Divergence dans le positionnement d'un phosphate par fermeture. Les rectangles représentent des bases azotées adjacentes, et les cercles représentent les phosphates. À gauche sont illustrées la relation d'adjacence implicitement exprimée lors de la construction (en noir) et la relation choisie pour positionner le phosphate (en gris). Au centre sont représentés les deux scénarios de positionnement selon la base azotée choisie comme référentiel pour le positionnement du phosphate. À droite est illustré le résultat des deux scénarios de positionnement. Les phosphates positionnés de cette façon sont en noir, le phosphate en trait pointillé représente une position intermédiaire potentielle.

D'un point de vue algorithmique, la méthode de modélisation par fragments que nous avons utilisé pour modéliser un ARN de transfert au chapitre 2 et une partie du domaine I d'un intron du groupe II au chapitre 4 se présente comme une technique d'optimisation de la méthode de résolution du problème de satisfaction de contraintes. Nous avons expliqué

l'impact de la fragmentation sur les variables de la représentation en CSP. En fait, dans l'ensemble des variables $V = \{v_1, v_2, \ldots, v_n\}$, la fragmentation des variables v_2, v_3 et v_4, par exemple, représente un renforcement de cohérence de problème entier. En effet, la modélisation indépendante des variables v_2, v_3 et v_4 a comme effet de filtrer la conjugaison de leur domaine respectif de tous les triplets d'instance qui ne satisfont pas aux contraintes applicables. Ainsi, le problème entier connaît dorénavant seulement des triplets $< v_2, v_3, v_4 >$ valides, leur instanciation sera toujours cohérente. Cette propagation des contraintes est un cas particulier de *k*-consistance [25] où seul un sous-ensemble ordonné particulier est de cohérence garantie. Dans le système *MC-Sym* à ce jour, la modélisation par fragments est réalisable, cependant l'usager doit lui-même créer les modélisations indépendantes des fragments puis les regrouper dans une modélisation finale. Ce processus devrait être automatisé, l'usager n'aurait qu'à marquer d'une certaine façon les fragments de structure qu'il veut voir se modéliser indépendamment. Mieux encore, des régions fortement structurées et redondantes, comme les tiges par exemple, pourraient être automatiquement détectées et fragmentées.

Références

[1] Eddy S. R., *Non-coding RNA genes and the modern RNA world*, Nature Reviews Genetics, 2, 2001, 919-929

[2] Gottesman S., *Stealth regulation: biological circuits with small RNA switches*, Genes & Development, Cold Spring Harbor Laboratory Press, 2002, 2829-2842

[3] Watson J. D., Crick F. H., *Molecular structure of nucleic acids: A structure for deoxyribonucleic acid*, Nature, 171, 1953, 694-967

[4] Cedergren R., Major F., *Modeling the tertiary structure of RNA*, RNA Structure and Function, R. W. Simon and M. Grunberg-Manago, Cold Spring Harbor Laboratory Press, 1998, 37-75

[5] Major F., Griffey R., *Computational methods for RNA structure determination*, Current Opinion in Structural Biology, 11, 2001, 282-286

[6] Ravelli R. B., McSweeney S. M., *The 'fingerprint' that x-rays can leave on structures*, Structure, 8, 2000, 315-328

[7] Zucker M., Sankoff D., *RNA secondary structures and their prediction*, Bulletin of Mathematical Biology, 46(4), 1984, 591-621

[8] Zucker M., *On finding all suboptimal foldings of an RNA molecule*, Science, 244, 1989, 48-52

[9] Rivas E., Eddy S. R., *A dynamic programming algorithm for RNA structure prediction including pseudoknots*, Journal of Molecular Biology, 285, 1999, 2053-2068

[10] Lavery R., Zakrzewska K., Sklenar H., *JUMNA (junction minimization of nucleic acids)*, Computer Physics Communications, 91, 1995, 135-158

[11] Olson W. K., Bansal M., Burley S. K., Dickerson R. E., Gerstein M., Harvey S. C., Heinemann U., Lu X. J., Neidle S., Shakked Z., Sklenar H., Suzuki M., Tung C. S., Westhof E., Wolberger C., Berman H. M., *A standard reference frame for the description of nucleic acid base pair geometry*, Journal of Molecular Biology, 313, 2001, 229-237

[12] Harvey S. C., Wang C., Teletchea S., Lavery R., *Motifs in nucleic acids: molecular mechanics restraints for base pairing and base stacking*, Journal of Computational Chemistry, 24, 2003, 1-9

[13] Massire C., Westhof E., *MANIP: an interactive tool for modeling RNA*, Journal of Molecular Graphics and Modeling, 16, 1998, 197-205

[14] Mueller F., Brimacombe R., *A new model for the three-dimensional folding of escherichia coli 16S ribosomal RNA. I. fitting the RNA to a 3D electron microscopic map at 20 Å*, Journal of Molecular Biology, 271(4), 1997, 524-544

[15] De Rijk P., Wuyts J., De Wachter R., *RnaViz2 : an improved representation of RNA secondary structure*, Bioinformatics, 19(2), 2003, 299-300

[16] Major F., *Building three-dimensional ribonucleic acid structures*, IEEE Computing in Science & Engineering, sept./oct., 2003, 44-53

[17] Major F., Gautheret D., Cedergren R., *Reproducing the three-dimensional structure of a tRNA molecule from structural constraints*, Proceedings of the National Academy of Sciences – Biochemistry, 90, 1993, 9408-9412

[18] Lemieux S., Major F. *RNA canonical and non-canonical base pairing types : a recognition method and complete repertoire*, NAR, 30(19), 2002, 4250-4263

[19] Saenger W. *Principles of nucleic acid structure*, Springer-Verlag, New York, USA, 1984.

[20] Altona C., Sundaralingman M. *Conformational analysis of the sugar ring in nucleosides and nucleotides: a new description using the concept of pseudorotation,* Journal of the American Chemical Society, 94, 1972, 8205-8212.

[21] Levitt M., Warshel A., *Extreme conformational flexibility of the furanose ring in DNA and RNA*, Journal of the American Chemical Society, 100, 1978, 2607-2613

[22] Parkinson G., Vojtechovsky J., Clowney L., Brünger A. T., Berman H., *New Parameters for the Refinement of Nucleic Acid*, Acta Cryst D52, 1996, 57-64

[23] Bazaraa M.S., Shetty C.M. *Nonlinear programming theory and algorithms*, John Wiley & Sons, 1979.

[24] Lemieux S., Major F. *Automatic 3-D modeling of RNA using the minimal cycle basis decomposition, in press*

[25] Dechter R., Frost D., *Backjump-based backtracking for constraint satisfaction problems*, Artificial Intelligence, 136, 2002, 147-188

[26] Johnsonbaugh R., *Discrete Mathematics*, Prentice Hall, 1997

[27] Major F., Turcotte M., Gautheret D., Lapalme G., Fillion E., Cedergren R., *The combination of symbolic and numerical computation for three-dimensional modeling of RNA*, Science, 253, 1991, 1255-60.

[28] Lemieux S., Oldziej S., Major F., *Nucleic acids: qualitative modeling*, Encyclopedia of Computational Chemistry, N.L. Allinger et al. eds., John Wiley & Sons, West Sussex, England, 1998.

[29] Gendron P., Lemieux S., Major F., *Quantitative analysis of nucleic acid three-dimensional structures*, Journal of Molecular Biology, 308(5), 2001, 919-36.

[30] Leontis N. B., Westhof E., *Geometric nomenclature and classification of RNA base pairs*, RNA, 7, 2001, 499–512.

[31] Brassard G., Bratley P., *Algorithmique : conception et analyse*, Masson, Les Presses de l'Université de Montréal, 1987.

[32] Sosič R., Gu J., *Efficient local search with conflict minimization: a case study of the n-queens problem*, IEEE Transactions on Knowledge and Data Engineering, 6(5), 1994, 661-8.

[33] Erbas C., Sarkeshik S., Tanik M. M., *Different perspectives of the n-queens problem*, Proceedings of the ACM 1992 Computer Science Conference, 1992.

[34] Kabsch W., *A solution for the best rotation to relate two sets of vectors*. Acta Crystallographica A, 32, 1976, 922-923.

[35] Kabsch W., *A discussion of the solution for the best rotation to relate two sets of vectors*. Acta Crystallographica A, 1978, 34, 827-828.

[36] Kraulis P. J., *Molscript: a program to produce both detailed and schematic plots of protein structures*, Journal of Applied Crystallography, 24, 1991, 946-950.

[37] Merritt E. A., Bacon D. J., *Raster3D: photorealistic molecular graphics*, Methods in Enzymology, 277, 1997, 505-524.

[38] RnaViz
http://rrna.uia.ac.be/rnaviz

[39] RasMol
http://www.umass.edu/microbio/rasmol

[40] MC-Sym : Macromolecular Conformations by SYMbolic programming
http://www-lbit.iro.umontreal.ca/mcsym/

[41] Mc-Core
http://sourceforge.net/projects/mccore

[42] Nucleic Acid Standards – Ideal Geometries :
http://ndbserver.rutgers.edu/standards/ideal_geometries.html

[43] The R Project for Statistical Computing :
http://www.r-project.org

[44] Molscript

http://www.avatar.se/molscript

[45] Raster3D

http://www.bmsc.washington.edu/raster3d

[46] RCSB Protein Data Bank

http://www.pdb.org

[47] Rutgers Nucleic Acid Database

http://ndbserver.rutgers.edu