Université de Montréal

Prédiction structurale de biomolécules à l'aide d'une construction d'automates cellulaires simulant la dynamique moléculaire

par
André Caron

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de Philosophiae Doctor (Ph.D.)
en informatique

Avril 2008

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée :

Prédiction structurale de biomolécules à l'aide d'une construction d'automates cellulaires simulant la dynamique moléculaire

Présentée par :
André Caron

a été évaluée par un jury composé des personnes suivantes :

Dr Michel Boyer
président-rapporteur

Dr François Major
directeur de recherche

Dr Pascale Legault
membre du jury

Dr Michel Dumontier
examinateur externe

Dr Thérèse Cabana
représentante du doyen de la FES

Résumé

À l'aide d'automates cellulaires, cette recherche propose une nouvelle approche par auto-organisation pour simuler la dynamique moléculaire de l'ARN. La simulation utilise une représentation simplifiée de l'ARN, où chaque nucléotide est considéré comme une particule autonome et une molécule d'ARN comme une chaîne de ces particules. Une simulation type s'initialise avec la forme étirée de la molécule et produit, suite à un processus d'auto-organisation, un état d'équilibre duquel émerge la structure secondaire native. Le mouvement des particules est de type Brownien et il est soumis à la contrainte de règles locales simples qui n'impliquent aucun algorithme d'optimisation. La modélisation se limite aux structures secondaires de l'ARN et exclut les structures tertiaires. Aucune partition ou réduction n'est faite sur l'espace des structures secondaires.

Le développement de ce modèle a nécessité la mise au point d'un nouveau concept nommé « force relative de rétention ». Cette force relative de rétention se veut une représentation approximative, sous forme de pourcentage, de l'effet global des forces d'attraction et de répulsion impliquées au niveau moléculaire de l'ARN. Son comportement s'est avéré conforme aux observations générales faites sur la dénaturation de l'ARN.

Les résultats ont été concluants pour plusieurs petites molécules constituées de 10 à 48 nucléotides, dont une molécule avec pseudonoeuds, sept molécules formées de deux brins et une molécule formée de trois brins. Deux structures intermédiaires, actuellement considérées comme importantes dans le repliement de la structure avec pseudonoeuds, sont également ressorties comme structures dominantes dans les simulations. Ces résultats démontrent qu'il est possible de prédire les structures secondaires de molécules d'ARN par un simple processus dynamique d'auto-organisation.

Mots clés : ARN, structure secondaire, repliement, hybridation, force relative de rétention, mouvement Brownien, auto-organisation, émergence.

Abstract

Using cellular automata, this research proposes a new approach by self-organization to simulate the molecular dynamics of RNA. The simulation uses a coarse-grained representation of RNA, where every nucleotide is considered to be an autonomous particle and a molecule of RNA is a chain of these particles. Each simulation starts with the stretched form of the molecule and, by a self-organization process, reaches an equilibrium state where the native secondary structure emerges. The particles are animated by a Brownian motion and are constrained by simple local rules that implicate no optimization algorithm. Modeling is limited to RNA secondary structures and excludes tertiary structures. No partition or reduction is made on the secondary structure space.

The development of this model required a new concept called "relative retention force". This relative retention force is meant to be an approximate value, expressed as a percentage, for the attraction and repulsion force implicated at the molecular level of RNA. Its behavior correlated with the general observations made on RNA denaturation.

The results were conclusive for several small molecules comprising from 10 to 48 nucleotides, including a molecule with pseudoknots, seven molecules with two strands and one molecule with three strands. Two intermediate structures, considered important in this pseudoknot folding, also emerged as dominant structures in the simulations. These results prove that it is possible to predict secondary structures of RNA molecules by a simple dynamics of self-organization.

Keywords: RNA, secondary structure, folding, hybridization, relative retention force, Brownian motion, self-organization, emergence.

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

AA : acide aminé

AC : automate cellulaire

ADN : acide désoxyribonucléique

ARN : acide ribonucléique

CMB : contrainte minimale de boucle

DM : dynamique moléculaire

FRR : force relative de rétention

nt : nucléotide

TE : taux d'émergence

3D : tridimensionnelle

Dédicace

Cette thèse est dédiée à ma conjointe, Sylvie, qui m'a toujours apporté son support et ses encouragements. Sans elle, je n'aurais jamais eu l'opportunité de réaliser ce projet.

Tu m'auras permis de vivre et de réaliser un rêve.

Remerciements

Je tiens à remercier mon directeur de thèse, François Major, qui a toujours su faire preuve d'une très grande ouverture d'esprit tout en conservant son sens critique. Son excellente maîtrise du domaine de la bio-informatique, associée à sa personnalité dynamique, ont été une source constante d'inspiration tout au long de ma recherche.

1 Introduction

1.1 Le contexte

La découverte en 1953, par Watson et Crick [1], de la structure tridimensionnelle (3D) de l'acide désoxyribonucléique (ADN) marqua le début d'une nouvelle ère de recherche en biologie. Dans les années qui suivirent, plusieurs percées majeures eurent lieu en biologie structurale et en génétique. Plus près de nous, durant les années 1990, on observa une croissance des trois axes suivants : l'amélioration des méthodes et des équipements de laboratoire; l'accumulation de données génétiques provenant du projet Génome Humain; l'augmentation de la performance des ordinateurs et des bases de données. De la convergence de ces trois axes résulta une véritable explosion du nombre de structures biomoléculaires publiées. La croissance de la base de données structurales PDB (*Protein Data Base*) illustre bien ce phénomène. Cette base de données qui ne contenait que quelques centaines de structures au début des années 1990, en contient aujourd'hui près de 50 000 [1].

Les chercheurs s'intéressent à la structure des biomolécules essentiellement parce que la fonction biologique de la biomolécule est intimement liée à sa structure [2]. Au fil des années, plusieurs découvertes sont venues enrichir les domaines de la dynamique moléculaire et de la prédiction structurale des molécules. Toutefois, même après plusieurs décennies de recherche, le problème de la prédiction de la conformation 3D de biomolécules à partir de la connaissance de la séquence seulement est encore aujourd'hui un sujet d'actualité. Il continue de mobiliser un grand nombre de chercheurs multidisciplinaires (biologie, informatique, mathématique, physique, chimie, etc), ce qui démontre bien la nature complexe du problème.

[1] Source : www.pdb.org

1.2 Portée de la recherche

Il existe plusieurs approches applicables au problème de la prédiction structurale. Cedergren et Major [3], dans un chapitre consacré spécifiquement à la modélisation de l'acide ribonucléique (ARN), font une évaluation des différentes approches et des résultats obtenus avec ces approches. Aucune d'entre elles ne peut prétendre être universelle. Elles présentent des forces et des faiblesses différentes. Aussi sont-elles souvent utilisées en combinaison pour obtenir de meilleures prédictions.

Notre recherche propose une nouvelle approche pour simuler la dynamique moléculaire de l'ARN. Partant de l'idée, généralement acceptée, que la structure d'une biomolécule émerge d'un ensemble d'interactions au niveau moléculaire, nous avons choisi d'explorer une approche basée sur l'auto-organisation. Pour ce faire, nous avons développé une construction d'automates cellulaires permettant de simuler la dynamique moléculaire et de prédire les structures qui résultent du repliement de l'ARN. Notre modèle se limite aux structures secondaires de l'ARN.

La décision d'utiliser les automates cellulaires repose essentiellement sur leur capacité à modéliser les propriétés émergentes d'un système dynamique complexe en utilisant des règles locales simples. Quelques chercheurs ont déjà utilisé des automates cellulaires pour la modélisation de polymères [4, 5] ou la modélisation d'interactions protéiniques [6 – 9]. Mais, à notre connaissance, il n'existe actuellement aucune modélisation, avec automates cellulaires, de la dynamique moléculaire de l'ARN.

1.3 Concepts de base en biologie moléculaire

L'étude de la structure des biomolécules nécessite une compréhension de l'organisation de la matière vivante et de la biologie moléculaire. Afin de bien cerner la problématique, les éléments nécessaires à la compréhension de la partie biologique du problème sont présentés dans ce chapitre.

Il existe quatre grandes classes de biomolécules, soit les protéines, les glucides ou hydrates de carbone, les lipides et les acides nucléiques [2, 10]. Tous, sauf les lipides qui sont polymorphes, sont constitués de plusieurs maillons élémentaires similaires. On parle alors de polymères ou, plus précisément, de biopolymères. Les protéines sont des biopolymères formés par l'assemblage d'acides aminés (AA). Une chaîne de plus de 20 AA est aussi appelée un polypeptide ou une chaîne polypeptidique. Les glucides sont des biopolymères formés de un ou plusieurs sucres; lorsqu'ils ont plus d'un sucre, on les appelle des polysaccharides et, dans le cas contraire, des monosaccharides. Finalement, les acides nucléiques (ADN et ARN) sont des biopolymères formés de plusieurs nucléotides. On parle alors de polynucléotides ou, plus spécifiquement, d'oligonucléotides lorsqu'il y a moins de 25 nucléotides.

Les biopolymères peuvent avoir des caractéristiques qui diffèrent nettement de celles de leurs monomères (résidus) constitutifs. Par exemple, l'amidon, qui est une chaîne formée par l'assemblage de plusieurs unités de glucose, n'est pas soluble dans l'eau et n'est pas sucré, bien qu'il soit constitué de sucre [2]. C'est donc dire que la somme des caractéristiques de chaque constituant d'une biomolécule ne fournit pas nécessairement les caractéristiques globales de la biomolécule. La diversité des biomolécules est très grande. Les plus petites contiennent quelques atomes, alors que les plus grandes peuvent être formées de dizaines de milliers d'atomes.

La molécule d'ARN est une biomolécule qui est présente dans la cellule vivante. On peut facilement comparer la cellule à une gigantesque usine automatisée. À l'intérieur de la cellule, des structures fonctionnelles, appelées *organites*, jouent le rôle d'outils robotisés et les biomolécules, le rôle des différentes matières (matières premières, produits intermédiaires, produits finis, produits de transformation, déchets, etc). Cette usine fabrique elle-même sa propre machinerie moléculaire (les organites) et elle a la capacité de s'autoreproduire (la division cellulaire). Cette usine est de taille microscopique, la plupart des cellules mesurant entre 1 et 100 μm (micromètre) [11].

Lorsqu'il est question d'usine automatisée, nous, informaticiens, comprenons qu'il existe des éléments plus ou moins sophistiqués de programmation qui donnent les instructions de fonctionnement aux machines-outils de l'usine. Dans le cas de la cellule, c'est le code génétique qui joue le rôle de programmation. Souvent présentée comme le dogme central de la biologie moléculaire, l'information génétique qui est contenue dans l'ADN peut être répliquée (copiée) pour être utilisée dans une autre cellule ou bien transcrite sous une forme de messager, soit une molécule d'ARN appelée ARN messager (ARNm). Ce messager transporte ensuite l'information du code génétique vers une machinerie moléculaire spécifique, dénommée *ribosome*. Par un processus de traduction (décodage), le ribosome produit la protéine spécifiée au départ par l'ADN [11].

La figure 1.1 illustre le dogme central de la biologie moléculaire.

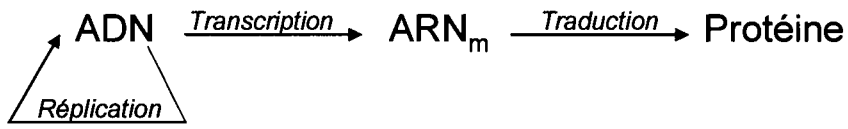

Figure 1.1 Dogme central de la biologie moléculaire
La molécule d'ADN peut se répliquer ou bien être transcrite sous la forme d'une molécule d'ARN messager (ARN_m) qui sera traduite sous la forme d'une protéine.

1.4 Les définitions structurales de l'ARN

L'ARN est un polynucléotide, c'est-à-dire, une longue chaîne formée de plusieurs nucléotides attachés les uns aux autres de manière contiguë. On identifie le début de la chaîne par la notation 5′ et la fin de la chaîne par la notation 3′. Dans le cas de l'ARN, chaque nucléotide est formé par la liaison de trois composantes : un groupement phosphate, un sucre et une base. La base peut être l'une des quatre bases suivantes : l'Adénine (A), l'Uracil (U), la Cytosine (C) ou la Guanine (G). Les lettres

A, U, C ou G sont utilisées pour identifier les différents nucléotides de l'ARN. Cette chaîne possède plusieurs angles de torsion [12], ce qui lui confère une grande flexibilité et la possibilité théorique d'adopter une très grande variété de formes dans un espace 3D. Toutefois, *in vivo*, cette longue chaîne de nucléotides va se recroqueviller sur elle-même pour former une structure 3D qui lui conférera sa forme et sa fonctionnalité biologique. Ce repliement (*folding*) est principalement causé par les forces d'attraction et de répulsion des molécules en présence. De plus, cette structure 3D n'est pas nécessairement statique. Elle peut subir des changements de forme durant les processus biologiques où elle est impliquée et osciller entre certaines formes.

L'étude de la structure de l'ARN se fait selon trois niveaux de définition. On les appelle structure primaire, structure secondaire et structure tertiaire [12]. Ultimement, ce qui intéresse les chercheurs, c'est la connaissance de la dynamique fonctionnelle des biomolécules. Puisque cette dernière est intimement liée à la composition et à la conformation 3D de la biomolécule, la détermination des structures primaire, secondaire et tertiaire fait partie des stratégies utilisées pour étudier la dynamique de l'ARN.

1.4.1 La structure primaire

Les nucléotides de l'ARN sont attachés les uns aux autres de manière contiguë. Ils sont reliés entre eux par des liaisons dites *liaisons covalentes*. Une liaison covalente existe quand deux atomes partagent une ou plusieurs paires d'électrons de valence, c'est-à-dire des électrons situés sur le dernier niveau énergétique [11]. Ce type de liaison est très fort et est responsable de la formation de la majorité des molécules. L'ordre dans lequel les nucléotides sont reliés de manière contiguë dans la chaîne est appelée la séquence d'ARN ou la structure primaire de l'ARN.

À chaque molécule d'ARN corrrespond une et une seule structure primaire. Par exemple, la molécule d'ARN identifiée 1BN0 dans la base de données PDB est composée de 20 nucléotides et sa structure primaire est la suivante :

$$^{5'}\text{GGACUAGCGGAGGCUAGUCC}^{3'}$$

1.4.2 La structure secondaire

En plus des liaisons covalentes, il existe aussi des liaisons dites *liaisons hydrogènes*. Ces liaisons sont le résultat de l'attraction d'un atome d'hydrogène chargé positivement avec un autre atome chargé négativement, souvent un atome d'oxygène ou d'azote. La liaison hydrogène est beaucoup plus faible que la liaison covalente et peut donc être brisée plus facilement. Elle est très présente dans les processus chimiques de la vie [11] puisqu'elle demande peu d'énergie pour se faire et se défaire.

Diverses liaisons hydrogènes peuvent se former entre les bases de la chaîne, engendrant un rapprochement entre les bases appariées et imposant des contraintes structurelles à l'ARN. Il existe plusieurs appariements possibles entre les bases. Les plus connus et les plus forts sont ceux de type Watson-Crick qui peuvent se former entre les bases G et C pour former une paire GC ou entre les bases A et U pour former une paire AU. On les appelle Watson-Crick puisqu'ils ont été identifiés pour la première fois par Watson et Crick. L'appariement GC, ou lien GC, est plus fort que l'appariement AU, ou lien AU. Auparavant, seuls les liens GC et les liens AU étaient considérés dans l'étude de la structure secondaire. Aujourd'hui, les études sur la structure secondaire intègrent aussi un appariement entre les bases G et U (paire GU). L'appariement GU, ou lien GU, est plus faible que le lien AU.

L'ensemble de ces différents liens entre les nucléotides de la chaîne est appelé la structure secondaire de l'ARN. Théoriquement, à partir d'une séquence d'ARN donnée, il est possible de former plus d'une structure secondaire. Par exemple, il

existe 4 476 combinaisons possibles de structures secondaires avec la séquence de 1BN0. Toutefois, *in vivo*, la molécule 1BN0 adopte principalement une seule structure secondaire et cette dernière est appelée la structure secondaire native. Une partie du problème de prédiction consiste donc à trouver la structure secondaire native parmi toutes celles possibles.

La figure 1.2 illustre trois structures secondaires possibles pour la molécule 1BN0. La structure I, qui est aussi la structure secondaire native, est composée de cinq liens GC et de trois liens AU. La structure II est composée de deux liens GC et de deux liens AU. Finalement, la structure III est composée de trois liens GC et d'un lien GU.

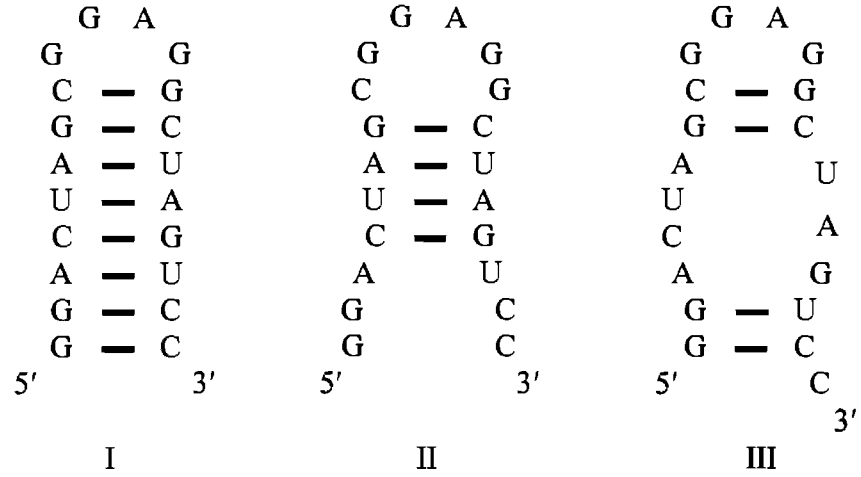

Figure 1.2 Trois structures secondaires possibles pour la molécule 1BN0
Les traits foncés représentent les liens entre les bases appariées. La structure I correspond à la structure secondaire native.

Par commodité, une structure secondaire est souvent représentée dans la littérature à l'aide de parenthèses, le pairage d'une parenthèse ouvrante avec une parenthèse fermante indiquant un lien secondaire entre deux nucléotides. Nous utiliserons aussi cette représentation et nous indiquerons l'absence de lien par le caractère deux-points.

Le tableau 1.1 montre la représentation, à l'aide des parenthèses, des structures secondaires I, II et III de la figure 1.2.

Structure	Représentation
primaire	GGACUAGCGGAGGCUAGUCC
I	((((((((::::)))))))))
II	:::(((((:::::::))))::::
III	((::::((::::)):::))::

Tableau 1.1 Représentation de structures secondaires à l'aide de parenthèses. La première ligne montre la structure primaire de 1BN0. Les lignes suivantes représentent les structures secondaires de la figure 1.2. Le pairage d'une parenthèse ouvrante avec une parenthèse fermante indique un lien entre les deux nucléotides et le caractère deux-points indique une absence de lien.

1.4.3 La structure tertiaire

La structure tertiaire correspond à la conformation 3D de la molécule d'ARN [10]. Les liens tertiaires sont des interactions ou liaisons, autres que primaires ou secondaires, qui existent entre différentes parties de la molécule et qui contribuent à la formation de la conformation 3D. La molécule d'ARN peut adopter différentes conformations 3D selon le processus biologique où elle est impliquée [10].

Afinsen [13] démontra en 1973, grâce à ses travaux sur la dénaturation des protéines (chaîne polypeptidique), qu'une protéine que l'on a déformée en laboratoire peut, dans une solution adéquate, se replier sur elle-même pour retrouver sa structure 3D initiale. Ainsi, la séquence de la structure primaire contient l'information suffisante pour retrouver la structure tertiaire. Il est donc clair qu'il existe des règles qui déterminent comment une chaîne d'acides aminés donnée va se replier sur elle-même pour former une structure 3D [14]. Toutefois, même après plusieurs décennies de recherche, cette dynamique de repliement est encore un sujet de recherche important, ce qui démontre bien la nature complexe du problème.

1.5 Introduction aux automates cellulaires

Tel que mentionné dans la portée de la recherche (section 1.2), nous avons choisi d'explorer une nouvelle approche basée sur l'auto-organisation pour développer un modèle permettant de simuler la dynamique moléculaire et de prédire les structures secondaires qui résultent du repliement de l'ARN. Notre modèle ayant été développé à l'aide d'une construction d'automates cellulaires, nous jugeons à propos de présenter ici une brève introduction à la théorie des automates cellulaires.

1.5.1 Historique

Les premiers automates cellulaires (AC) apparurent à la fin des années 1940 avec les travaux de Stanislaw Ulam et de John Von Neumann [15]. Ce dernier cherchait à modéliser les systèmes naturels complexes. Il travailla, entre autres, sur la possibilité de créer une machine ayant la capacité de se recréer elle-même. En 1952, il compléta la description d'un automate cellulaire capable de s'autoreproduire. Ces travaux furent publiés par Burks [16].

Puis, dans les années 1970, le concept d'automates cellulaires devint du domaine populaire suite à la publication, dans la revue *Scientific American*, d'un jeu réalisé à l'aide d'un AC à deux dimensions [17, 18]. Il s'agissait du jeu « *Life* » de John Conway, mathématicien à l'Université de Cambridge. Dans les années 1980, Stephen Wolfram démontra par une série de publications que les AC étaient beaucoup plus qu'une simple curiosité mathématique et qu'ils pouvaient servir à modéliser des systèmes physiques complexes [19].

Différents phénomènes naturels ont déjà été modélisés en utilisant l'approche AC. Par exemple, Lent et al. ont modélisé des interactions atomiques [20], Brewster a modélisé le flux de molécules liquides [21], et Wainer s'est attaqué à des systèmes environnementaux [22].

1.5.2 Définition

Les différents auteurs regroupent souvent les AC selon trois catégories. D'abord, les AC dits classiques, c'est-à-dire, ceux influencés par les travaux de Von Neumann. Ensuite, les AC dits modernes, soit ceux influencés par les travaux de Wolfram et les autres chercheurs des années 1970 et 1980. Finalement, la dernière catégorie regroupe les jeux à base d'AC.

En s'inspirant de Wolf-Gladrow [23] et Sarkar [15], un AC classique peut se définir ainsi : un AC est un arrangement régulier de plusieurs cellules[2] de même type; chaque cellule peut prendre un nombre fini d'états discret; les états de chaque cellule sont mis à jour simultanément à des intervalles de temps discret; les règles de mise à jour sont déterministes et uniformes dans le temps et l'espace; finalement, l'évolution d'une cellule est fonction de ses voisins seulement.

Illustrons cette définition à l'aide d'un exemple simple. Soit un tableau à une dimension qui est infini dans les deux directions. Chaque cellule en position i du tableau ne peut contenir que la valeur 0 ou bien la valeur 1. Au temps $t = 0$, le tableau est dans une configuration dite initiale. Au temps $t = 1$, le contenu de chaque cellule du tableau est recalculé à l'aide d'une fonction f dont les arguments sont la valeur actuelle de la cellule, la valeur de la cellule à sa gauche et finalement la valeur de la cellule à sa droite. Soit $q_i(t)$ l'état de la cellule i au temps t et $q_i(t+1)$ l'état de la cellule i au temps $t + 1$. Alors, on peut exprimer l'état d'une cellule au temps $t + 1$ avec l'équation suivante :

$$q_i(t+1) = f[q_i(t), q_{i+1}(t), q_{i-1}(t)]$$

La figure 1.3 illustre cet exemple lorsque la fonction f est égale à

$$q_i(t) \oplus (q_{i+1}(t) \oplus q_{i-1}(t))$$

où $\oplus$ est l'opérateur OU exclusif.

[2] Le terme « cellule » fait référence aux cellules d'un tableau.

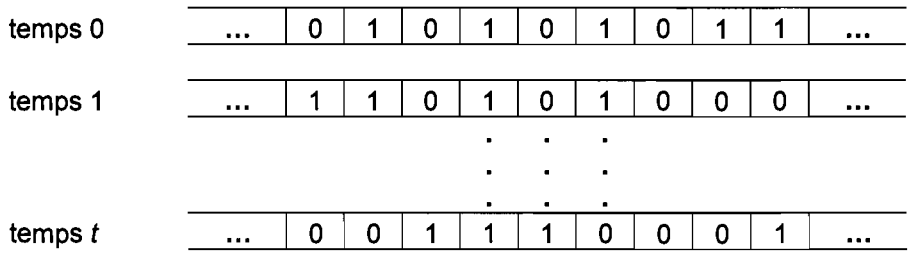

Figure 1.3 Évolution d'un automate cellulaire à une dimension

La fonction f qui calcule le nouvel état pour chaque cellule est appelée la *règle locale*, la *règle de l'AC* ou bien la *règle*. Lorsque l'AC n'utilise pas de données de l'extérieur en entrée (*input*), il est dit *autonome*. L'ensemble des états de toutes les cellules à un temps t quelconque est appelé une *configuration* ou bien un *état global* de l'AC. L'ensemble des états de toutes les cellules au temps $t = 0$ est appelé la *configuration initiale* ou bien l'*état global initial*. L'application de la règle locale à chaque cellule produit la transformation d'une configuration vers une autre. Cette transformation est appelée la *règle globale* ou la *carte globale* de l'AC [15, 23, 24]. Si on considère un voisinage de trois cellules (gauche, centre, droite) et la possibilité de deux états par cellule, il y a donc une possibilité de 2^3 i.e. huit configurations distinctes pour ce voisinage. À partir de ces huit configurations de voisinage, la cellule centrale peut prendre deux états distincts au temps $t + 1$. Donc, il existe une possibilité de 2^8 règles globales différentes.

La multiplicité des recherches effectuées sur les AC a généré une multitude de variantes à partir du modèle classique. Ces variantes se retrouvent dans la géométrie de l'espace utilisé qui peut être de différentes formes et de dimensions variées, fini ou infini. Elles se retrouvent aussi dans le type de règles locales utilisées, dans le choix des ensembles d'états des cellules, dans le choix du voisinage des cellules, et ainsi de suite.

Voici donc la définition de quelques variantes [15, 16, 19, 23, 24].

Définition 1. Si la règle locale est la même pour chaque cellule, alors l'AC est dit *uniforme* ou *homogène*. Sinon, il est dit *hybride* ou *hétérogène*.

Définition 2. Si l'ensemble des états possibles n'est pas le même pour chaque cellule, alors l'AC est dit à *plusieurs ensembles d'états*.

Définition 3. Si l'ensemble des cellules et des interconnexions entre les cellules est fixe, alors l'AC est dit *statique*. Sinon, il est dit *dynamique*.

Définition 4. Si le nombre de voisins en entrée est le même que le nombre de voisins en sortie, alors l'AC est dit *balancé*. Sinon, il est dit *non balancé*. La plupart du temps, c'est la géométrie elle-même qui détermine le balancement.

Définition 5. Si la règle locale peut varier dans le temps, alors l'AC est dit *programmable*. On utilise aussi le terme *automates mosaïques*. La plupart du temps, la réalisation de ce type d'AC se fait grâce à une ligne d'entrée distribuée sur chaque cellule. Cette entrée sert à choisir la règle qui sera appliquée parmi un ensemble fini de règles contenues dans la cellule.

1.5.3 Classification moderne

Dans les années 1980, le physicien Stephen Wolfram entreprit une analyse expérimentale des motifs générés par les AC. Il évalua plusieurs paramètres statistiques sur l'évolution des motifs dans l'espace et dans le temps afin de vérifier la capacité des AC à modéliser des systèmes physiques complexes. Entre autres, il observa un phénomène d'auto-organisation dans certains cas. En partant d'une configuration au hasard avec un maximum d'entropie, l'AC peut évoluer vers des configurations de moindre entropie, ce qui semble à priori contraire à la deuxième loi de la thermodynamique qui statue que des systèmes réversibles évoluent toujours vers des états d'entropie maximum. Ce comportement d'auto-organisation est en fait causé par l'irréversibilité microscopique de ces AC [15]. Les travaux de Wolfram ont démontré que dans certaines conditions, l'entropie part d'un maximum pour diminuer, tandis que dans d'autres cas, elle augmente. Il développa donc une

classification intuitive basée sur la mesure de l'entropie et qui est largement utilisée aujourd'hui.

Voici les quatre classes qualitatives proposées par Wolfram [19] :

(1) Les AC qui évoluent vers une configuration homogène.

(2) Les AC qui évoluent vers un ensemble diversifié de structures simples qui sont stables ou périodiques.

(3) Les AC qui évoluent vers des motifs chaotiques.

(4) Les AC qui évoluent vers des structures complexes bien localisées et qui sont parfois persistantes.

1.5.4 Phénomène d'auto-organisation

L'auto-organisation n'est pas un phénomène exclusif aux AC. Ce phénomène a d'abord été observé dans plusieurs systèmes naturels. En 1977, I. Prigogine obtint un prix Nobel de chimie pour avoir découvert que, malgré l'apparente contradiction avec la seconde loi de la thermodynamique, des systèmes physico-chimiques loin de l'équilibre thermodynamique ont tendance à s'auto-organiser en exportant de l'entropie et ainsi former ce qu'il a appelé des structures dissipatives [25]. Camazine et al. [26] définissent l'auto-organisation comme suit :

> « *Self-organization is a process in which pattern at the global level of a system emerges solely from numerous interactions among the lower-level components of the system* ».

L'émergence d'un ordre dans ces systèmes est un phénomène complexe qui intrigue les scientifiques de toutes les disciplines. La complexité provient, entre autres, du fait que de l'interaction entre les composantes de base peut émerger des propriétés globales non présentes dans les composantes. Ce qui intrigue particulièrement est l'observation que les règles entre les composantes peuvent être très simples, même lorsque les propriétés émergentes sont très sophistiquées [26].

2 Prédiction et dynamique moléculaire de l'ARN

La recherche en prédiction structurale de l'ARN peut être regroupée en deux grandes catégories, soit : la recherche qui s'intéresse à la prédiction de la structure secondaire à partir de la structure primaire; et la recherche qui s'intéresse à la prédiction de la structure tertiaire, généralement à partir de la structure secondaire. Dans les deux cas, les travaux peuvent viser trois objectifs : la prédiction statique de la structure finale; la prédiction de structures intermédiaires impliquées dans le repliement; ou la prédiction de la dynamique de repliement.

Notre recherche vise plus particulièrement la dynamique de repliement menant à la structure secondaire. La prédiction de la dynamique de repliement se fait actuellement à l'aide de méthodes de dynamique moléculaire. Ces méthodes seront présentées à la section 2.2.

2.1 Les grandes approches de prédiction

Il existe actuellement trois grandes approches au problème de la prédiction structurale, soit l'approche par heuristique, l'approche par homologie et l'approche énergétique [3, 27, 28]. Parmi ces trois grandes approches, seule l'approche énergétique est applicable à la prédiction de la dynamique de repliement. Les deux autres approches seront quand même présentées ici à titre informatif.

2.1.1 L'approche heuristique

L'approche heuristique restreint l'espace des solutions à l'aide des données déjà disponibles de manière à rendre la solution accessible par des méthodes traditionnelles. Ces données peuvent aussi bien être des données obtenues à partir des expériences en laboratoire que des données obtenues par d'autres méthodes. Cette approche est donc orientée vers la recherche de la structure qui satisfait le mieux les contraintes imposées par les données disponibles.

Il est important de souligner que les données obtenues à partir des expériences en laboratoire ont été, et sont encore aujourd'hui, une source importante d'information sur les structures primaire, secondaire et tertiaire de l'ARN. De plus, les données expérimentales servent à valider les résultats théoriques obtenus à partir des différentes méthodes de prédiction.

Cette approche est principalement utilisée pour la prédiction statique de la structure tertiaire. Le logiciel MC-Sym, développé par Major, est basé sur ce type d'approche [3, 27, 29 – 31]. De plus, MC-Sym intègre des algorithmes de résolution de problèmes par satisfaction de contraintes. L'inclusion de contraintes structurales permet de diminuer la taille de l'arbre de recherche en éliminant les sous-ensembles de structures qui sont incompatibles avec les contraintes.

2.1.2 L'approche par homologie

L'approche dite par homologie consiste à comparer la séquence de la molécule étudiée avec les séquences de molécules dont la structure secondaire est déjà connue. La prémisse de base est que des séquences semblables donneront des structures semblables. On effectue donc une analyse comparative de la séquence étudiée afin d'y retrouver des parties de séquence correspondant à des éléments de structure déjà répertoriés. Cette approche est aussi utilisée pour la prédiction statique de la structure tertiaire à partir de la connaissance de la structure secondaire. Il s'agit donc d'une approche orientée vers la recherche d'une structure similaire. Les méthodes utilisant cette approche produisent habituellement de très bonnes prédictions.

2.1.3 L'approche énergétique

L'approche la plus ancienne et la plus répandue est l'approche énergétique. Cette approche est basée essentiellement sur les lois de la thermodynamique. Ces dernières statuent que lorsqu'une molécule est stable, elle est aussi à son plus bas niveau d'énergie [3], c'est-à-dire que la molécule cherche à adopter une structure qui minimise une certaine fonction d'énergie. Il s'agit donc de définir cette fonction

d'énergie et d'utiliser une technique d'optimisation sur cette fonction. Plus la fonction d'énergie sera représentative de la réalité, meilleure sera la prédiction. Cette approche est donc orientée vers la recherche d'une structure bien précise, soit la structure présentant le minimum d'énergie.

Plusieurs méthodes, basées sur la programmation dynamique, utilisent l'approche énergétique pour faire de la prédiction statique. Le programme Mfold [32] en est un exemple. Ces méthodes ne sont applicables que pour la prédiction de la structure secondaire. Si on exclut la présence de pseudonoeuds, ces méthodes sont très efficaces pour trouver la structure secondaire présentant le minimum d'énergie. Par contre, la performance diminue grandement à mesure que les méthodes essaient d'intégrer différents pseudonoeuds.

L'approche énergétique est aussi utilisée pour la prédiction d'intermédiaires impliqués dans le repliement menant à la structure secondaire. Les deux méthodes les plus connues sont la méthode de repliement thermodynamique [33] et la méthode de repliement cinétique [34]. La première simule la dénaturation thermique d'une molécule dans l'hypothèse de faire ressortir des intermédiaires importants impliqués dans le repliement. La seconde effectue un repliement stochastique qui vise à faire ressortir les intermédiaires les plus fréquents parmi plusieurs simulations.

Finalement, l'approche énergétique est la seule approche utilisée pour la prédiction de la dynamique de repliement. Les méthodes actuellement utilisées sont des méthodes de dynamique moléculaire qui sont présentées à la section suivante.

2.2 La dynamique moléculaire

La dynamique moléculaire (DM) s'intéresse aux déplacements dans l'espace et dans le temps des atomes d'une molécule [12, 35]. La première application de la dynamique moléculaire à des macromolécules biologiques remonte en 1977 avec les

travaux de McCammon sur une petite protéine animale [36, 37]. Depuis ce temps, une multitude d'études et de travaux ont été accomplis dans le but de simuler la dynamique moléculaire de biomolécules et ainsi être capable de prédire leurs comportements et leurs conformations structurales.

Selon Van Gunsteren et al. [38], la simulation de la dynamique moléculaire nécessite d'abord de préciser les éléments de base suivants :

- Quelles seront les forces à retenir et celles à ne pas considérer? Par exemple, si la molécule est dans une solution, est-ce que l'on tiendra compte des forces externes provenant du solvant?
- Quelle sera la fonction d'énergie à utiliser? Quels seront les termes de la fonction d'énergie? On retrouve souvent les angles de liaison, les angles de torsion, les forces de Van Der Waals et les forces de Coulomb parmi les termes de la fonction d'énergie.
- Quelle sera l'équation utilisée pour simuler le mouvement? Par exemple, le mouvement peut être calculé en utilisant la loi de Newton, ou les équations de Lagrange, ou l'équation de Schrödinger en mécanique quantique, ou encore l'équation stochastique de Langevin.
- Quelles seront les limites spatiales ou thermodynamiques de la simulation? Est-ce que ces limites auront un effet de distorsion sur les atomes près des frontières?

Toujours selon Van Gunsteren et al. [38], plusieurs facteurs peuvent limiter la qualité de la simulation. Par exemple :

- les contraintes ou les restrictions imposées au système ont amené une distorsion du système ou d'un atome en particulier;
- les forces qui n'ont pas été prises en compte dans le modèle avaient un effet non négligeable;
- la fonction d'énergie ou la fonction de mouvement n'était pas de qualité suffisante pour reproduire le comportement de manière précise;
- la durée de la simulation était inadéquate.

On peut aussi introduire des contraintes externes sur le système afin de limiter les mouvements des molécules. Par contre, l'utilisation de contraintes ou de restrictions de nature spatiale ou thermodynamique peut avoir des effets distordants sur le comportement dynamique des atomes et sur les propriétés du système [38, 39].

2.2.1 Mécanique et dynamique moléculaire

La mécanique et dynamique moléculaire est le modèle classique de dynamique moléculaire. Ce modèle représente les noyaux atomiques et les électrons d'un atome comme une seule particule (sphère) qui possède une masse et une charge. L'interaction entre deux particules est basée sur les calculs classiques de champs électriques ou sur les calculs statistiques de la mécanique moderne. Le lien entre deux particules est considéré de longueur flexible à l'image d'un ressort. C'est l'ensemble des interactions entre les particules en présence qui détermine leurs positions relatives dans l'espace. La mécanique moléculaire considère donc la distance d'une liaison, l'angle de la liaison, les angles de torsion entre les particules et, finalement, les forces d'attraction ou de répulsion entre des particules non liées. Pour calculer ces déplacements, elle utilise la deuxième loi de Newton. La détermination d'une structure se base sur le principe que la molécule cherche à adopter une structure qui minimise une certaine fonction d'énergie. Habituellement, ces fonctions d'énergie sont calculées à l'aide d'équations différentielles ordinaires ou partielles et tiennent compte de la position 3D des particules, ainsi que des interactions entre les particules [3, 35, 39].

Actuellement, la mécanique et dynamique moléculaire est surtout limitée par la capacité de traitement. À cause des forces en jeu, les équations de mouvement doivent être intégrées sur de très petits intervalles de temps (environ 10^{-15} seconde). Dans un tel contexte, même avec la puissance de calcul des ordinateurs d'aujourd'hui, on ne peut simuler que des mouvements de quelques nanosecondes (10^{-9}) pour une protéine de grosseur moyenne. Pour simuler de plus longs mouvements, il faut donc simplifier le modèle tout en espérant ne pas perdre d'aspects physiques importants

[38, 40]. Les chercheurs font donc face au dilemme suivant : ils doivent simplifier les modèles de simulation pour obtenir des résultats significatifs avec de grosses molécules et, en même temps, ils doivent éviter la perte de précision engendrée par la simplification des modèles. Il existe plusieurs logiciels de simulation qui utilisent la mécanique et dynamique moléculaire. En voici quelques-uns : AMBER [41], FEDER/2 [42], FOCUS [43], GROMACS [44], MDScope [45], MOIL [46], nMOLDYN [47].

Pour l'ARN, la mécanique et dynamique moléculaire est surtout utilisée comme méthode de raffinement appliquée sur les résultats obtenus par les autres méthodes. Elle ne permet pas de simuler le repliement complet d'une molécule d'ARN.

2.2.2 Dynamique moléculaire avec modèles simplifiés

Dans la mécanique et dynamique moléculaire, les interactions sont exprimées comme des forces qui s'exercent en continu. Dans ce contexte, la résolution des équations de mouvements est l'une des étapes les plus exigeantes. Une des simplifications qui peut être apportée est l'utilisation de forces d'interactions par paliers (*step-wise potential*). L'utilisation de forces par paliers amène une discrétisation qui accélère de manière importante le calcul des mouvements. On parle alors de dynamique moléculaire discrète. Toutefois, même après ce gain de traitement, la dynamique moléculaire discrète est elle aussi limitée par la capacité de traitement.

Ding et Dokholyan [48] préconisent d'utiliser la dynamique moléculaire discrète avec une approche de modélisation dite « intuitive ». Cette approche vise à simplifier la représentation moléculaire en fonction des sujets d'intérêt. L'utilisation de modèles simplifiés avec la dynamique moléculaire discrète permet des simulations jusqu'alors inaccessibles.

En 2008, Ding et al. [49] ont appliqué la dynamique moléculaire avec modèles simplifiés à l'étude de molécules d'ARN. Ils ont modélisé un nucléotide à l'aide de trois sphères reliées entre elles, une pour le groupement phosphate, une pour le sucre

et une pour la base. La chaîne d'ARN est alors représentée par plusieurs de ces groupes de trois sphères. Les groupes sont attachés les uns aux autres par un lien entre la sphère phosphate et la sphère base. Le modèle inclut les appariements de bases, les empilements et d'autres types d'interactions dites *hydrophobiques*.

Une simulation commence avec la forme allongée de la molécule d'ARN et recherche la conformation 3D correspondant au minimum d'énergie. L'échantillonnage de l'espace des conformations est réalisé à l'aide d'un réseau de huit ordinateurs. L'utilisation d'un modèle simplifié conjugué à un échantillonnage en parallèle permet d'explorer rapidement l'espace des conformations. Une simulation de 2×10^6 unités de temps de simulation pour une molécule de 36 nucléotides demande environ cinq heures de temps de traitement avec huit ordinateurs.

Cette étude a simulé le repliement de 153 molécules d'ARN, contenant de 10 à 100 nucléotides, et a obtenu de très bons résultats avec 150 de ces molécules. Les auteurs notent que ces résultats n'ont été obtenus qu'après avoir corrigé un effet secondaire indésirable, induit par la modélisation, en introduisant un traitement stochastique lors du traitement des boucles. Cette étude est la première à simuler le repliement complet de la conformation 3D de molécules d'ARN sans avoir utilisé la connaissance préalable de la structure native.

2.2.3 Dynamique moléculaire avec automates cellulaires

D'autres chercheurs ont utilisé les automates cellulaires pour la modélisation d'interactions protéiniques [6 – 9] ou la modélisation de polymères [4, 5]. Ainsi, Ostrovsky et al. [4] ont utilisé un AC à deux dimensions pour étudier l'agrégation et la désagrégation d'un polymère. Shirvanyanz et al. [5] ont utilisé un AC à trois dimensions pour étudier l'auto-organisation d'un système comprenant plusieurs polymères similaires possédant de fortes zones d'attraction. Mais, à notre connaissance, il n'existe actuellement aucune modélisation avec automates cellulaires de la dynamique moléculaire de l'ARN.

Comme le décrit Kier [50], les AC sont des systèmes informatiques dynamiques, discrets dans le temps, l'espace et les états, dont le comportement global est uniquement généré par des règles locales. Les AC tentent de modéliser les propriétés émergentes d'un système dynamique complexe en substituant aux calculs numériques, souvent impraticables dans ces cas, des règles locales faciles à calculer. Les automates cellulaires sont donc un outil de choix pour explorer une approche par auto-organisation.

Dans les simulations de dynamique moléculaire actuelles, le cheminement dans l'espace des solutions est dirigé vers la structure d'énergie minimum. Pour ce faire, à chaque unité de temps, la simulation doit calculer une valeur d'énergie pour chacune des structures suivantes possibles. La structure ayant la plus faible valeur devient la structure suivante. La simulation se termine lorsqu'il n'est plus possible de trouver une structure avec une valeur d'énergie plus faible.

Une simulation par auto-organisation n'est pas orientée vers la recherche d'une structure particulière. Il n'y a pas de calculs à faire pour choisir la structure suivante. La simulation, ne possédant aucune information sur la structure native, ne peut se terminer seule. En fait, la structure native émerge d'elle-même après un certain temps. La difficulté majeure inhérente à cette approche est la détermination du modèle et des règles locales à utiliser pour recréer le processus d'auto-organisation.

3 Modélisation de l'ARN avec l'approche AC

Ce chapitre présente le modèle de prédiction structurale de biomolécules développé en utilisant une approche par automates cellulaires. La première partie de ce chapitre est consacrée aux descriptions formelles qui sous-tendent le développement du modèle. La seconde partie porte plus spécifiquement sur le modèle, les concepts qui lui sont propres, les règles locales appliquées et la paramétrisation utilisée.

3.1 Descriptions formelles

3.1.1 Description formelle de la structure primaire

Nous appellerons *lien covalent* le lien qui relie deux nucléotides de manière contiguë et nous représenterons formellement la structure primaire comme une chaîne de caractères de longueur n, où n est égal au nombre de nucléotides (nt) de l'ARN. Chaque caractère de la chaîne appartient à l'alphabet {G, C, A, U} et est numéroté séquentiellement de 1 jusqu'à n, la première position de la chaîne représentant le nucléotide en position 5′ de l'ARN et la position n de la chaîne celui de la position 3′ de l'ARN. Nous utiliserons la notation P_n pour signifier une séquence d'ARN de longueur n et $P(x)$ pour signifier le nucléotide à la $x^{ième}$ position.

3.1.2 Description formelle de la structure secondaire

Nous appellerons *lien secondaire* un appariement admis dans la structure secondaire et nous utiliserons la définition suivante de structure secondaire :

> Soit une structure primaire P_n, alors une structure secondaire S de P_n est définie comme étant un ensemble de paires (i, j) avec $i < j$ tel que les conditions suivantes sont satisfaites :
>
> pour n'importe quelles paires (i, j) et (i′, j′) avec $i \leq i'$
>
> 1) $i = i'$ si et seulement si $j = j'$
>
> 2) $i < j < i' < j'$ ou bien $i < i' < j' < j$

et pour toutes les paires (i, j)

3) $j - i > 3$

4) $P(i)$ forme un lien secondaire avec $P(j)$

La première condition signifie que chaque nucléotide ne peut former une paire qu'avec un seul autre nucléotide. La deuxième condition impose une restriction sur les liens possibles en interdisant ceux que l'on nomme pseudonoeud. Ces derniers sont alors considérés comme de niveau tertiaire même s'ils peuvent être de type Watson-Crick. La principale motivation à cette exclusion est de diminuer le niveau de complexité du problème de prédiction [51]. La troisième condition représente des contraintes stériques qui empêchent la formation d'un lien secondaire entre deux nucléotides trop près l'un de l'autre. Ici, la condition stipule qu'il doit y avoir un minimum de trois nucléotides entre deux nucléotides qui sont liés.

Soit une paire (i, j) appartenant à une structure secondaire S de P_n, si $P(i) = \{G\}$ et que $P(j) = \{C\}$, alors nous pourrons aussi représenter cette paire (i, j) par le symbolisme G=C ou i=j, et respectivement,

par C=G ou i=j si $P(i) = \{C\}$ et que $P(j) = \{G\}$,

par A-U ou i-j si $P(i) = \{A\}$ et que $P(j) = \{U\}$,

par U-A ou i-j si $P(i) = \{U\}$ et que $P(j) = \{A\}$,

par G~U ou i~j si $P(i) = \{G\}$ et que $P(j) = \{U\}$,

par U~G ou i~j si $P(i) = \{U\}$ et que $P(j) = \{G\}$.

3.1.3 L'espace des structures secondaires

L'ensemble de toutes les structures S compatibles avec une séquence P_n est appelé l'espace des structures secondaires de P_n, et sera identifié par C ou plus précisément C de P_n.

La cardinalité de C, $|C|$, est fonction du contenu de la séquence et croît de manière exponentielle avec la longueur de la séquence [52]. À titre d'exemple, pour la

séquence $(ACGU)_2$, i.e. ACGUACGU, il existe seulement quatre structures secondaires distinctes en accord avec la définition de la section précédente. Pour la séquence $(ACGU)_3$, il en existe 34 et avec la séquence $(ACGU)_{10}$, il est possible de former 157 795 462 structures secondaires différentes.

Il n'existe pas de formule mathématique qui donne la cardinalité exacte de C pour une séquence particulière. Il existe seulement des formules récursives qui calculent la cardinalité de C pour des classes de séquences. Par exemple, avec la formule récursive de Waterman et Smith [53] où les séquences sont classées selon la longueur, la cardinalité de C est la même pour la séquence $(AAAA)_5$ et la séquence $(AUAU)_5$. Or, en réalité, la séquence $(AAAA)_5$ qui ne contient que des nucléotides A, ne peut pas former de structures secondaires, tandis qu'il existe 5 985 possibilités avec la séquence $(AUAU)_5$.

Pour connaître la valeur réelle de $|C|$, il faut donc dénombrer, une à la fois, toutes les structures secondaires possibles. Ce dénombrement est réalisable avec de petites séquences mais il devient impraticable au fur et à mesure que la longueur de la séquence augmente. Plusieurs auteurs ont donc développé des estimateurs asymptotiques qui sont utilisés pour avoir un aperçu de la valeur de $|C|$ [53 – 59].

Un des estimateurs couramment utilisé est celui de Zuker et Sankoff [54] qui ajoute une approche stochastique afin de tenir compte de la composition de la séquence. La probabilité que deux bases quelconques forment une paire de bases est définie comme étant $p = 2(\,p(\mathrm{A})\,p(\mathrm{U}) + p(\mathrm{C})\,p(\mathrm{G}))$, où $p(\mathrm{A})$, $p(\mathrm{C})$, $p(\mathrm{G})$ et $p(\mathrm{U})$ sont respectivement les probabilités d'occurrence des nucléotides A, C, G et U dans la séquence. Pour une séquence de longueur n, l'estimateur du nombre de structures secondaires est alors $E(n) \sim hn^{-3/2}\alpha^n$

$$\text{où} \quad \alpha = \left(\frac{1+\sqrt{1+4\sqrt{p}}}{2}\right)^2 \qquad \text{et} \qquad h = \frac{\alpha\left(1+4\sqrt{p}\right)^{\frac{1}{4}}}{2\sqrt{\pi}\,p^{3/4}} \,.$$

Si les quatre types de nucléotides sont représentés dans les mêmes proportions à l'intérieur de la séquence, alors $p = 2(\ 0.25*0.25 + 0.25*0.25) = 0.25$, $\alpha = 1.866$, $h = 1.959$ et, dans ce cas particulier, $E(n) \sim (1.9n^{-3/2})1.8^n$. Précisons que cet estimateur, hormis qu'il suppose une fréquence identique des quatre types de bases dans la séquence, ne tient pas compte du contenu exact de la séquence. Par exemple, avec cet estimateur, la cardinalité de C pour la séquence AAUUGGCC est estimée à 13 structures secondaires alors qu'en réalité, aucune n'est possible vu la distance entre les nucléotides complémentaires. Il est à noter aussi que cet estimateur ne considère pas les liens GU.

Dans le cadre de notre étude, il était préférable d'utiliser la valeur réelle de $|C|$ plutôt qu'une valeur estimée afin d'évaluer correctement la proportion de l'espace C parcourue lors des simulations. Pour ce faire, nous avons programmé une fonction qui effectue le décompte, une structure à la fois, de toutes les structures possibles avec une séquence donnée. Avec un ordinateur personnel, cette fonction peut dénombrer jusqu'à un maximum de 500 millions de structures secondaires différentes, ce qui nous a permis d'obtenir la valeur réelle de $|C|$ pour plusieurs séquences étudiées.

Le tableau 3.1 présente la valeur exacte de $|C|$ pour quelques séquences de longueur n croissante. Ces séquences sont tirées de l'article de Cupal et al. [56] et nos décomptes sont identiques à ceux de l'article. Le tableau 3.1 présente également, à titre comparatif, la valeur obtenue avec l'estimateur de Zuker et Sankoff [54].

L'admission des pseudonoeuds dans les structures secondaires ne fait pas qu'augmenter la complexité du problème [51], elle augmente également la cardinalité de l'espace. L'espace des structures secondaires lorsque les pseudonoeuds sont admis sera identifié par C_{pk}. Le tableau 3.1 présente les cardinalités de C et C_{pk} pour quelques séquences. Ainsi, pour la séquence $(ACGU)_6$, la cardinalité de C_{pk} est de 82 792 077 soit plus de 4 000 fois supérieure à la cardinalité de C qui est de 20 183. On observe également que pour une valeur croissante de n, la cardinalité de C_{pk} croît

beaucoup plus vite que celle de C. La présence d'un astérisque signifie que la cardinalité de C_{pk} excède le maximum de 500 millions.

Séquence	n	Zuker	$\|C\|$	$\|C_{pk}\|$
$(ACGU)_2$	8	13	4	5
$(ACGU)_3$	12	84	34	214
$(ACGU)_4$	16	662	274	10 988
$(ACGU)_5$	20	5 740	2 298	815 156
$(ACGU)_6$	24	52 947	20 183	82 792 077
$(ACGU)_7$	28	509 439	183 430	*
$(ACGU)_8$	32	5 055 622	1 711 400	*
$(ACGU)_9$	36	51 370 766	16 298 434	*
$(ACGU)_{10}$	40	531 802 059	157 795 462	*

Tableau 3.1 Impact des pseudonoeuds sur l'espace C
C_{pk} est l'espace des structures secondaires avec possibilité de pseudonoeuds. La présence d'un astérisque indique que la cardinalité de C_{pk} excède le maximum de 500 millions. Dans la colonne Zuker se trouve la valeur obtenue avec l'estimateur de Zuker et Sankoff [54].

3.1.4 Le repliement

Le repliement d'une molécule d'ARN fait référence au processus par lequel une molécule d'ARN dénaturée va retrouver sa forme dite native. Dans notre modèle, le repliement est la succession, dans le temps, des structures secondaires formées avant d'atteindre la S_{native}, par exemple $S_0 \rightarrow S_1 \rightarrow S_2 \rightarrow \ldots \rightarrow S_{native}$ [3], où S_0 représente une structure secondaire complètement étirée. Il s'agit donc d'un chemin parcouru dans l'espace C.

La probabilité de trouver S_{native} en choisissant un chemin au hasard dans l'espace C devient extrêmement petite à mesure que $|C|$ augmente et le temps nécessaire pour trouver S_{native} devient alors démesurément grand [60]. La plupart des modèles de repliement vont utiliser les équations de la cinétique et/ou les équations de la

[3] L'utilisation de la flèche (→) indique le passage d'une structure vers une autre, sans aucun sens mathématique.

thermodynamique pour simuler le chemin de $S_0 \rightarrow S_{native}$. Ils vont aussi diminuer la taille de C en élaguant ou en ne conservant que certaines conformations du paysage énergétique (*energy landscape*), quitte à ne représenter que quelques étapes intermédiaires du cheminement, comme par exemple, $S_0 \rightarrow ? \rightarrow S_{50} \rightarrow ? \rightarrow S_{native}$.

3.2 Le modèle CA-RNA

Le modèle CA-RNA (*Cellular Automata – RNA*) utilise une représentation simplifiée de la chaîne d'ARN et un mouvement Brownien soumis à la contrainte de règles locales simples qui n'impliquent aucun algorithme d'optimisation. Le modèle n'effectue aucune partition ou réduction de l'espace des conformations secondaires. Il ne bénéficie d'aucune information lui permettant de circonscrire le problème ou de progresser vers la structure native. Une simulation ne possède aucune information sur la structure native et elle n'est même pas en mesure de la reconnaître lorsque cette dernière se présente.

3.2.1 L'environnement

En utilisant l'approche AC, nous avons programmé un modèle simulant le repliement de l'ARN. Plusieurs éléments ont été paramétrés afin d'en faire un meilleur outil de simulation. CA-RNA intègre une interface permettant de suivre l'évolution des différentes configurations adoptées par l'AC. Ce modèle a été programmé avec Visual Basic sur la plate-forme PC.

CA-RNA simule l'espace tridimensionnel à l'aide d'un treillis régulier à trois dimensions. Les dimensions du treillis sont 70 x 70 x 70, pour un total de 343 000 cellules. La taille du treillis a été choisie de façon arbitraire afin de fournir l'espace nécessaire pour le chargement de la forme étirée de courtes chaînes d'ARN et à son libre mouvement. Le modèle se limite aux structures secondaires de l'ARN, mais il utilise un treillis à trois dimensions parce qu'il représente mieux la réalité cellulaire qu'un treillis à deux dimensions.

Dans ce treillis, chaque cellule peut être vide ou bien occupée par un nucléotide. Si la cellule est occupée par un nucléotide, son type (A, U, C ou G) est alors conservé dans une variable de la cellule. Chaque cellule possède plusieurs variables, à savoir :

- le type du nucléotide qu'elle contient (typeN);
- la position de ce nucléotide dans la chaîne 5′ - 3′ (noNuc);
- la position du nucléotide précédent dans la chaîne 5′ - 3′ (noCov1);
- la position du nucléotide suivant dans la chaîne 5′ - 3′ (noCov2);
- la position, dans la chaîne 5′ - 3′, du nucléotide formant un lien secondaire avec le nucléotide contenu dans la cellule (noSec).

La configuration initiale est obtenue en chargeant une chaîne d'ARN, généralement sa forme étirée, dans l'espace du treillis et en initialisant les variables des cellules à partir des informations de la structure primaire de la chaîne d'ARN. La position occupée par un nucléotide est représentée par une sphère centrée aux coordonnées de la cellule. Le type du nucléotide est indiqué par un code de couleur. Ainsi, la sphère d'un nucléotide de type G est de couleur verte, celle d'un nucléotide de type C est rouge, celle pour le type A est bleue, et finalement, celle pour le type U est jaune. Il est également possible de visualiser les liens covalents ainsi que les liens secondaires.

La figure 3.1 présente la forme étirée de la séquence {GGA CUA GCG GAG GCU AGU CC} après son chargement initial dans le treillis. La figure ne présente qu'une petite portion à deux dimensions du treillis à trois dimensions.

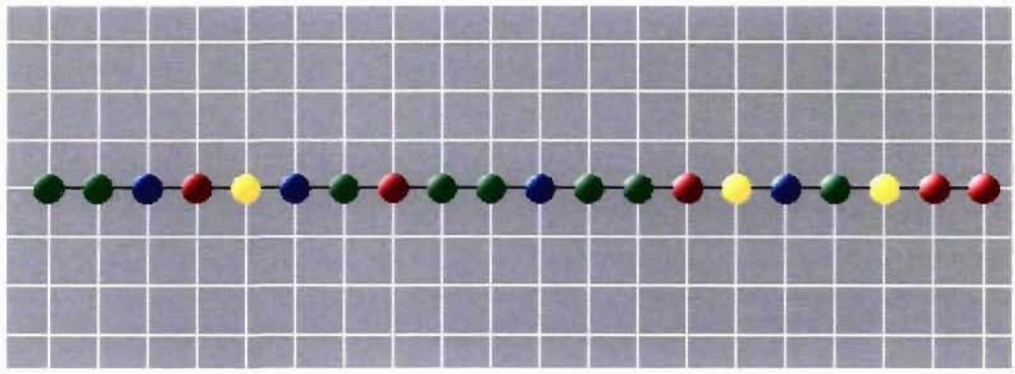

Figure 3.1 Chargement initial d'une séquence étirée
Les sphères vertes représentent les nucléotides G, les rouges les nucléotides C, les bleues les nucléotides A et les jaunes les nucléotides U. Les sphères sont reliées par un trait noir qui représente le lien covalent.

Une fois la configuration initiale établie, on peut lancer la simulation pour l'intervalle de temps souhaité et observer le repliement de la chaîne suite à la dynamique de création des liens secondaires. Il est possible de modifier certains paramètres, mais seulement avant la simulation. CA-RNA est un AC autonome et statique.

3.2.2 Définition des concepts et des paramètres

Exécution, simulation et unité de temps (laps)

Dans un AC, à chaque intervalle de temps discret, l'ensemble des cellules est parcouru dans un ordre prédéterminé et les règles locales sont appliquées sur chaque cellule, une à la fois. Lorsqu'il sera question d'un seul intervalle de temps, nous désignerons ce processus sous le terme d'exécution. Lorsque le processus se produira sur plusieurs intervalles de temps, nous parlerons alors de simulation. Une simulation est donc créée par plusieurs exécutions.

Nous utiliserons le terme *laps* pour désigner un intervalle de temps et nous utiliserons les abréviations usuelles K et M pour les nombres. Ainsi, une simulation de 500K laps signifiera une simulation d'une durée de 500 000 intervalles de temps.

La force relative de rétention (FRR)

Le développement de ce modèle a nécessité la mise au point d'un nouveau concept nommé « force relative de rétention ». Cette force relative de rétention se veut une représentation approximative de l'effet global des forces d'attraction et de répulsion impliquées au niveau moléculaire de l'ARN. À chaque type de lien secondaire est attribué une valeur de FRR. Par défaut, une valeur de 98 est attribuée aux liens GC (GC98), une valeur de 91 aux liens AU (AU91) et une valeur de 80 aux liens GU (GU80). Ces valeurs ont été déterminées empiriquement et la section 4.2.2 montre les résultats utilisés pour attribuer une valeur de 98 à la FRR des liens GC.

Les valeurs des FRR sont utilisées comme des probabilités, dans le temps, de stabilité des liens secondaires. Ces FRR permettent de travailler avec des règles locales simples.

Le vortex

Le vortex est un paramètre que nous avons nommé ainsi parce que son effet est un peu le même que celui d'un appareil vortex dans un laboratoire. Il augmente les interactions en favorisant le mouvement des particules. Il injecte, en quelque sorte, de l'énergie cinétique dans le système. Cet effet est obtenu par l'ajout d'une méta-règle locale qui stipule que, si après l'application des règles locales, la résultante est l'absence de mouvement du nucléotide, alors on recommence l'application, et ce, tant qu'il n'y aura pas de mouvement ou tant que le nombre maximum spécifié par le paramètre vortex n'aura pas été atteint. La valeur par défaut de ce paramètre a été fixée à 8, ce qui signifie que l'application des règles locales peut être répétée au maximum huit fois. Cette valeur a été déterminée empiriquement à l'aide des résultats présentés à la section 4.2.3.

La contrainte minimale de boucle (CMB)

La valeur du paramètre de contrainte minimale de boucle (CMB) est le nombre minimal de nucléotides qu'il peut y avoir entre deux nucléotides reliés par un lien secondaire. Le paramètre CMB correspond à la troisième condition de la définition

formelle d'une structure secondaire (section 3.1.2), soit celle représentant les contraintes stériques, qui stipule qu'il doit y avoir un minimum de trois nucléotides entre deux nucléotides qui sont liés. La valeur par défaut de CMB est donc 3.

Taux d'émergence (TE), structure dominante et structure émergente

Le taux d'émergence correspond au pourcentage d'occurrences d'une structure secondaire. Par exemple, un TE de 75% signifie que la structure secondaire est apparue dans 75% du temps. À chaque exécution, CA-RNA calcule un TE pour chacune des structures secondaires rencontrées. Une structure dominante est une structure ayant présenté le TE le plus élevé à un moment quelconque de la simulation. La structure émergente est celle ayant le TE le plus élevé à la fin de la simulation.

3.2.3 Les règles locales

CA-RNA est un AC de type hétérogène. Dépendamment du type de nucléotide qu'une cellule contient et si un lien secondaire existe avec ce nucléotide, des règles différentes seront appliquées sur la cellule.

Les règles de CA-RNA peuvent être classées en deux grandes catégories. Il y a celles qui affectent principalement la mobilité de la chaîne d'ARN et celles qui affectent principalement les liens secondaires entre les nucléotides. Dans les deux cas, ce sont des règles locales simples qui n'impliquent aucun algorithme d'optimisation.

Règle 1. Le mouvement Brownien (*Brownian*)

Le mouvement d'un nucléotide est considéré comme aléatoire. La dynamique Brownienne est souvent utilisée à l'intérieur de différentes approches [61 – 66]. Cela permet une simplification importante des calculs nécessaires pour déterminer la prochaine position d'une particule. Dans CA-RNA, cette position ne peut être que contigüe à la position actuelle et toutes les positions contigües sont, à priori, acceptables. Dans un treillis 3D, cela signifie que le treillis local des déplacements est

un treillis FCC « *face-centered-cubic* » [67, 68]. Les autres règles ajoutent des restrictions au déplacement Brownien.

Règle 2. L'exclusion de volume - collision

Un nucléotide ne peut se déplacer que vers une cellule libre.

Règle 3. L'exclusion de volume - voisinage

La distance entre deux positions dans le treillis est calculée en utilisant la formule de Thebyshev, soit $D_{Theb} = max(|x_2 - x_1|, |y_2 - y_1|, |z_2 - z_1|)$. Deux nucléotides doivent toujours être éloignés d'une distance $D_{Theb} \geq 3$. Font exception à cette règle, les paires de nucléotides reliés par covalence, les deux nucléotides reliés par covalence à un même nucléotide, les paires de nucléotides reliés par un lien secondaire et finalement les nucléotides dits délinquants (voir règle 6). Dans ces cas, les nucléotides doivent être éloignés d'une distance $D_{Theb} \geq 2$.

Règle 4. L'attraction – lien covalent

La distance D_{Theb} entre deux nucléotides reliés par un lien covalent doit toujours être égale à un. Les liens covalents sont définis lors du chargement de la configuration initiale et sont immuables.

Règle 5. L'attraction – lien secondaire

Lorsque deux nucléotides complémentaires sont à une distance $D_{Theb} \leq 3$, et qu'ils n'ont aucun lien secondaire, un lien secondaire est alors établi entre les deux, pourvu que les conditions qui définissent un lien secondaire soient satisfaites. Lorsqu'un tel lien existe, une nouvelle contrainte est ajoutée au déplacement des deux nucléotides impliqués. Toutefois, l'application de cette contrainte obéit à une probabilité dans le temps. Il s'agit de la force relative de rétention (FRR). À chaque type de lien secondaire est attribuée une valeur de FRR. La contrainte ajoutée stipule que la distance D_{Theb} entre les deux nucléotides ne doit pas augmenter. Lorsque la contrainte n'est pas appliquée, la distance D_{Theb} peut augmenter. Si la distance D_{Theb} devient plus grande que trois, le lien secondaire est brisé.

Le lien secondaire est donc modélisé comme une force d'attraction, élastique et pouvant se briser. Il peut présenter des forces d'attraction différentes selon le type du lien secondaire. Comme la règle 5 varie dans le temps, CA-RNA est donc un AC programmable.

Règle 6. L'empilement

Les liens secondaires peuvent être stabilisés par d'autres interactions atomiques. On appelle ce phénomène l'empilement (*stacking*) des paires de bases [2]. L'empilement se produit entre les bases adjacentes de deux liens secondaires voisins. Il a pour effet, entre autres, de compacter la molécule, ce qui augmente sa stabilité. La règle 6 tente de reproduire, en partie seulement, cet effet de tassement en autorisant un plus grand rapprochement entre les nucléotides impliqués dans deux liens secondaires voisins. Cette règle autorise une distance $D_{Theb} \geq 2$, au lieu de $D_{Theb} \geq 3$, entre les nucléotides impliqués.

L'utilisation de cette règle peut engendrer la stagnation d'un nucléotide. Cela se produit lorsque les nucléotides sont à une distance D_{Theb} égale à deux, suite à une situation d'empilement, et qu'un des deux liens secondaires se brise. Un nucléotide peut alors se retrouver dans une position où tous ses mouvements subséquents seraient normalement fautifs par rapport à la règle 3. Ce nucléotide est alors considéré comme délinquant par la règle 3, ce qui permet d'éviter tout effet de stagnation permanente.

3.2.4 La simulation

Lors d'une simulation, il n'y a aucune destruction ou création de nucléotides. Le nombre de nucléotides est fixe et il occupe une infime partie des 343 000 cellules du treillis. Puisque les règles locales ne concernent que les cellules contenant un nucléotide, nous avons implanté, dans CA-RNA, une liste des cellules non vides. Ainsi, au lieu de parcourir les 343 000 cellules à chaque exécution, nous parcourons seulement les cellules pointées par la liste. L'ordre de cette liste correspond à la

séquence 5′ - 3′ de la chaîne d'ARN. Pour chaque temps *t* pair, la liste est parcourue dans l'ordre, et pour chaque temps *t* impair, la liste est parcourue en ordre inversé. Cela évite l'introduction d'une régularité dans le processus stochastique. Un accès aléatoire à la liste aurait été préférable, mais aussi un peu plus exigeant en temps de traitement.

Certains paramètres peuvent être modifiés avant une exécution. Il s'agit principalement de la valeur du vortex, de la valeur de la CMB, des FRR pour chaque type de lien secondaire admis et de l'admissibilité ou non des pseudonoeuds comme lien secondaire.

Par défaut, une exécution utilise les paramètres suivants :

- Les pseudonœuds ne sont pas considérés.
- La CMB est de 3 nucléotides.
- Seulement les liens GC, AU et GU sont admis et les valeurs de FRR sont respectivement GC98, AU91 et GU80.
- La valeur du vortex est fixée à 8 fois.

Les valeurs des FRR et du vortex ont été déterminées de façon empirique. Le choix de ces valeurs sera discuté plus loin. Pour toutes les simulations décrites, s'il n'est pas fait mention de la valeur de ces paramètres, il faut tenir pour acquis que les valeurs par défaut ont été utilisées.

Après chacune des exécutions, les résultats sont compilés et la structure secondaire peut être affichée. Pour une simulation donnée, on peut donc connaitre toutes les structures secondaires qui sont apparues, dans quel ordre, à quel moment et pendant combien de temps. Ces données permettent de calculer le TE de toutes les structures rencontrées. Il est important de noter que ce sont les structures secondaires qui sont comptabilisées et non pas les différentes conformations spatiales d'une même structure secondaire.

La procédure est toujours la même pour chaque simulation. D'abord, CA-RNA est initialisé en positionnant la forme étirée de la séquence d'ARN dans le centre du treillis. Puis, la durée souhaitée pour la simulation est indiquée et le programme est lancé. Pour une même séquence, plusieurs simulations ont été faites en utilisant les mêmes valeurs de paramètres. Chaque simulation est identifiée par son nom de séquence suivi d'un numéro (ex. 1BN0#8).

3.3 Analyse du programme

3.3.1 Les principaux algorithmes

Les principaux algorithmes utilisés dans CA-RNA sont des algorithmes d'automates cellulaires et sont donc très simples. Le premier algorithme contient une boucle qui exécute l'automate cellulaire le nombre de fois déterminé pour la simulation. L'algorithme est le suivant :

Algorithme SIMULATION
Initialiser les paramètres
Initialiser les cellules de l'AC avec la structure primaire de l'ARN
totalLaps = la durée totale de la simulation
Pour *laps* = 1 jusqu'à *totalLaps*
    Faire EXECUTION_AC
    Afficher la nouvelle configuration de l'AC
    Comptabiliser la structure secondaire présente dans le treillis
Prochain *laps*

Le deuxième algorithme concerne l'exécution proprement dite de l'automate cellulaire. Comme il est expliqué à la section 3.2.4, au lieu de parcourir les 343 000 cellules du treillis à chaque exécution de l'automate cellulaire, seules les cellules occupées par la chaîne de nucléotides sont parcourues. L'algorithme est le suivant :

Algorithme EXECUTION_AC
Faire pour chaque nucléotide (nt) de la chaîne (en commençant par 5' lorsque *laps* est pair et 3' lorsque *laps* est impair)
    xyz = 0 et *vortex* = 8
    Faire tant que *xyz* = 0 et que *vortex* > 0
        vortex = *vortex -1*
        xyz = Calculer une nouvelle position pour le nt (règle 1)
        SI *xyz* est invalide selon les règles 2,3,4,5 ou 6 ALORS *xyz* = 0
    FinFaire
    SI *xyz* <> 0 ALORS déplacer le nt à la position *xyz*
    Faire MAJ_LIEN
FinFaire

L'algorithme montre que le temps d'exécution est fonction du nombre de nucléotides et que la valeur du paramètre vortex peut affecter ce temps d'exécution par un facteur, dans le pire des cas, correspondant à la valeur du vortex. Le calcul et la validation d'une nouvelle position font référence aux règles locales décrites à la section 3.2.3. Ces règles explorent le voisinage d'un seul nucléotide, ce qui n'ajoute pas à l'ordre de grandeur du temps d'exécution.

La création et la destruction des liens secondaires obéissent à la règle 5 et la mise à jour est effectuée selon l'algorithme suivant :

Algorithme MAJ_LIEN
SI le nt possède déjà un lien secondaire ALORS
    SI la distance D_{theb} > 3 (règle 5) ALORS
        Défaire le lien
    SINON
        ntMeilleur = Chercher le plus court lien possible (règle 5)
        SI *ntMeilleur* a déjà un lien ALORS
            SI le lien possible entre le nt et *ntMeilleur* est plus court que le lien actuel de nt et aussi plus court que le lien actuel de *ntMeilleur*
            ALORS
                Défaire le lien de ntMeilleur
                Remplacer le lien de nt
            FINSI
        SINON
            SI le lien possible entre le nt et *ntMeilleur* est plus court que le lien actuel de nt ALORS remplacer le lien de nt
        FINSI
    FINSI

```
FINSI
SI le nt n'a pas de lien secondaire ALORS
    ntMeilleur = Chercher le plus court lien possible (règle 5)
    SI ntMeilleur a déjà un lien ALORS
        SI le lien possible entre le nt et ntMeilleur est plus court que le
        lien actuel de ntMeilleur ALORS
            Défaire le lien de ntMeilleur
            Créer un lien entre le nt et ntMeilleur
        FINSI
    SINON
        Créer un lien entre le nt et ntMeilleur
    FINSI
FINSI
```

Cet algorithme n'ajoute pas à l'ordre de grandeur du temps d'exécution puisque la règle 5 ne fait qu'explorer le voisinage d'un nucléotide.

En conclusion, le temps d'exécution du programme est dans **O**(n), i.e. d'ordre linéaire et fonction du nombre de nucléotides. La taille du treillis n'a pas d'incidence sur le temps d'exécution. L'allocation mémoire du programme est de l'ordre d'une constante qui est déterminée selon les dimensions du treillis. Le programme comptabilise, sur espace disque, la fréquence de chacune des structures secondaires rencontrées. Il est clair que le nombre de structures rencontrées sera influencé par la longueur de la séquence, mais il est impossible de préciser un ordre de grandeur.

3.3.2 Durée d'une simulation et temps de traitement

Une simulation peut avoir deux objectifs, soit : simuler jusqu'à ce qu'un état d'équilibre donné soit atteint; ou simuler jusqu'à ce qu'une structure secondaire particulière soit atteinte. Selon la molécule, le premier objectif peut nécessiter de plusieurs minutes à plusieurs heures de traitement, tandis que le deuxième objectif peut s'effectuer en quelques secondes ou quelques minutes seulement.

Le modèle ne permet pas actuellement de déterminer si un état d'équilibre est atteint. La durée de la simulation doit donc être spécifiée au départ de la simulation. La

plupart des simulations présentées dans cette recherche sont d'une durée de 500K laps. Pour plusieurs molécules, l'état d'équilibre a été atteint beaucoup plus tôt dans la simulation. Toutefois, puisqu'une durée trop courte peut faire manquer l'état d'équilibre, nous avons toujours opté pour des durées amplement suffisantes.

Le temps réel de traitement pour une simulation est d'ordre linéaire et fonction du nombre de nucléotides. Par exemple, sur un ordinateur personnel, une simulation de 500K laps prend environ 40 minutes pour 10 nucléotides, 80 minutes pour 20 nucléotides et 160 minutes pour 40 nucléotides. Une simulation de 1M laps prend environ le double de temps.

La valeur du paramètre vortex peut également augmenter le temps réel de traitement de façon importante. Dans le pire des cas, le temps réel de traitement est augmenté d'un facteur correspondant à la valeur du vortex.

4 Résultats et discussions

Nous avons effectué plusieurs simulations en utilisant 20 molécules constituées de 10 à 48 nucléotides, dont une molécule avec pseudonoeuds, sept molécules formées de deux brins et une molécule formée de trois brins. Les résultats sont présentés et discutés dans ce chapitre.

Pour chacune des 20 molécules utilisées, on retrouve en annexe la structure émergente obtenue par la simulation. Les figures sont des impressions d'écran de CA-RNA. Cette application a été développée de façon à pouvoir suivre à l'écran le mouvement de la chaîne, la formation des liens et le repliement. L'affichage sur écran se fait en trois dimensions. L'impression des figures en annexe étant en deux dimensions, il peut arriver que l'effet de profondeur ne soit pas toujours bien rendu.

4.1 La structure émergente

Les premières simulations ont été réalisées en utilisant la séquence identifiée 1BN0 dans la base de données PDB. Sa longueur est de 20 nt et ses différents niveaux structuraux sont très bien connus. Sa structure P_{20} est {GGA CUA GCG GAG GCU AGU CC} et sa structure S_{native} est {1=20, 2=19, 3-18, 4=17, 5-16, 6-15, 7=14, 8=13}. La cardinalité de son espace C est de 4 476 structures. La première simulation sera dénommée 1BN0#1. Pour chaque structure, les liens de la S_{native} sont présentés entre crochets.

Le tableau 4.1 énumère les structures secondaires qui ont obtenu un TE $\geq$ 1% après une simulation d'une durée de 500K laps.

Rang	TE %	Structure secondaire
R1	69.0	[1=20], [2=19], [3-18], [4=17], [5-16], [6-15], [7=14], [8=13]
R2	19.7	[1=20], [2=19], [3-18], [4=17], [5-16], [6-15], [7=14], 8=12
R3	4.4	4=10, 5~9, 12=20, 13=19

Tableau 4.1 Taux d'émergence pour la simulation 1BN0#1
Simulation de 500K laps. Les liens entre crochets sont ceux de la S_{native}.

La structure au premier rang (R1), illustrée à la figure A.1, est la S_{native} avec un TE de 69.0%. Elle se distingue nettement des autres, la deuxième (R2) ayant un TE de 19.7%. À partir d'un AC stochastique, sans aucune connaissance de la S_{native}, sans aucune partition ou réduction de l'espace C, sans aucune directive de cheminement et sans aucun algorithme d'optimisation, sur une possibilité de 4 476 structures, c'est la S_{native} qui émerge et de manière significative.

CA-RNA étant un programme stochastique, on peut s'interroger quant à savoir si l'émergence de la S_{native} est une résultante de l'effet du hasard. Les tableaux 4.2 à 4.5 énumèrent les structures secondaires qui ont obtenu un TE $\geq$ 1% pour quatre autres simulations identiques de 500K laps effectuées avec 1BN0. Les résultats obtenus sont similaires. À chaque fois, la S_{native} émerge de manière significative. Plus encore, chaque simulation présente la même structure en R2 et les TE observés sont de même grandeur. Ainsi, l'émergence de la S_{native} n'est pas juste un coup de chance.

Rang	TE %	Structure secondaire
R1	75.1	[1=20], [2=19], [3-18], [4=17], [5-16], [6-15], [7=14], [8=13]
R2	19.6	[1=20], [2=19], [3-18], [4=17], [5-16], [6-15], [7=14], 8=12

Tableau 4.2 Taux d'émergence pour la simulation 1BN0#2
Simulation de 500K laps. Les liens entre crochets sont ceux de la S_{native}.

Rang	TE %	Structure secondaire
R1	74.3	[1=20], [2=19], [3-18], [4=17], [5-16], [6-15], [7=14], [8=13]
R2	19.1	[1=20], [2=19], [3-18], [4=17], [5-16], [6-15], [7=14], 8=12
R3	1.2	[1=20], [2=19], [3-18], 4=10, 5~9, 11-15

Tableau 4.3 Taux d'émergence pour la simulation 1BN0#3
Simulation de 500K laps. Les liens entre crochets sont ceux de la S_{native}.

Rang	TE %	Structure secondaire
R1	74.5	[1=20], [2=19], [3-18], [4=17], [5-16], [6-15], [7=14], [8=13]
R2	19.1	[1=20], [2=19], [3-18], [4=17], [5-16], [6-15], [7=14], 8=12
R3	1.5	4=10, 5~9, 12=20, 13=19

Tableau 4.4 Taux d'émergence pour la simulation 1BN0#4
Simulation de 500K laps. Les liens entre crochets sont ceux de la S_{native}.

Rang	TE %	Structure secondaire
R1	78.4	[1=20], [2=19], [3-18], [4=17], [5-16], [6-15], [7=14], [8=13]
R2	16.2	[1=20], [2=19], [3-18], [4=17], [5-16], [6-15], [7=14], 8=12

Tableau 4.5 Taux d'émergence pour la simulation 1BN0#5
Simulation de 500K laps. Les liens entre crochets sont ceux de la S_{native}.

Le tableau 4.6 montre le nombre de structures pour les cinq simulations. On y retrouve le nombre de structures secondaires différentes rencontrées avant d'atteindre la S_{native} pour la première fois et le nombre total de structures secondaires différentes rencontrées durant toute la simulation.

	1BN0#1	1BN0#2	1BN0#3	1BN0#4	1BN0#5
Nombre de structures secondaires différentes rencontrées avant d'atteindre la S_{native}	85	53	54	81	54
Nombre de structures secondaires différentes rencontrées durant toute la simulation	180	132	123	131	133

Tableau 4.6 Nombre de structures dans les simulations 1BN0#1 à 1BN0#5
Les simulations 1BN0#1 à 1BN0#5 sont des simulations de 500K laps.

Malgré un espace de 4 476 structures possibles, on constate que pour chaque simulation, il a suffi de parcourir seulement quelques dizaines de structures dans C pour trouver la S_{native}. Puisque cela représente à peine 2% de C, cela démontre que

l'émergence observée n'est pas le résultat d'une quelconque distribution de probabilité qui aurait été induite par les FRR. Cette émergence correspond plutôt à un phénomène d'auto-organisation qui tend vers la S_{native}. Cela concorde avec la quatrième classe de la classification de Wolfram où les AC, suite à un phénomène d'auto-organisation, évoluent vers des structures complexes bien localisées.

La figure 4.1 permet d'avoir une vue d'ensemble de la distribution des TE obtenus pour toutes les structures de la simulation 1BN0#1. L'utilisation d'un seul graphique avec échelle logarithmique n'offrant pas une représentation adéquate des résultats, cette figure se compose donc de trois graphiques utilisant des échelles différentes. Les structures sont numérotées par ordre de leur première apparition et la S_{native} est la structure numéro 86, soit S_{86}.

Même si le TE de la S_{native} est plus de 100 fois supérieur à la presque totalité des autres structures, il est intéressant d'observer que la distribution n'est pas homogène et qu'il y a un regroupement autour de la S_{86} et un autre autour de la S_{18}, cette dernière étant la structure en R3 du tableau 4.1.

Le système évolue donc vers un état d'équilibre autour de la S_{native}. Afin de vérifier si cet état d'équilibre est persistant, une autre simulation a été faite, mais cette fois-ci, pour une durée 20 fois plus longue, soit 10M laps. La figure 4.2 montre l'évolution, sur 10M laps, de S_0, S_{native}, S_{R2} ainsi que le nombre total de structures. L'échelle du temps est logarithmique.

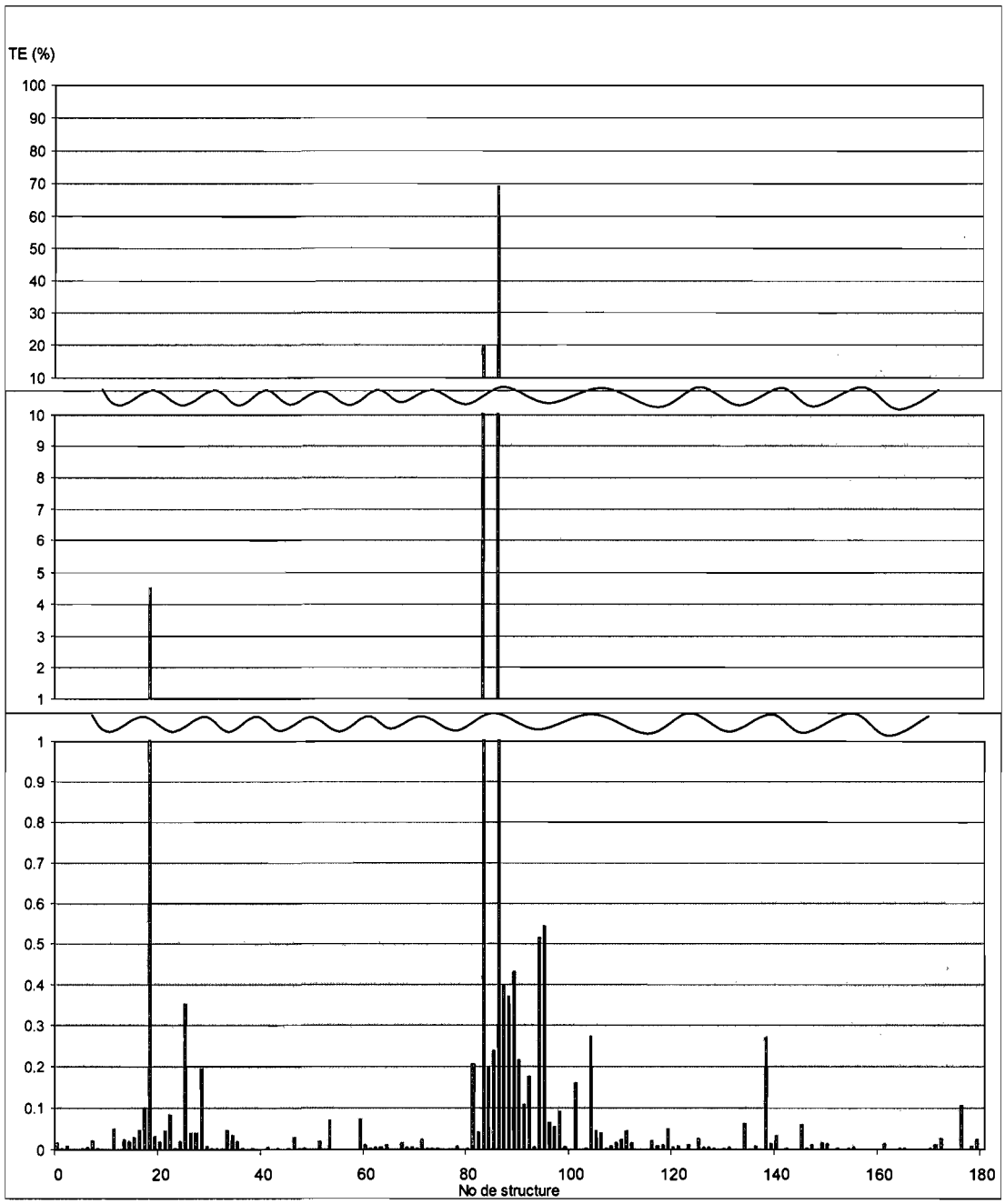

Figure 4.1 Distribution des TE pour la simulation 1BN0#1
Le graphique utilise trois échelles différentes. Les structures sont numérotées par ordre de leur première apparition et la S_{native} est la structure no 86.

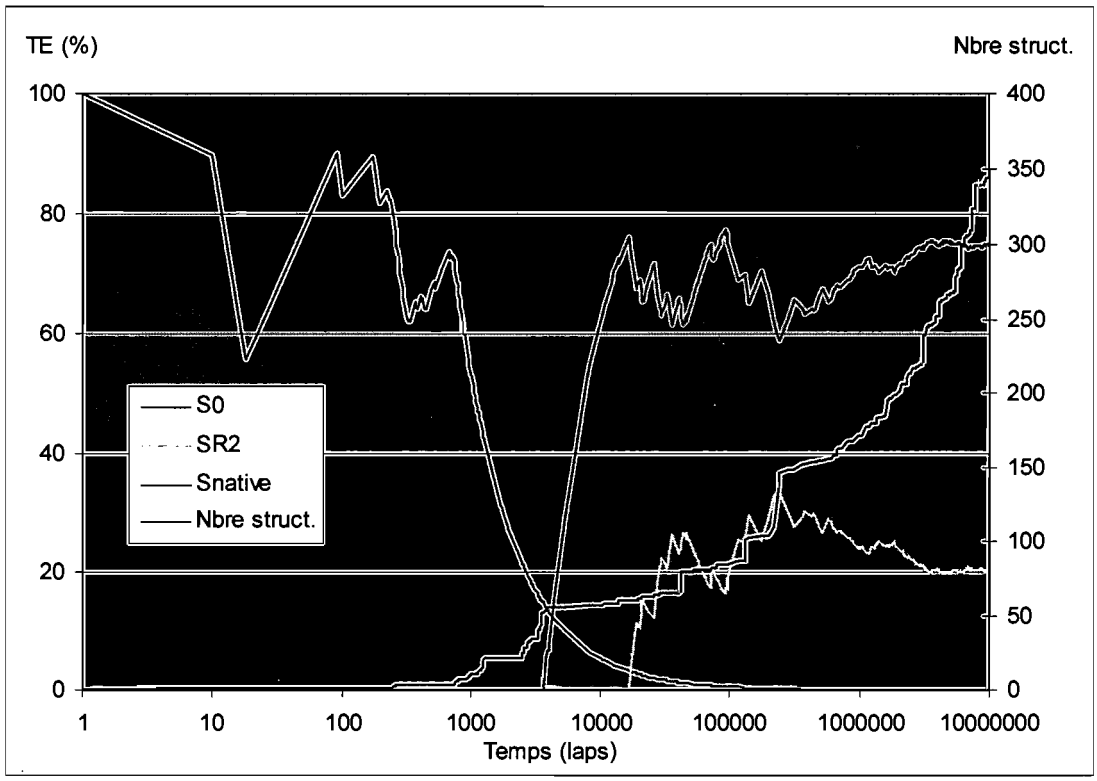

Figure 4.2 Évolution dans le temps des TE et du nombre de structures Simulation de 10M laps avec 1BN0.

Évidemment, la S_0 est la structure qui domine au début puisqu'elle correspond à la configuration initiale. Mais, avec le temps, la S_{native} devient la structure dominante et son TE atteint un plateau. Même la structure au deuxième rang, S_{R2}, qui n'est jamais dominante, atteint aussi un plateau. Cet équilibre relatif persiste jusqu'à la fin de la simulation, même si le nombre de structures continue à augmenter dans le temps. Cette augmentation est conséquente du fait que le système est dynamique et bouge continuellement. Ce que nous obtenons est un état d'équilibre autour de la S_{native} et non pas une S_{native} figée. Les structures se font et se défont, incluant la S_{native}. Par contre, avec le temps, le système revient toujours vers la S_{native} et préfère y demeurer plus longtemps que n'importe quelle autre structure. Cette particularité est intéressante, puisqu'elle indique que l'état d'équilibre semble indépendant de la configuration initiale.

4.2 Effets des paramètres

4.2.1 Les pseudonoeuds

Nous avons exclu les pseudonoeuds de nos simulations uniquement afin de permettre la comparaison avec les autres approches. En effet, les pseudonoeuds sont généralement exclus des structures secondaires parce qu'ils augmentent de façon très importante la cardinalité de C et la complexité du traitement mathématique [69 – 73]. Dans CA-RNA, l'ajout des pseudonoeuds ne nécessite aucune modification de traitement. Il s'agit seulement d'enlever une contrainte. L'exécution obéit aux mêmes règles locales, une contrainte en moins.

Le tableau 4.7 compare, pour une même séquence, les moyennes des résultats obtenus pour cinq simulations sans pseudonoeuds (1BN0#1 à 1BN0#5) par rapport à cinq autres simulations où les pseudonoeuds ont été admis (1BN0#7 à 1BN0#11). L'admission des pseudonoeuds fait passer la $|C|$, pour 1BN0, de 4 476 à 1 237 420 structures. Même après l'ajout des pseudonoeuds, la S_{native} demeure toujours la structure émergente. La seule différence notable est l'augmentation du nombre de structures rencontrées qui s'accroit d'un facteur de quatre, ce qui est faible comparativement à l'augmentation, d'un facteur supérieur à 250, du nombre de structures possibles.

Simulations	**Cardinalité de l'espace C**	**TE moy. de la S_{native}**	**Nbre moy. de struct. avant S_{native}**	**Nbre total moy. de structures**
1BN0#1 à 1BN0#5	4 476	74.26%	65.4	139.8
1BN0#7 à 1BN0#11	1 237 420	67.12%	262.8	566.2

Tableau 4.7 Résultats avec et sans pseudonoeuds
Les simulations 1BN0#1 à 1BN0#5, de 500K laps, ont été réalisées sans admission de pseudonoeuds et les simulations 1BN0#7 à 1BN0#11, de 500K laps, ont été réalisées avec la possibilité de créer des pseudonoeuds.

4.2.2 La force relative de rétention (FRR)

Il est commode d'imaginer les variations des FRR comme des variations de température. En effet, une diminution des FRR amène un relâchement des liens comme le ferait une augmentation de la température. La figure 4.3 présente l'effet d'une diminution des FRR sur le TE de la S_{native} et de la S_0, ainsi que sur le nombre total de structures. Les unités en abscisse sont décroissantes, la première unité correspond aux valeurs par défaut des FRR, soient GC98, AU91 et GU80 et la dernière unité correspond à des valeurs nulles soient GC0, AU0 et GU0. Pour chaque unité en abscisse, une simulation de 500K laps avec 1BN0 a été réalisée en utilisant les FRR indiquées.

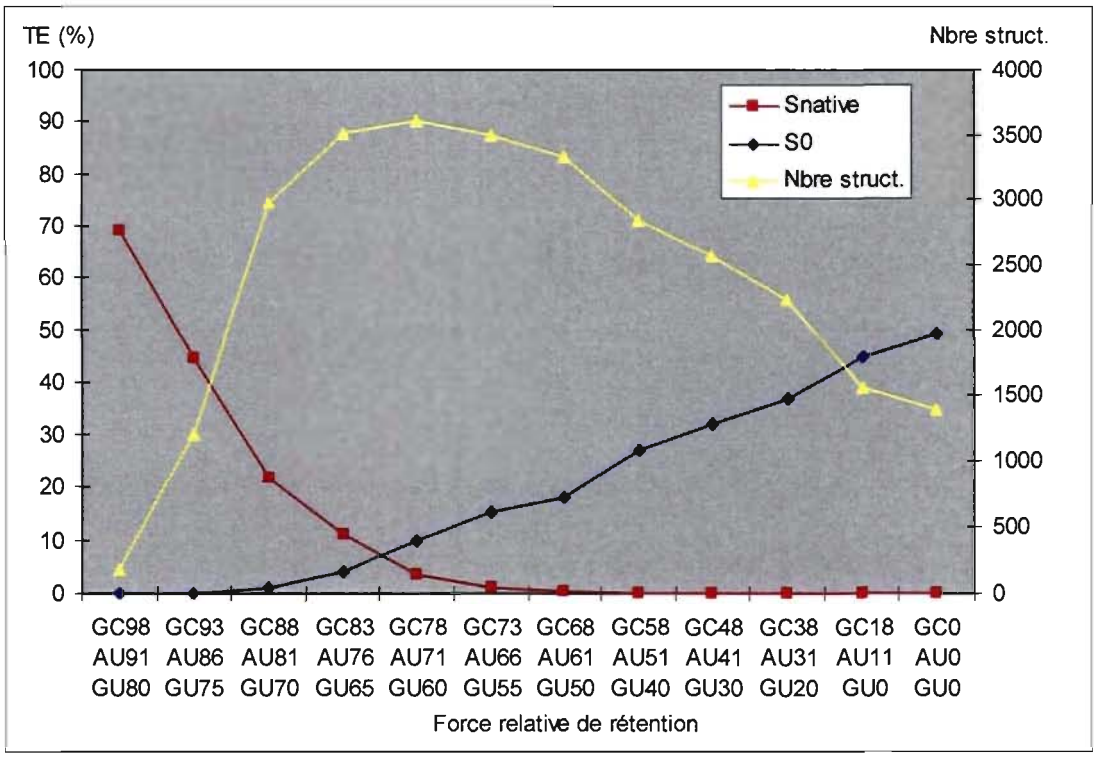

Figure 4.3 Effets des FRR sur les TE et le nombre de structures
À chaque unité en abscisse correspond une simulation de 500K laps avec 1BN0 en utilisant les FRR indiquées.

On remarque que, avec la diminution des FRR, l'émergence de la S_{native} diminue et celle de la S_0 augmente, ce qui est conforme avec les expériences de dénaturation de l'ARN [2]. On remarque aussi que la diminution des FRR provoque d'abord une augmentation rapide du nombre de structures, qui atteint presque la $|C|$, pour ensuite diminuer progressivement à mesure que l'état d'équilibre se déplace autour de la S_0. Ainsi, la transition entre l'état d'équilibre autour de la S_{native} et celui autour de la S_0 est progressive et passe par des états intermédiaires plus anarchiques, ce qui est aussi conforme avec l'aspect dynamique de l'ARN [2].

Les valeurs par défaut des FRR ont été déterminées de manière empirique d'après les résultats obtenus sur quelques séquences d'ARN. Les figures 4.4 et 4.5 montrent l'ajustement, avec la séquence 1BN0, de la FRR pour les liens GC lorsque les FRR des liens AU et GU sont constantes (AU91, GU80). On y retrouve le nombre de structures rencontrées avant l'apparition de la S_{native}, ainsi que le nombre de structures après 100K laps et 500K laps. Chaque point représente la valeur moyenne de cinq simulations.

Dans la figure 4.4, les minimums et les maximums sont également tracés, ce qui permet d'évaluer la dispersion des résultats. On constate que la dispersion est relativement faible sur toute la longueur de la courbe de 500K laps, tandis que la dispersion de 100K laps est beaucoup plus variable, ne s'affaiblissant qu'aux valeurs élevées de FRR. Donc, peu importe la valeur de la FRR, le nombre total de structures semble s'uniformiser à mesure que la durée des simulations augmente. Ces deux courbes sont sensiblement de même forme, celle de 500K laps étant seulement un peu plus accentuée. Elles présentent une légère inflexion vers le bas, à la valeur 20, puis elles augmentent progressivement, pour finalement atteindre un point de chute vers la valeur 70.

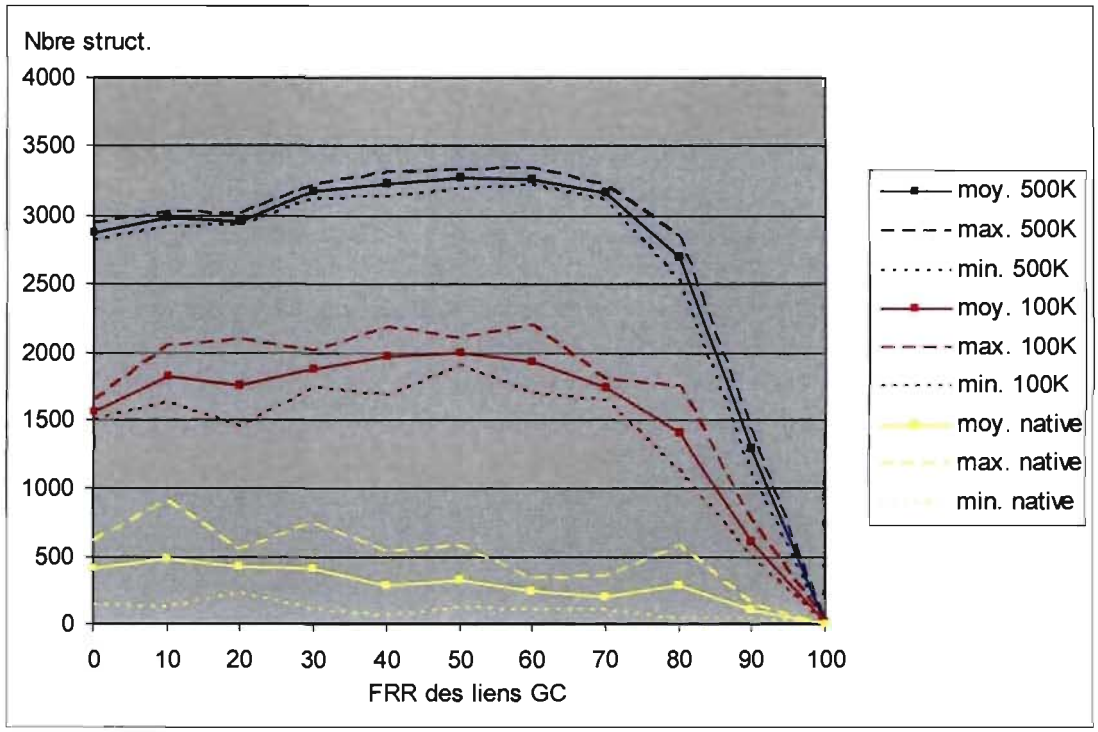

Figure 4.4 Détermination approximative de la FRR pour les liens GC
Cinq simulations de 500K laps avec 1BN0 ont été réalisées pour 11 valeurs différentes de la FRR. Chaque simulation fournit le nombre de structures rencontrées avant l'apparition de la S_{native}, le nombre de structures rencontrées à 100K laps et le nombre de structures rencontrées à 500K laps. Le graphique présente les courbes des valeurs moyennes obtenues, des valeurs maximums et des valeurs minimums.

Avec une FRR de 100, tout lien GC qui se forme ne peut plus se défaire. C'est exactement comme si la température venait d'être abaissée au point de congélation. Les premières structures figent littéralement et le nombre total de structures est au plus bas. À l'inverse, une valeur de 0 n'apporte aucune contrainte sur les liens GC et, par conséquent, aucune contrainte sur les structures. Le nombre de structures est alors très élevé, mais sans toutefois être à son maximum.

Lorsque la FRR commence à augmenter, le nombre de structures se met aussi à augmenter légèrement. Ce phénomène peut s'expliquer de la façon suivante : à mesure que la FRR augmente, les liens GC sont maintenus plus longtemps, ce qui permet l'exploration de nouvelles structures dans C, donc une augmentation du nombre de structures rencontrées. Cet effet se poursuit jusqu'à une FRR de 70.

Dépassée cette valeur, les liens GC exercent une contrainte tellement forte que les structures perdent de la flexibilité et le nombre de structures chute alors de manière importante.

La figure 4.5 permet de zoomer sur l'aire entre les valeurs 90 et 100.

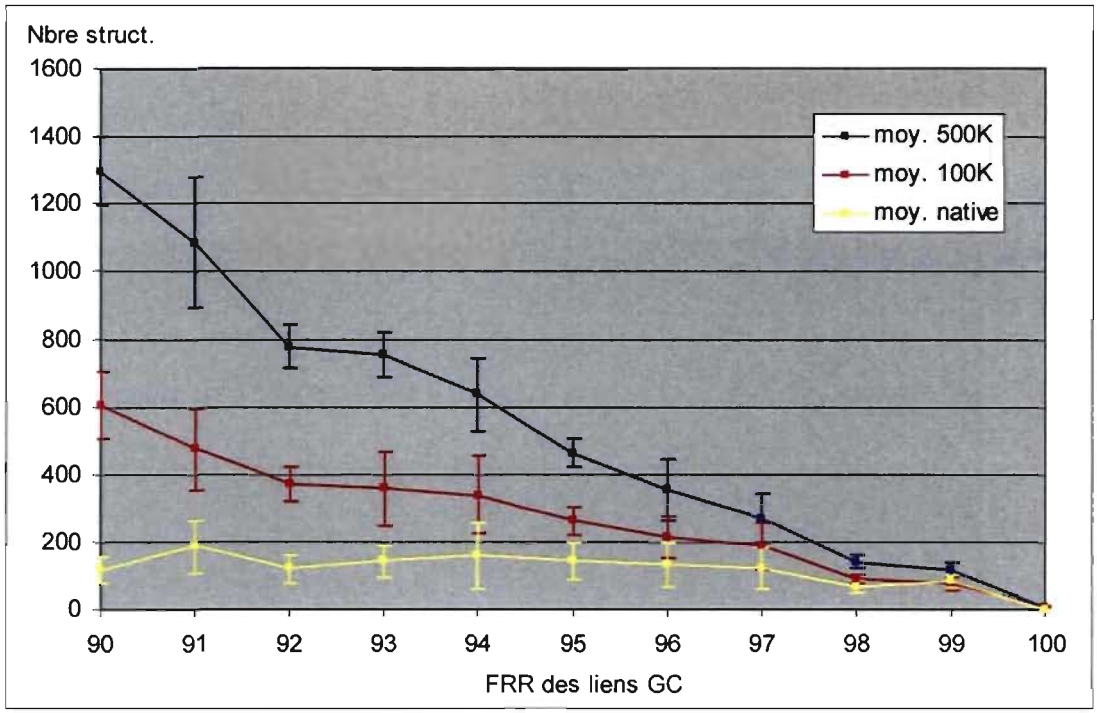

Figure 4.5 Détermination précise de la FRR pour les liens GC
Cinq simulations de 500K laps avec 1BN0 ont été réalisées pour 11 valeurs différentes de la FRR. Chaque simulation fournit le nombre de structures rencontrées avant l'apparition de la S_{native}, le nombre de structures rencontrées à 100K laps et le nombre de structures rencontrées à 500K laps. Le graphique présente les courbes des valeurs moyennes obtenues et les écarts types.

Nous sommes intéressés par une valeur de FRR qui permettra d'atteindre, le plus rapidement possible, l'équilibre autour de la S_{native} tout en minimisant le nombre de structures. La figure 4.5 montre que les valeurs 96, 97 et 98 sont toutes acceptables, la valeur 99 présentant trop de rigidité et la valeur 95 un peu trop de liberté. Nous avons opté pour une valeur de 98 dans nos simulations afin d'obtenir le minimum de

structures. Les valeurs par défaut des autres FRR ont été choisies de manière similaire.

4.2.3 Le vortex

Comme nous l'avons indiqué précédemment, le paramètre vortex sert à augmenter les interactions en favorisant le mouvement des particules. Il est donc utilisé pour accélérer la formation des structures dans une simulation. Cependant, il n'est pas sans effets secondaires. Le TE de certaines structures, plus fragiles ou moins stables, peut diminuer de manière importante si on utilise une valeur de vortex trop élevée.

La figure 4.6 montre l'effet du vortex sur le TE de la S_{native} de 1BN0. Chaque point est le résultat moyen de cinq simulations. Avec un vortex de 1, une simulation de 1BN0 a besoin de près de 1M laps pour atteindre la stabilité. C'est pour cela qu'une baisse marquée du TE est observée dans la courbe de 500K laps, le plein potentiel du TE n'étant pas encore atteint. En augmentant le vortex, la stabilité est atteinte plus vite. C'est un peu comme si on agitait une éprouvette pour accélérer la réaction. Mais dépassé un certain seuil, le TE diminue progressivement à mesure que la valeur du vortex augmente. C'est comme si on agitait trop l'éprouvette au point de perturber la réaction.

Ce gain de mouvement va avoir un effet direct sur le nombre de structures. Plus on augmente le vortex, plus le système est dynamique et plus on favorise la formation de structures différentes (figure 4.7). Cela va aussi permettre d'atteindre plus rapidement la S_{native}. La figure 4.8 montre la relation entre la valeur du vortex et le moment de la première apparition de la S_{native}.

De manière générale, une modification de vortex ne déplace pas l'équilibre vers une autre structure complètement différente comme peut le faire une modification de la FRR, mais elle peut déstabiliser les structures et faire varier les TE. Nous avons

utilisé un vortex de 8 dans nos simulations parce qu'il nous semblait un bon compromis entre le gain en laps et la diminution des TE.

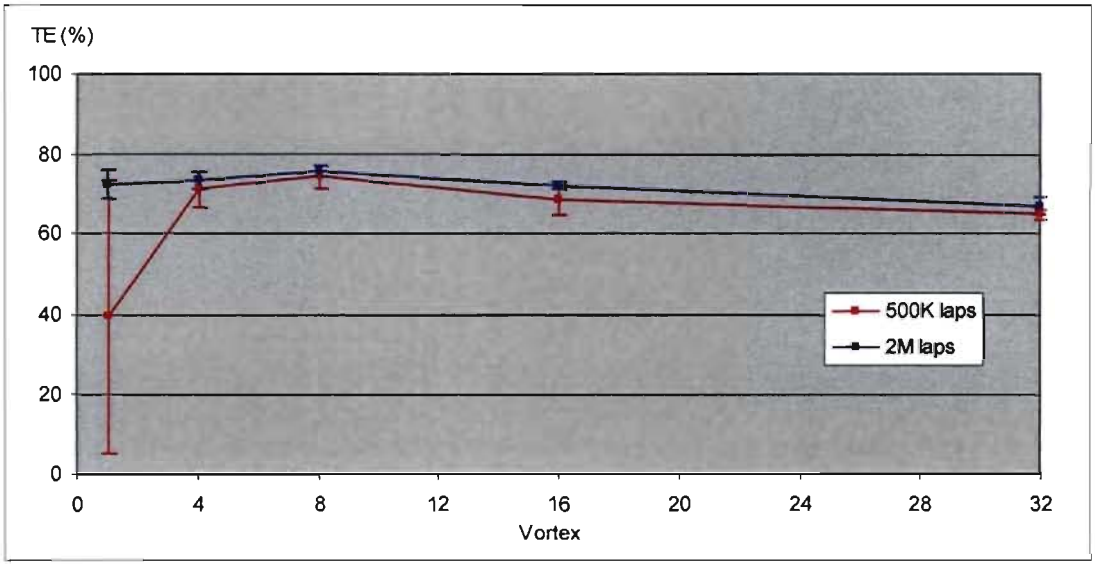

Figure 4.6 Effet du vortex sur le TE de la S_{native}
Chaque point représente le TE moyen obtenu pour cinq simulations avec 1BN0. Pour la valeur de vortex égale à un, une des simulations de 500K laps n'a jamais trouvé la S_{native}, i.e. que la valeur du TE a été nulle. Les écarts types sont affichés.

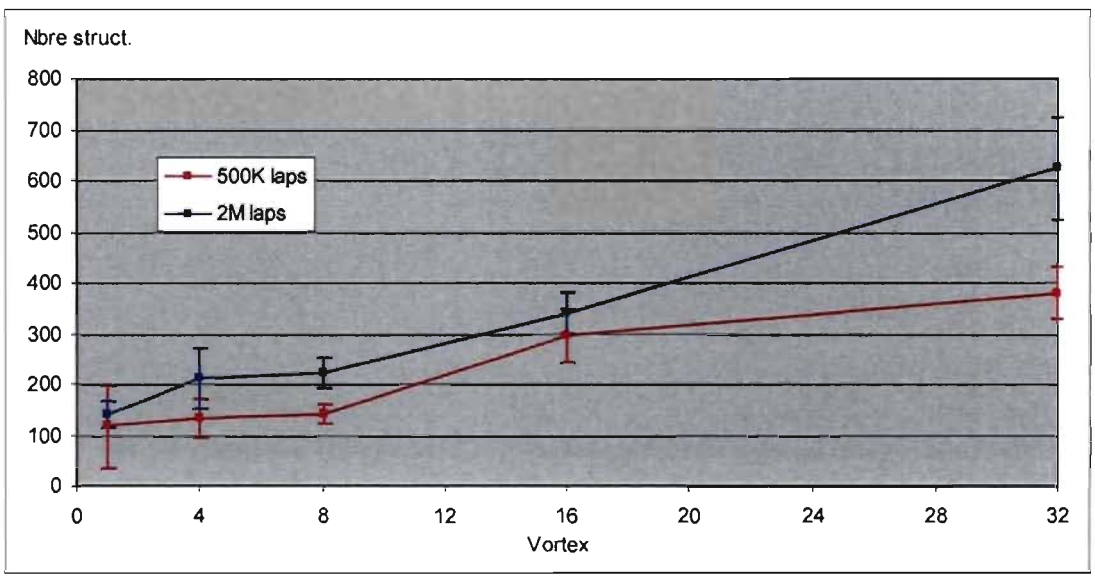

Figure 4.7 Effet du vortex sur le nombre total de structures
Chaque point représente la valeur moyenne du nombre total de structures rencontrées pour cinq simulations avec 1BN0. Les écarts types sont affichés pour chaque point.

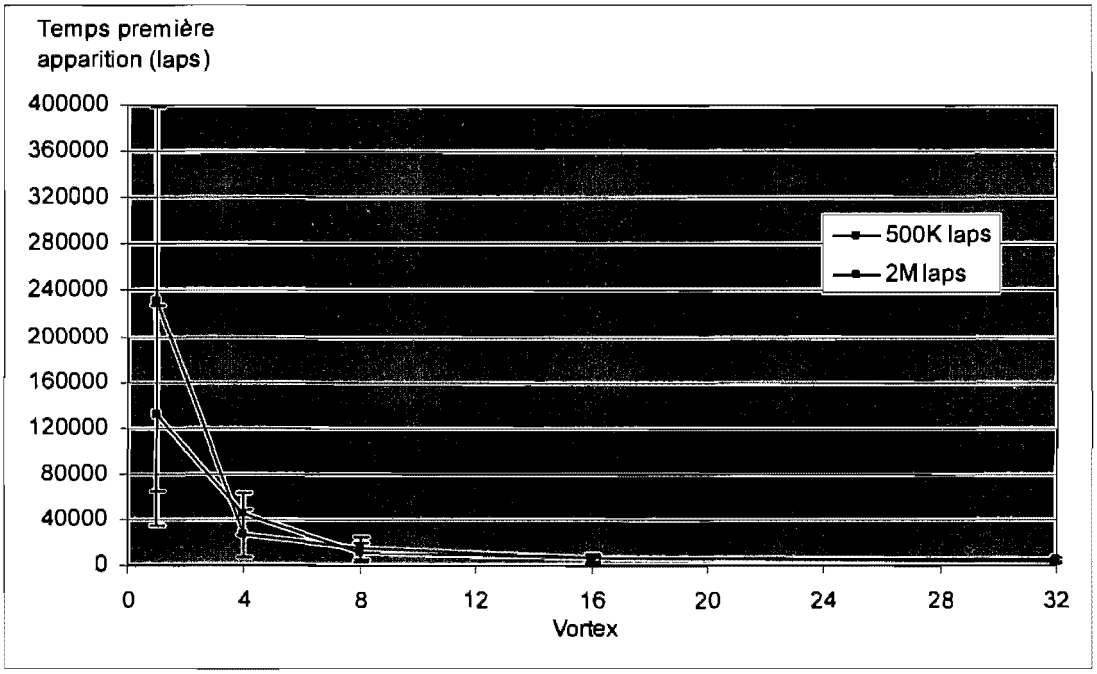

Figure 4.8 Effet du vortex sur le moment d'apparition de la S_{native}
Chaque point représente la valeur moyenne du temps (laps) nécessaire pour atteindre la S_{native} lors de cinq simulations avec 1BN0. Pour la valeur de vortex égale à un, une des simulations de 500K laps n'a jamais atteint la S_{native}. Ce point représente donc la moyenne de quatre simulations au lieu de cinq. Les écarts types sont affichés pour chaque point.

4.2.4 La contrainte minimale de boucle (CMB)

L'augmentation de la valeur de la CMB affecte directement la cardinalité de *C*. Par exemple, lorsqu'on augmente la valeur de la CMB de 3 à 4 pour la séquence 1BN0, la cardinalité de C diminue de 4 476 à 2 538. CA-RNA n'accepte que des valeurs de CMB $\geq$ 3. Le tableau 4.8 présente la moyenne des résultats de 10 simulations de 500K laps avec 1BN0 pour des CMB de 3 et de 4.

	TE S_{native}	**Nbre struct. avant S_{native}**	**Nbre struct. total**	**Nbre laps avant S_{native}**
CMB = 3	75.1%	65.7	144.3	11227.2
CMB = 4	93.6%	95.1	136.7	14591.9

Tableau 4.8 Effets de la CMB
Résultats moyens de 10 simulations de 500K laps avec 1BN0.

L'augmentation du TE pour une CMB de 4 est attribuable au fait que la structure que l'on trouve normalement en R2 (tableau 4.1 à 4.5) et qui avait un TE d'environ 19%, ne fait plus partie de l'espace C. Ce concurrent en moins, le TE de la S_{native} augmente.

Le chemin vers la S_{native} semble plus rapide (Nbre laps avant S_{native}) et plus direct (Nbre struct. avant S_{native}) avec une CMB de 3. Cela peut s'expliquer par le fait que cette valeur de CMB permet des liens entre des nucléotides plus rapprochés, provoquant ainsi un repliement plus rapide.

4.3 Diverses séquences d'ARN

Le tableau 4.9 présente les structures P et les S_{native} de neuf autres séquences d'ARN que nous avons utilisées, ainsi que les TE qui ont été obtenus avec ces séquences. Ce sont toutes des séquences de longueur inférieure à 20. Toutes ces simulations ont été réalisées en conservant une paramétrisation à valeur identique dans tous les cas. Les structures y sont représentées à l'aide de parenthèses.

Pour toutes ces simulations, la S_{native} est aussi la structure qui émerge, et ce, de manière très significative (figures A.2 à A.10). Les TE observés varient de 60% à 97%, dont sept sur neuf sont supérieurs à 90%. De plus, pour les simulations 1VOP#1, 1K4B#1, 1OQ0#1, 1ATW#1 et 1J4Y#1, la S_{R2} est dans le voisinage de la S_{native}.

Molécule	*n* TE_{R1} TE_{R2}	Structure *P* Structure $S_{R1} = S_{native}$ Structure S_{R2}
1IDV	*10*	GGGCGUGCCC
	91.9%	(((::::)))
	2.3%	:((::::::))
1I46	*13*	GGUGCGUAGCACC
	97.7%	((((((:::))))))
	0.6%	:::(:((:::)))
1VOP	*13*	GACUGGGGCGGUC
	94.9%	((((((:::))))))
	1.8%	(((:(:::):)))
1IK1	*14*	GGUACUAUGUACCA
	97.5%	((((((:::)))))):
	0.5%	(((((:::))):)::
1K4B	*14*	GGUUCAGUUGAACC
	82.0%	((((((:::))))))
	3.1%	(((:(::::):)))
1ATW	*15*	GCUCCAGAUGGAGCG
	91.1%	((((((:::))))))):
	2.9%	(((:(:::::)))))):
1OQ0	*15*	GAGAGUUGGGCUCUC
	97.7%	((((((:::))))))
	1.3%	((((((:::::))))))
1J4Y	*17*	GGGGAUUGAAAAUCCCC
	94.8%	(((((((:::)))))))
	1.6%	(((((((:::::):)))))
1Z30	*18*	GGCGUUCGUUAGAACGUC
	60.7%	(((((((:::::)))))))
	8.8%	((:::)):::((:::))

Tableau 4.9 TE obtenus avec neuf autres molécules d'ARN
Chaque simulation est de 500K laps.

Le tableau 4.10 présente les résultats pour deux autres molécules particulières. La première, identifiée PK5, est une molécule artificielle [74] de 26 nucléotides dont la S_{native} contient des pseudonoeuds. Dans les simulations avec PK5, la présence des pseudonoeuds a donc été admise. Pour cette séquence, des parenthèses carrées ont été ajoutées afin d'illustrer correctement le pairage des pseudonoeuds. La deuxième molécule, identifiée 1A9L, tire son intérêt du fait qu'elle est beaucoup plus longue que les molécules précédentes. Cette molécule est constituée de 38 nucléotides.

Les TE observés ici sont plus faibles, mais dans les deux cas, l'espace C est aussi beaucoup plus grand. Il faut prendre en considération que la cardinalité de C est de plusieurs millions de structures pour PK5 et 1A9L, comparativement à quelques milliers pour les molécules précédentes. Malgré cela, la S_{native} est la structure émergente (figures A.11 et A.12) et la S_{R2} est dans le voisinage immédiat de la S_{native}.

Molécule	***n***	**Structure *P***
	TE_{R1}	**Structure $S_{R1} = S_{native}$**
	TE_{R2}	**Structure S_{R2}**
PK5	*26*	GCGAUUUCUGACCGCUUUUUUGUCAG
	44.3%	(((::::[[[[[)))::::::]]]]]
	2.3%	(((::::::[[[[)))::::::]]]]:
1A9L	*38*	GGGUGACUCCAGAGGUCGAGAGACCGGAGAUAUCACCC
	23.0%	((((((((((:::((((::::))))))))):::))))))
	13.7%	((((((((((:::((((::::))))))))):):::))))))

Tableau 4.10 TE obtenus avec PK5 et 1A9L
Chaque simulation est de 500K laps. Les crochets différencient les pseudonoeuds.

4.4 Hybridation de séquences

Une hybridation de séquences se produit lorsqu'il y a formation de liens entre des nucléotides de chaînes différentes. Dans le cas d'une hybridation, nous parlerons de brins au lieu de chaînes et nous parlerons de molécule hybride, ou d'hybride, pour une molécule formée de plusieurs brins. La structure secondaire native pour une molécule hybride sera identifiée $S_{hybride}$.

Avec les méthodes mathématiques, l'hybridation n'est pas toujours possible et si elle l'est, il s'agit d'un cas particulier qui nécessite un traitement supplémentaire. La seule modification apportée à CA-RNA a été de réduire le treillis à des dimensions de 30 x 30 x 30 afin d'augmenter la probabilité que les brins distincts entrent en contact.

La figure 4.9 montre un exemple de la disposition de brins distincts dans le treillis réduit. Il s'agit de la configuration initiale de la molécule 1EKW. On y voit les trois

brins étirés placés aux extrémités du treillis réduit, dont les limites sont tracées en rouge.

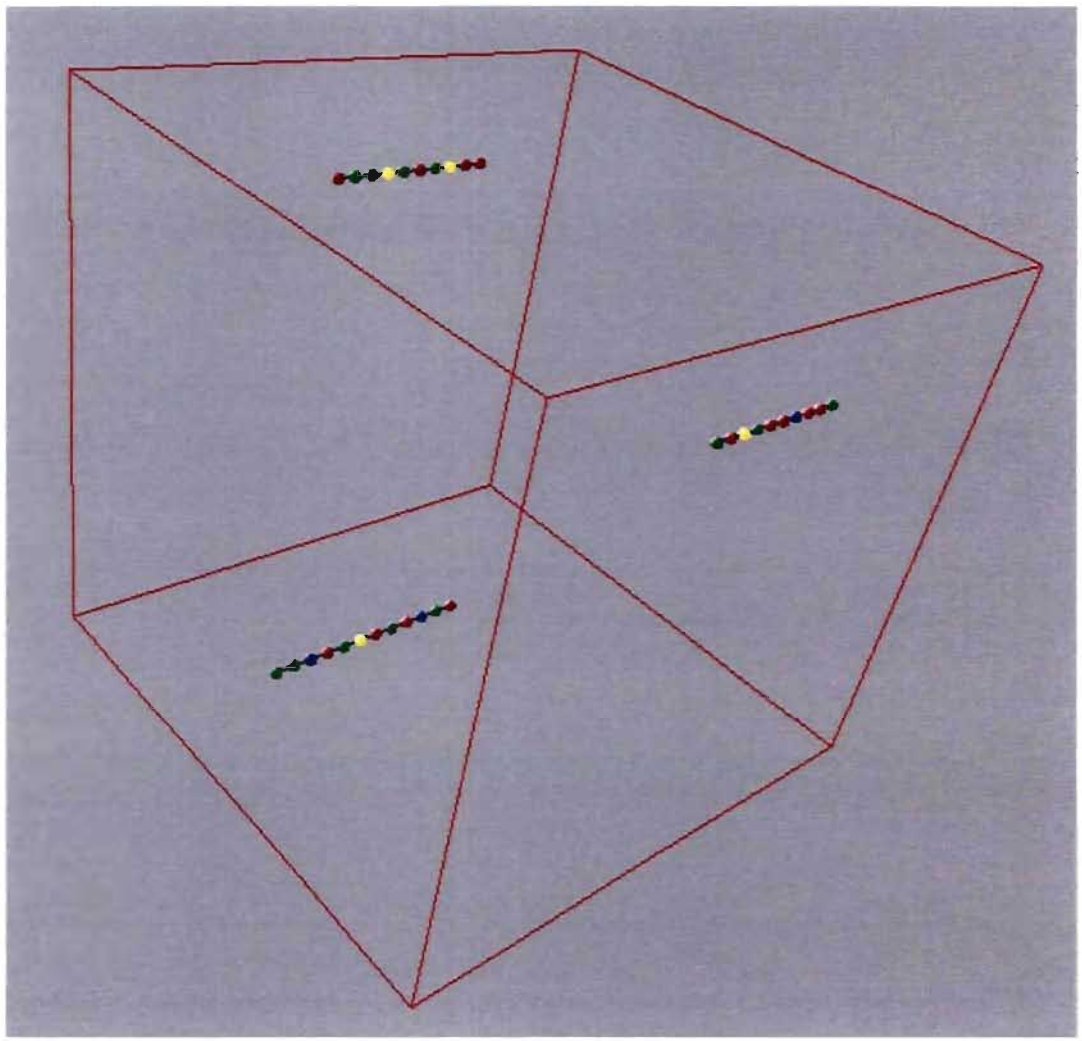

Figure 4.9 Configuration initiale pour la molécule 1EKW
Les trois brins sont placés aux extrémités d'un treillis réduit

Le tableau 4.11 présente les structures P et les structures S_{hybride} de quatre molécules hybrides à deux brins, ainsi que les TE qui ont été obtenus. Les couleurs noire et rouge permettent de distinguer les deux brins différents. La longueur des brins varie de 6 à 15 nucléotides et le nombre de nucléotides par molécule varie de 12 à 27 nucléotides. Les simulations étaient de 1M laps afin de donner suffisamment de temps aux brins de se rencontrer.

Molécule	*n1, n2* TE_{R1} TE_{R2}	**Structures *P*** **Structure $S_{R1} = S_{hybride}$** **Structure S_{R2}**
1PBM	*6,6* 93.5% 3.7%	CGCGCG CGCGCG (((((( )))))) ::((((::))))
1EKA	*8,8* 86.6% 2.3%	GAGUGCUC GAGUGCUC (((((((( )))))))) (((::((( )))::)))
1DQF	*9,10* 77.7% 12.1%	GCCACCCUG CAGGGUCGGC (((((((((( ))))))):))) (((:((((( )))))::)))
397D	*15,12* 40.3% 11.0%	GGCCAGAUCUGAGCG GCUCUCUGGCCC (((((((:::((((:))))))))))): ((((((:::((((:))))):))))):

Tableau 4.11 TE obtenus avec quatre molécules à deux brins
Chaque simulation est de 1M laps. Les brins différents sont de couleurs différentes.

Pour toutes ces simulations, les résultats sont concluants. La $S_{hybride}$ est la structure émergente (figures A.13 à A.16), et ce, de manière très significative. Les TE observés varient de 40% à 93%. Pour les simulations avec 1EKA, 1DQF et 397D, la S_{R2} est dans le voisinage de la $S_{hybride}$.

Le tableau 4.12 présente les résultats pour deux autres molécules particulières. La première, 1F27, est un hybride à deux brins dont la $S_{hybride}$ est constituée de pseudonoeuds, ce qui augmente considérablement son espace *C*. La deuxième, 1EKW, est un hybride formé à partir de trois brins. La molécule 1EKW est en fait une molécule d'ADN. Puisque le nucléotide U est remplacé par un nucléotide T (Thymine) dans l'ADN, nous avons remplacé le type AU par un type AT et désactivé les liens de type GU. La longueur des brins, pour les deux molécules, varie de 10 à 19 nucléotides et le nombre de nucléotides est de 30 pour la molécule 1F27 et de 32 pour 1EKW. Dans les deux cas, la $S_{hybride}$ est encore la structure émergente (figures A.17 et A.18) et la S_{R2} est dans le voisinage de la $S_{hybride}$.

Molécule	*n1, n2, n3* TE_{R1} TE_{R2}	**Structures *P*** **Structure $S_{R1} = S_{hybride}$** **Structure S_{R2}**
1F27	*19,11* 24.4% 3.4%	ACCGUCAGAGGACACGGUU AAAAAGUCCUC (((((::[[[[[[)))))): :::::]]]]]] ((((:::[[[[[[:)))))): :::::]]]]]]
1EKW	*10,12,10* 85.4% 3.9%	CGGUGCGUCC GGACGUCGCAGC GCUGCCACCG ((((((((((( ))))))::((((( )))))))))) ((((((((((( ))))))):)((((( ))))):))))

Tableau 4.12 TE pour un hybride avec pseudonoeuds et un hybride à trois brins
Chaque simulation est de 1M laps. Les brins différents sont de couleurs différentes. Les crochets différencient les pseudonoeuds.

Le tableau 4.13 présente les résultats pour 2B8R, un hybride avec pseudonoeuds de 46 nucléotides, qui présente une structure en forme de deux têtes d'épingle attachées par les têtes. La $S_{hybride}$ est la structure émergente (figure A.20), mais avec seulement une faible avance par rapport à la S_{R2} qui est dans le voisinage de la $S_{hybride}$. Le tableau 4.13 présente aussi les résultats pour 2P89, un hybride avec pseudonoeuds de 48 nucléotides, qui présente une structure de forme complexe. La $S_{hybride}$ de 2P89 est une structure qui présente un lien GU impossible à réaliser avec notre modèle. En effet, les nucléotides formant ce lien sont trop près l'un de l'autre. Deux nucléotides seulement les séparent, ce qui est en dessous de la valeur de la CMB. Malgré cela, la structure émergente (figure A.19) obtenue correspond à la $S_{hybride}$ moins ce lien.

Mol.	*n1, n2* TE_{R1} TE_{R2}	**Structures *P*** **Structure $S_{R1} = S_{hybride}$** **Structure S_{R2}**
2B8R	*23,23* 15.8% 13.8%	CUUGCUGAAGCGCGCACGGCAAG CUUGCUGAAGCGCGCACGGCAAG (((((((::[[[[[[:))))))) ((((((::]]]]]]:))))))) (((((((::([[[[):))))))) ((((((::(]]]]):))))))))
2P89	*34,14* 8.9% 1.9%	GGCCUUAGGAAACAGUUCGCUGUGCCGAAAGGUC UUCGGCUCUUCCUA ((((((((((((((::::))))))[[[[[[))))))]]]]]]::]]]]]] ((((((((((:((((::::))))))[[[[[[))))))]]]]]]:::]]]]]

Tableau 4.13 TE avec deux hybrides de 46 et 48 nucléotides
Chaque simulation est de 2M laps. Les brins différents sont de couleurs différentes. Les crochets différencient les pseudonoeuds.

Vingt simulations supplémentaires, de 2M laps, ont été effectuées avec 2B8R afin de vérifier la structure émergente. Trois structures différentes ont émergé. Le tableau 4.14 présente ces trois structures et le nombre de simulations où elles ont été dominantes. La première structure (figure A.20) est la S_{hybride} qui a été la structure dominante dans 12 simulations sur 20. La deuxième structure (figure A.21) est la structure S_{R2} du tableau 4.12, une structure dans le voisinage de la S_{hybride}, qui a été dominante dans cinq simulations. Finalement, la troisième structure (figure A.22) est une structure complètement différente. Elle correspond à la structure de minimum d'énergie calculée par le programme Mfold [32]. Cette structure, qui n'est pas considérée comme une structure secondaire native, a été la structure dominante dans trois simulations.

Nbre de simul.	**Structure**
	CUUGCUGAAGCGCGCACGGCAAG CUUGCUGAAGCGCGCACGGCAAG
12	(((((((::[[[[[[:))))))) (((((((::]]]]]]:)))))))
5	(((((((::([[[[):))))))) (((((((::(]]]]):)))))))
3	(((((((::((((((:((((((( )))))))::)))))):)))))))

Tableau 4.14 Fréquence absolue des trois structures dominantes avec 2B8R
Vingt simulations, de 2M laps, ont été effectuées. La première structure est la S_{hybride}, la deuxième est dans le voisinage de la S_{hybride}, et la troisième correspond à la structure de minimum d'énergie calculée par le programme Mfold [32]. Les brins différents sont de couleurs différentes. Les crochets différencient les pseudonoeuds.

Ainsi, le modèle a été capable de simuler l'hybridation pour sept molécules à deux brins et une molécule à trois brins. Puisque chaque nucléotide est considéré comme une particule autonome et que les règles sont locales, le modèle considère indifféremment le repliement d'une seule chaîne ou l'hybridation de plusieurs brins.

4.5 Repliement

Actuellement, certains chercheurs s'interrogent à savoir s'il existe un chemin universel de $S_0 \rightarrow S_{native}$, c'est-à-dire, est-ce qu'une séquence d'ARN utilise toujours un seul et même chemin de $S_0 \rightarrow S_{native}$. Les travaux de Zhang et Chen [75] suggèrent le contraire, i.e. qu'il y a plus d'un chemin possible et ce, même pour les structures simples.

Dans les simulations effectuées, nous avons pu observer que le résultat présentant la plus grande variabilité était, sans contredit, le nombre de structures rencontrées avant la S_{native}. Ceci indique clairement qu'il y a eu différents chemins, de différentes longueurs, partant de $S_0 \rightarrow S_{native}$. Toutefois, ces chemins ne sont pas distincts. Le tableau 4.15 présente les proportions de structures qu'ont en commun les simulations 1BN0#1 à 1BN0#5. On constate qu'entre 19% et 28% des structures rencontrées dans une simulation l'ont aussi été dans les quatre autres simulations. Par ailleurs, la proportion de structures uniques à une simulation varie seulement de 20% à 38%. Donc, même s'il existe plusieurs chemins, ces derniers présentent de bonnes similitudes.

Simulation	**Nbre struct. total**	**Nbre struct. dans les 5**	**Nbre struct. dans 4 autres**	**Nbre struct. dans 3 autres**	**Nbre struct. dans 2 autres**	**Nbre struct. unique**
1BN0#1	180	35 *19%*	21 *12%*	29 *16%*	37 *21%*	58 *32%*
1BN0#2	132	35 *27%*	13 *10%*	18 *14%*	16 *12%*	50 *38%*
1BN0#3	123	35 *28%*	21 *17%*	17 *14%*	26 *21%*	24 *20%*
1BN0#4	131	35 *27%*	19 *15%*	20 *15%*	29 *22%*	28 *21%*
1BN0#5	133	35 *26%*	18 *14%*	21 *16%*	22 *17%*	37 *28%*

Tableau 4.15 Nombre de structures communes dans 1BN0#1 à 1BN0#5 Chaque simulation est de 500K laps.

Les chercheurs s'interrogent aussi sur certaines structures intermédiaires, à savoir celles dont le passage est obligatoire et celles dont le passage a nécessité plus de temps. L'analyse détaillée du chemin de plusieurs simulations pourrait permettre d'identifier des structures obligatoires ou, du moins, des structures préférées. De plus, l'analyse des états d'équilibre dans le temps pourrait permettre d'identifier les structures dont la présence a été dominante. La figure 4.10 présente les différentes structures dominantes qui sont apparues lors de la simulation 1BN0#1 et la figure 4.11, celles qui sont apparues lors de la simulation 1BN0#3. Ces structures sont décrites dans les tableaux 4.16 et 4.18.

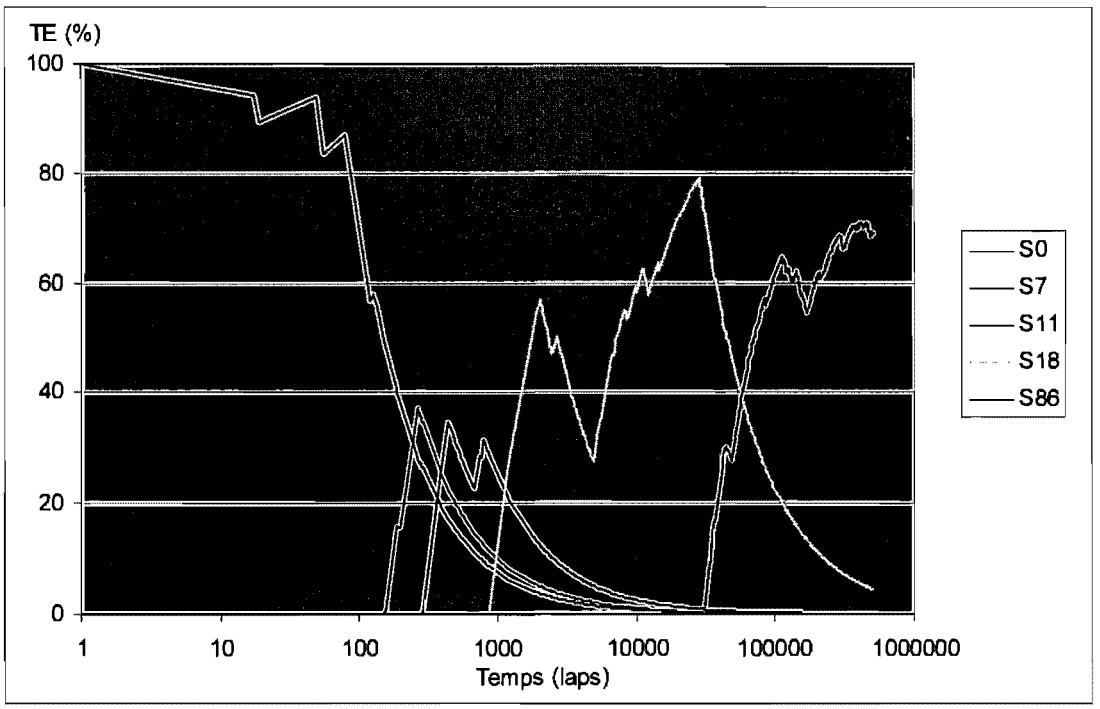

Figure 4.10 Structures dominantes dans 1BN0#1
Simulation de 500K laps. S_{86} correspond à la S_{native}. S_{18} = {4=10, 5~9, 12=20, 13=19}

On reconnait les courbes de la S_{native} en rouge et celles de la S_0 en bleu. Les courbes de couleur jaune représentent toutes deux la même structure soit {4=10, 5~9, 12=20, 13=19}. Ainsi, cette structure est apparue comme structure dominante dans ces deux simulations. L'examen des tableaux 4.16 à 4.20 nous permet de constater qu'elle a

aussi été dominante dans 1BN0#4 et 1BN0#5, soit quatre simulations sur cinq. Fait intéressant, elle était absente comme structure rencontrée dans la simulation 1BN0#2.

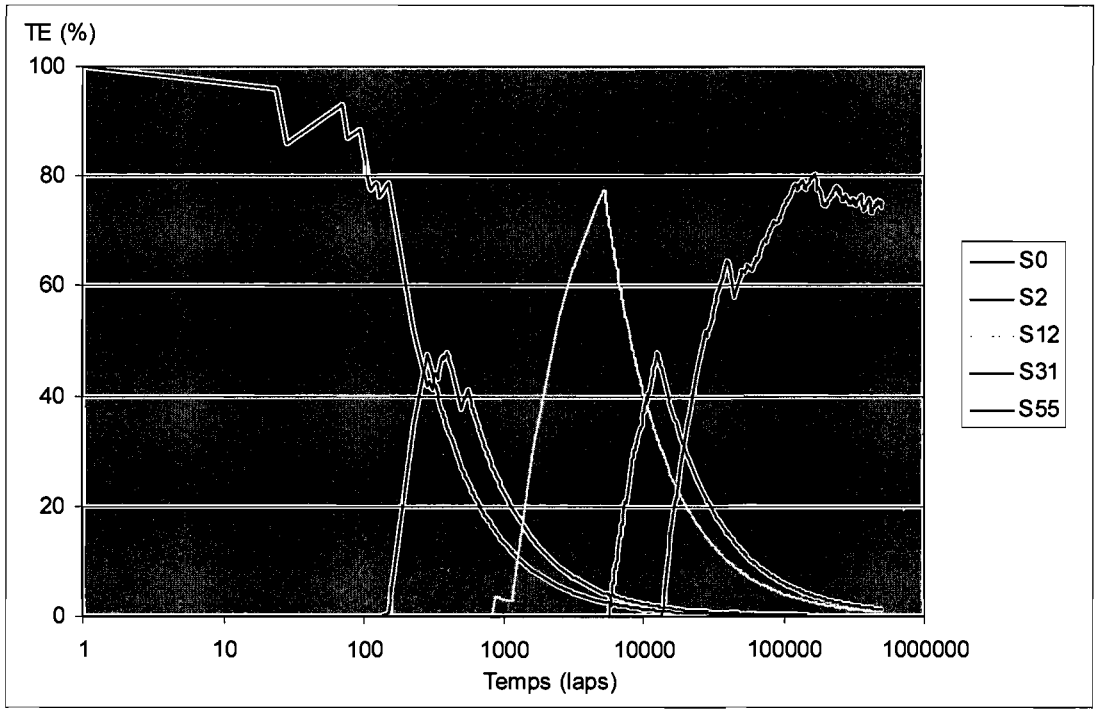

Figure 4.11 Structures dominantes dans 1BN0#3
Simulation de 500K laps. S_{55} correspond à la S_{native}. S_{12} = {4=10, 5~9, 12=20, 13=19}

No	Structure secondaire
7	2=8, 12=19, 13~18
11	2=8, 12=20, 13=19
18	4=10, 5~9, 12=20, 13=19
86	[1=20], [2=19], [3-18], [4=17], [5-16], [6-15], [7=14], [8=13]

Tableau 4.16 Structures dominantes dans 1BN0#1
Simulation de 500K laps. Les liens entre crochets sont ceux de la S_{native}.

No	Structure secondaire
44	[1=20], [2=19], [3-18], [4=17], [5-16], 11-15
51	[1=20], [2=19], [3-18], [4=17], [5-16], [6-15], [7=14], 8=12
54	[1=20], [2=19], [3-18], [4=17], [5-16], [6-15], [7=14], [8=13]

Tableau 4.17 Structures dominantes dans 1BN0#2
Simulation de 500K laps. Les liens entre crochets sont ceux de la S_{native}.

No	Structure secondaire
2	2=8
12	4=10, 5~9, 12=20, 13=19
31	[1=20], [2=19], [3-18], 4=10, 5~9, 11-15
55	[1=20], [2=19], [3-18], [4=17], [5-16], [6-15], [7=14], [8=13]

Tableau 4.18 Structures dominantes dans 1BN0#3
Simulation de 500K laps. Les liens entre crochets sont ceux de la S_{native}.

No	Structure secondaire
4	4=9, 13=20
12	4=10, 5~9, 12=20, 13=19
71	[1=20], [2=19], [3-18], [4=17], [5-16], [6-15], [7=14], 8=12
82	[1=20], [2=19], [3-18], [4=17], [5-16], [6-15], [7=14], [8=13]

Tableau 4.19 Structures dominantes dans 1BN0#4
Simulation de 500K laps. Les liens entre crochets sont ceux de la S_{native}.

No	Structure secondaire
9	4=10, 5~9, 12=20, 13=19
55	[1=20], [2=19], [3-18], [4=17], [5-16], [6-15], [7=14], [8=13]

Tableau 4.20 Structures dominantes dans 1BN0#5
Simulation de 500K laps. Les liens entre crochets sont ceux de la S_{native}.

Selon ces résultats, il semble y avoir plus d'un chemin possible pour le repliement de 1BN0. De plus, la structure {4=10, 5~9, 12=20, 13=19} apparaît comme une structure intermédiaire souvent impliquée dans ce repliement. Il est toutefois impossible de corroborer ces résultats car il n'existe aucune autre étude portant sur le repliement de 1BN0.

Par contre, il existe des études [33, 74, 76] portant sur les structures intermédiaires impliquées dans le repliement de PK5. Selon ces études, il semble que le repliement de PK5 pourrait faire intervenir deux structures intermédiaires particulières, soit { [1=15], [2=14], [3=13], 4:12, 5-11, 6~10 } et { [8=26], [9=25], [10=24], [11-23], [12=22] }, que nous désignerons respectivement comme S_A et S_B.

La structure S_A est composée d'un type de lien qui, normalement, n'entre pas dans la définition de structure secondaire. Il s'agit du lien 4:12, qui est une liaison entre des nucléotides A et C. La simulation déjà réalisée avec PK5 (tableau 4.9) est inutile puisqu'elle ne permettait pas la création de ce type de lien. Nous avons donc effectué de nouvelles simulations, en autorisant ce type de lien, pour vérifier si nous obtenions des résultats similaires. À défaut d'avoir effectué diverses expériences pour trouver une valeur plausible de la FRR des liens AC, nous lui avons arbitrairement attribué la valeur 0 afin de permettre les liens tout en limitant leurs effets. Les tableaux 4.21 à 4.25 présentent les structures qui ont été dominantes pour cinq simulations (PK5#2 à PK5#7) de 500K laps, lorsque les liens AC sont autorisés.

No	Structure secondaire
1	1~5
2	1~6
8	3~7, 20~26, 21-25
20	3=8, 20~26, 21-25
212	1=13, [2=14], 3=8, 6-11, 10=15, 20~26, 21-25
404	[1=15], [2=14], 3=8, 7-11, 20~26, 21-25
486	[1=15], [2=14], 3=8, 6-11, 20~26, 21-25
554	[1=15], [2=14], [3=13], 4:12, 5-11, 6~10, 20~26, 21-25
594	[1=15], [2=14], [3=13], 5-11, 6~10, 20~26, 21-25
602	[1=15], [2=14], [3=13], 4:8, 6-11, 20~26, 21-25
612	[1=15], [2=14], [3=13], 6-11, 20~26, 21-25
4076	[1=15], [2=14], [3=13], [8=26], [9-25], [10=24], [11-23], [12=22]

Tableau 4.21 Structures dominantes dans PK5#2
Simulation de 500K laps avec admission des liens AC (FRR=0). La structure 554 est comparable à S_A = { [1=15], [2=14], [3=13], 4:12, 5-11, 6~10 }.

Les S_A et S_B n'apparaissent pas dans la liste des structures dominantes des tableaux 4.21 à 4.25, mais des structures similaires s'y retrouvent. Il s'agit de la structure { [1=15], [2=14], [3=13], 4:12, 5-11, 6~10, 20~26, 21-25 }, que nous avons désignée comme $S_{A'}$ et représentée en rouge dans les tableaux 4.21, 4.22 et 4.25 et les structures { [8=26], [9=25], [10=24], [11-23], [12=22], 14~18 }, { [8=26], [9=25], [10=24], [11-23], [12=22], 14~19 } et { [8=26], [9=25], [10=24], [11-23], [12=22], 14~21 }, désignées respectivement comme S_{B1}, S_{B2} et S_{B3} et regroupées aussi sous la désignation $S_{B'}$. Ces dernières structures sont représentées en bleu dans les tableaux 4.23 et 4.24.

No	Structure secondaire
1	1~5
7	20~26, 21-25
755	[1=15], [2=14], [3=13], 4:12, 5-11, 6~10, 20~26, 21-25
964	[1=15], [2=14], [3=13], 6-11, 20~26, 21-25
965	[1=15], [2=14], [3=13], 4:8, 6-11, 20~26, 21-25
1908	[1=15], [2=14], [3=13], [8=26], [9-25], [10=24], [11-23], [12=22]

Tableau 4.22 Structures dominantes dans PK5#3
Simulation de 500K laps avec admission des liens AC (FRR=0). La structure 755 est comparable à S_A = { [1=15], [2=14], [3=13], 4:12, 5-11, 6~10 }.

No	Structure secondaire
17	3=8, 20~26, 21-25
694	[8=26], [9-25], [10=24], 11-21, [12=22], 14~18
702	7-11, [8=26], [9-25], [10=24], [12=22], 14~18
888	[8=26], [9-25], [10=24], [11-23], [12=22], 14~18
893	[8=26], [9-25], [10=24], [11-23], [12=22], 14~19
916	[8=26], [9-25], [10=24], [11-23], [12=22], 14~21
1728	[1=15], [2=14], [3=13], [8=26], [9-25], [10=24], [11-23], [12=22]

Tableau 4.23 Structures dominantes dans PK5#4
Simulation de 500K laps avec admission des liens AC (FRR=0). Les structures 888, 893 et 916 sont comparables à S_B = { [8=26], [9=25], [10=24], [11-23], [12=22]}.

No	Structure secondaire
3	3~7
4	1=8, 3~7
570	1=12, 2=10, 3~9, 4:8, 7-11, 15=26, 16-25, 18~22
847	[8=26], [9-25], [10=24], [11-23], [12=22], 14~19
857	[8=26], [9-25], [10=24], [11-23], [12=22], 14~21
977	[1=15], [2=14], [3=13], [8=26], [9-25], [10=24], [11-23], [12=22]

Tableau 4.24 Structures dominantes dans PK5#5
Simulation de 500K laps avec admission des liens AC (FRR=0). Les structures 847 et 857 sont comparables à S_B = { [8=26], [9=25], [10=24], [11-23], [12=22]}.

No	Structure secondaire
12	1~5, 4:8, 6~10, 7-11, 20~26, 21-25
23	1=8, 3~7, 6~10
33	1=12, 3~7, 5-11, 6~10
144	[1=15], [2=14], [3=13], 4:12, 5-11, 6~10, 20~26, 21-25
668	[1=15], [2=14], [3=13], [8=26], [9-25], [10=24], [11-23], [12=22]

Tableau 4.25 Structures dominantes dans PK5#6
Simulation de 500K laps avec admission des liens AC (FRR=0). La structure 144 est comparable à S_A = { [1=15], [2=14], [3=13], 4:12, 5-11, 6~10 }.

Les figures 4.12 et 4.13 montrent l'évolution dans le temps des structures dominantes de PK5#5 et PK5#6.

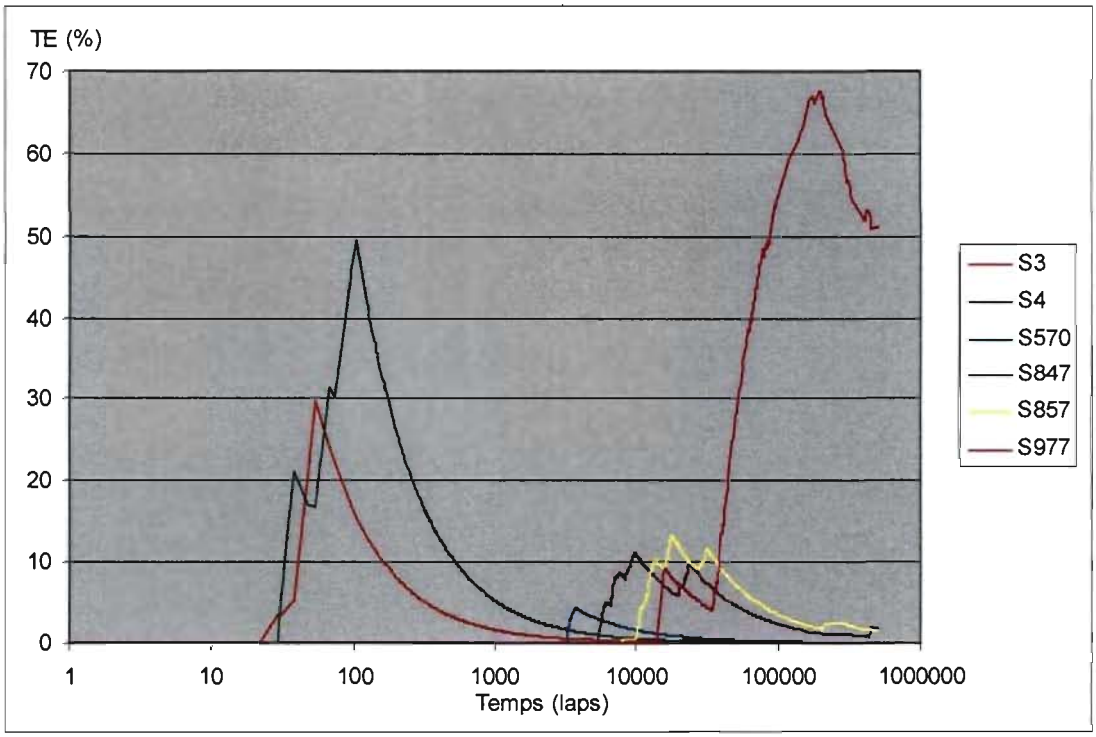

Figure 4.12 Structures dominantes dans PK5#5
Simulation de 500K laps avec admission des liens AC (FRR=0). S977 correspond à la S_{native}. S_{847} et S_{857} sont les structures comparables à S_B = {[8=26], [9=25], [10=24], [11-23], [12=22]}. Les autres structures sont décrites dans le tableau 4.24.

Les études précédentes [33, 74, 76] ne font aucune mention des structures $S_{A'}$ et $S_{B'}$, mais il n'y rien d'étonnant à cela. Les études biophysiques effectuées par Wyatt, Puglisi et Tinoco [74, 76] procèdent par dénaturation progressive, ce qui favorise les liens regroupés par rapport aux liens isolés. En effet, à forces égales, un lien solitaire va se défaire plus facilement qu'un lien groupé qui bénéficie d'une cohésion plus grande due, entre autres, à l'effet d'empilement. Les liens additionnels observés dans $S_{A'}$ sont une paire de liens GU et AU qui se dénature plus rapidement que le regroupement GC, GC, GC, AC, AU GU de S_A. Les liens additionnels observés dans $S_{B'}$ sont des liens GU solitaires qui se dénaturent beaucoup plus rapidement que le regroupement GC, AU, GC AU, GC de S_B. Il est raisonnable de croire que nos liens additionnels ne pouvaient êtres observés dans les études de Wyatt, Puglisi et Tinoco.

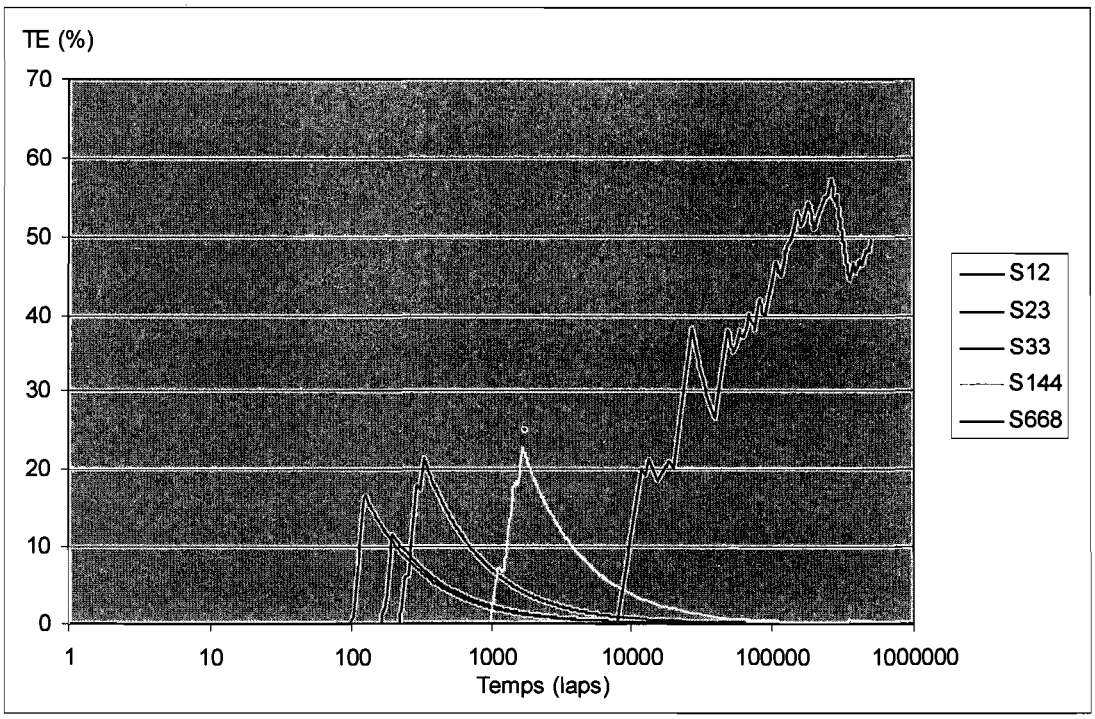

Figure 4.13 Structures dominantes dans PK5#6
Simulation de 500K laps avec admission des liens AC (FRR=0). S_{668} correspond à la S_{native}. S_{144} est la structure comparable à S_A = { [1=15], [2=14], [3=13], 4:12, 5-11, 6~10 }. Les autres structures sont décrites dans le tableau 4.25.

Quant à l'étude de Cao et Chen [33], il s'agit d'une simulation de repliement thermodynamique (dénaturation) qui utilise une approche ne retenant que les différents arrangements d'empilements possibles, ce qui réduit son espace C à 1 505 structures seulement. La structure $S_{B'}$, avec son lien isolé, n'existe pas dans cet espace C. Pour la structure $S_{A'}$, qu'elle existe ou non dans cet espace C, il est raisonnable de croire que la dénaturation simulée ne permette pas plus l'expression de la paire de liens GU et AU que la dénaturation en laboratoire.

Puisque les S_A et S_B sont mutuellement incluses dans $S_{A'}$ et $S_{B'}$ et que ces dernières ne sont pas incompatibles avec les études précédentes, nous considérons que les résultats obtenus pour nos cinq simulations vont dans le même sens que les données actuelles sur le repliement.

4.6 Comparaison avec Mfold

Mfold [32] est l'un des programmes couramment utilisés pour la prédiction statique de la structure secondaire. Il utilise les techniques de programmation dynamique pour calculer, à partir de la séquence, la structure secondaire qui présente le minimum d'énergie en fonction de différents paramètres énergétiques. La structure d'énergie minimum ainsi calculée est la prédiction de Mfold.

Rappelons que le modèle CA-RNA s'intéresse à la dynamique de repliement menant à la structure secondaire. Pour ce faire, il utilise une construction d'AC où les forces et les contraintes sont définies localement pour chaque nucléotide. En aucun moment, il ne calcule l'énergie totale d'une structure. Les différentes structures rencontrées durant le repliement ainsi que la structure émergente sont les prédictions de CA-RNA.

Le tableau 4.26 présente les prédictions obtenues avec Mfold pour 12 des 20 molécules que nous avons utilisées dans les simulations. On y observe que, pour les molécules 1IDV, 1I46, 1VOP, 1IK1, 1K4B, 1ATW, 1J4Y, 1Z30, 1BN0 et 1A9L, les structures de minimum d'énergie calculées par Mfold correspondent aux structures S_{native}, ainsi qu'aux structures émergentes obtenues avec CA-RNA, i.e. que les prédictions de Mfold et CA-RNA sont identiques.

Pour la molécule 1OQ0 (tableau 4.26), Mfold prédit deux structures possibles. La première, celle de minimum d'énergie, est la structure S_{R2} du tableau 4.9, une structure dans le voisinage de la S_{native}. La deuxième structure, avec une énergie légèrement supérieure à la première, correspond à la S_{native}. Avec CA-RNA, la structure émergente est la S_{native} avec un TE de 97.7% tandis que la S_{R2} a un TE de 1.3%. La prédiction obtenue pour 1OQ0 s'avère donc meilleure avec CA-RNA.

Dans toutes nos simulations avec la molécule PK5, la structure émergente a été la S_{native} et ce, même si elle est constituée de pseudonoeuds. La structure prédite par Mfold pour PK5 (tableau 4.26) ne correspond pas à la S_{native}. Il s'agit d'une structure

avec seulement cinq des huit liens secondaires présents dans la S_{native}. Il n'y a rien d'étonnant à cela puisque le programme Mfold n'autorise pas les pseudonoeuds.

Molécule	S_{native}	**Structure *P* Structure prédite par Mfold**
1IDV	 oui	GGGCGUGCCC ((:::::))
1I46	 oui	GGUGCGUAGCACC (((((:::)))))
1VOP	 oui	GACUGGGGCGGUC (((((:::)))))
1IK1	 oui	GGUACUAUGUACCA (((((:::)))))::
1K4B	 oui	GGUUCAGUUGAACC (((((:::)))))
1ATW	 oui	GCUCCAGAUGGAGCG (((((:::)))))::
1OQ0	 **non** oui	GAGAGUUGGGCUCUC (((((:::::))))) ((((((:::))))))
1J4Y	 oui	GGGGAUUGAAAAUCCCC (((((((:::)))))))
1Z30	 oui	GGCGUUCGUUAGAACGUC (((((((:::::)))))))
1BN0	 oui	GGACUAGCGGAGGCUAGUCC ((((((((:::::))))))))
PK5	 **non**	GCGAUUUCUGACCGCUUUUUUGUCAG :::::::((((:::::::::::))))
1A9L	 oui	GGGUGACUCCAGAGGUCGAGAGACCGGAGAUAUCACCC ((((((((((:::(((((:::)))))))))):::))))))

Tableau 4.26 Structures prédites par Mfold
Les prédictions ont été faites en utilisant les paramètres par défaut. Mfold prédit deux structures pour la molécule 1OQ0. Ces deux structures présentent des niveaux d'énergie voisins, mais c'est la première structure qui correspond au minimum d'énergie.

Le tableau 4.27 présente les prédictions obtenues avec Mfold pour l'hybridation des huit autres molécules que nous avons utilisées dans les simulations. On y observe que, pour les molécules 1EKA, 1DQF et 397D, les structures de minimum d'énergie calculées par Mfold correspondent aux structures $S_{hybride}$, ainsi qu'aux structures émergentes obtenues avec CA-RNA. Par contre, pour les molécules 1PBM, 1F27,

2B8R et 2P89, les structures prédites par Mfold ne correspondent pas aux structures $S_{hybride}$, alors que les résultats obtenus avec CA-RNA sont concluants. Il est à noter que les structures $S_{hybride}$ de 1F27, 2B8R et 2P89 intègrent des pseudonoeuds.

Il est impossible d'utiliser le programme Mfold avec la molécule 1EKW (tableau 4.27) puisque cette dernière est formée de trois brins et que le programme Mfold calcule l'hybridation pour deux brins seulement. Cette contrainte n'existe pas dans CA-RNA.

Molécule	$S_{hybride}$	**Structure *P*** **Structure prédite par Mfold**
1PBM	**non**	CGCGCG CGCGCG :((((: :)))):
1EKA	oui	GAGUGCUC GAGUGCUC (((((((( ))))))))
1DQF	oui	GCCACCCUG CAGGGUCGGC (((((((((( ))))))):)))
397D	oui	GGCCAGAUCUGAGCG GCUCUCUGGCCC (((((((:::((((:))))))))))):
1F27	**non**	ACCGUCAGAGGACACGGUU AAAAAGUCCUC :::::::((((((:::::: :::::))))))
1EKW	**non**	CGGUGCGUCC GGACGUCGCAGC GCUGCCACCG **calcul impossible**
2B8R	**non**	CUUGCUGAAGCGCGCACGGCAAG CUUGCUGAAGCGCGCACGGCAAG (((((((::((((((:(((((((( )))))))::))))))):)))))))
2P89	**non**	GGCCUUAGGAAACAGUUCGCUGUGCCGAAAGGUC UUCGGCUCUUCCUA :::::(((((:::((:::::::::(((((((:::::))))))):)))))))

Tableau 4.27 Hybridations prédites par Mfold
Les prédictions ont été faites en utilisant les paramètres par défaut. Les brins différents sont de couleurs différentes. Le programme Mfold ne peut être utilisé avec la molécule 1EKW puisqu'elle est formée de trois brins et que le programme n'accepte que deux brins seulement.

Finalement, avec CA-RNA, nous avons vérifié si les structures S_A et S_B apparaissaient comme structures intermédiaires importantes dans les simulations avec PK5. Cela est impossible à faire avec Mfold puisque ce dernier ne peut simuler un repliement ou prédire des structures intermédiaires.

Mfold et CA-RNA utilisent des approches totalement différentes et visent des objectifs différents. Le modèle CA-RNA se distingue de Mfold avant tout par le fait qu'il simule une dynamique de repliement. Les temps de traitement de CA-RNA sont plus longs que les temps de traitement de Mfold, ce qui est normal puisque Mfold ne fait que de la prédiction statique. Pour ce qui est de la qualité de la prédiction, sur les 20 molécules utilisées, Mfold n'a pu prédire correctement la structure S_{native} de deux molécules et la structure $S_{hybride}$ de cinq molécules. Avec CA-RNA, les résultats ont été concluants avec les 20 molécules.

5 Conclusion

Dans le cadre de cette recherche, nous avons démontré qu'il était possible d'obtenir les structures secondaires de molécules d'ARN par un simple processus dynamique d'auto-organisation, ce qui constitue une percée dans le domaine de la prédiction structurale. Sans aucun algorithme d'optimisation, sans aucune information sur la structure native, sans partition ou réduction de l'espace des conformations secondaires, le modèle de simulation a réussi, pour les 20 molécules étudiées, à faire émerger de façon significative la structure native en conservant une paramétrisation à valeur identique dans tous les cas. À partir de la forme étirée de la molécule, le modèle a produit, après un certain temps, un état d'équilibre duquel émerge la structure secondaire native. L'émergence de la structure native est le résultat d'un processus d'auto-organisation généré uniquement par des interactions locales. À notre connaissance, c'est la première fois que des structures secondaires natives sont obtenues par un processus d'auto-organisation.

Le développement de ce modèle a nécessité la mise au point d'un nouveau concept nommé « force relative de rétention ». Cette force relative de rétention se veut une représentation approximative, sous forme de pourcentage, de l'effet global des forces d'attraction et de répulsion impliquées au niveau moléculaire de l'ARN. Son comportement s'est avéré conforme aux observations générales faites sur la dénaturation de l'ARN.

Notre modèle est un modèle simplifié et, comme tous les modèles simplifiés, il tente de reproduire l'essentiel d'un comportement en utilisant un minimum d'efforts et de ressources. Son objectif n'est donc pas de pouvoir traiter tous les cas particuliers que le problème peut présenter. Malgré cela, le modèle s'est avéré être d'une très grande souplesse, acceptant autant les structures usuelles, les structures avec pseudonoeuds, que les structures hybrides à deux et trois brins.

L'admission des pseudonoeuds augmente de façon très importante la cardinalité de l'espace des conformations. Pour la plupart des méthodes utilisées actuellement, elle accroit également la complexité du traitement mathématique. Dans notre modèle, le traitement des pseudonoeuds ne nécessite aucune modification particulière et il n'a pas d'impact sur la durée des simulations. Nous avons vu que, même en permettant les pseudonoeuds pour une molécule donnée, la structure native est demeurée émergente de façon significative.

Par ailleurs, avec les méthodes mathématiques, l'hybridation n'est pas toujours possible. Si elle l'est, il s'agit d'un cas particulier qui nécessite un traitement supplémentaire. Notre modèle considère indifféremment le repliement d'une seule chaîne ou l'hybridation de plusieurs brins. Nous avons effectué des simulations sur sept molécules à deux brins et, à chaque fois, c'est la structure native qui a émergé. Nous avons poussé l'expérience jusqu'à faire une simulation avec une molécule à trois brins et, encore là, nous avons obtenu l'hybridation et l'émergence de la structure native.

Parmi les autres méthodes de dynamique moléculaire, il n'y a que celle de Ding et al. [49] qui permet de simuler le repliement de molécules d'ARN sans aucune connaissance préalable de la structure native. Toutefois, l'approche utilisée est différente de la nôtre. Les simulations de Ding sont des processus qui sont orientés vers la recherche de la structure présentant le minimum d'énergie alors que nos simulations sont des processus d'auto-organisation, ce qui permet, entre autres, de simuler facilement le processus d'hybridation. De plus, la simplicité de notre modèle nous permet d'obtenir des temps de traitement inférieurs à ceux de Ding.

En outre, comme notre modèle fournit la totalité des structures rencontrées lors d'une simulation, nous nous sommes intéressés au chemin entre la forme étirée et la structure native, même si cela ne faisait pas partie intégrante de la portée de la recherche. Dans les simulations effectuées, nous avons pu observer que même s'il

existe plusieurs chemins pour une molécule donnée, ceux-ci présentent de bonnes similitudes. De plus, dans nos simulations avec une molécule contenant des pseudonoeuds (PK5), deux des structures dominantes étaient pratiquement identiques aux structures intermédiaires actuellement considérées comme importantes dans le repliement de cette molécule. Ces quelques constats laissent entrevoir la possibilité d'utiliser ce modèle pour des analyses de parcours et pour l'étude des structures intermédiaires.

Notre modèle est le premier de son genre, ce qui laisse place à beaucoup d'améliorations et d'innovations. Ainsi, les règles pourraient être modifiées afin de permettre un plus grand rapprochement des nucléotides lors des situations d'empilement, ce qui augmenterait l'effet stabilisateur. Le modèle pourrait également être modifié pour simuler des chaînes polypeptidiques ou d'autres types d'interactions moléculaires. Finalement, lors des simulations, les structures secondaires ont souvent adopté des conformations 3D faisant penser à des formes hélicoïdales. Il serait intéressant, quoique beaucoup plus difficile, d'ajouter les types de liens impliqués dans les structures tertiaires et voir si la modélisation tertiaire est possible.

Références

[1] J.D. Watson et F.H.C. Crick (1953). A structure for Deoxyribose Nucleic Acid. *Nature*, **171**, p. 737.

[2] H.R. Horton, L.A. Moran, R.S. Ochs, J.D. Rawn et K.G. Scrimgeour (1994). *Principes de Biochimie*, traduit par C. François. De Boeck-Wesmael, Bruxelles.

[3] R. Cedergren et F. Major (1998). Modeling the Tertiary Structure of RNA. In: *RNA Structure and Function*, Monograph 35, Simons et Grunberg-Manago ed.,Cold Spring Harbor Laboratory Press, New York, pp. 37—75.

[4] B. Ostrovsky, G. Crooks, M.A. Smith et Y. Bar-Yam (2001). Cellular automata for polymer simulation with application to polymer melts and polymer collapse including implications for protein folding. *Parallel Computing*, **27**, pp. 613–641.

[5] D.G. Shirvanyanz, A.S. Pavlov et P.G. Khalatur (2000). Self-organization of comblike copolymers with end-functionalized side chains: A cellular-automaton-based simulation. *J. Chem. Phys.*, **112**, pp. 11069–11079.

[6] L.B. Kier, C.-K. Cheng et B. Testa (2003). Studies of Ligand Diffusion Pathways over a Protein Surface. *J. Chem. Inf. Comput. Sci.*, **43**, pp. 255–258.

[7] D.L. Mobley, D.L. Cox, R.R.P. Singh, R.V. Kulkarni et A. Slepoy (2003). Simulations of Oligomeric Intermediates in Prion Diseases. *Biophysical Journal*, **85**, pp. 2213–2223.

[8] X. Xiao, S. Shao, Y. Ding, Z. Huang et K.-C. Chou (2006). Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. *Amino Acids*, **30**, pp. 49–54.

[9] Y. Diao, D. Ma, Z. Wen, J. Yin, J. Xiang et M. Li (2008). Using pseudo amino acid composition to predict transmembrane regions in protein: cellular automata and Lempel-Ziv complexity. *Amino Acids*, **34**, pp. 111–117.

[10] T.M. Devlin (2002). *Textbook of Biochemistry with Clinical Correlations*, fifth edition. Wiley-Liss, New-York.

[11] N.A. Campbell (1995). *Biologie*, adaptation par R. Mathieu. Édition du Renouveau Pédagogique, Saint-Laurent, Québec.

[12] V.A. Bloomfield, D.M. Crothers et I. Tinoco Jr (2000). *Nucleic Acids: Structures, Properties, and Functions*, University Science Books, California.

[13] C.B. Afinsen (1973). Principles that Govern the Folding of Protein Chains. *Science*, **181**(4096), pp. 223–230.

[14] L.M. Gierasch et J. King (1990). *Protein Folding: Deciphering the second half of the genetic code*. American Association for the Advancement of Science, Washington, DC.

[15] P. Sarkar (2000). A Brief History of Cellular Automata. *ACM Computing Surveys*, **32**(1), pp. 80–107.

[16] A.W. Burks (1970). *Essays on Cellular Automata*. University of Illinois Press, USA.

[17] M. Gardner (1970). The fantastic combinations of John Conway's new solitaire game "Life". *Scientific American.*, **223**, pp. 120–123.

[18] M. Gardner (1971). On cellular automata, self-reproduction, the Garden of Eden and the game of "Life". *Scientific American.*, **224**, pp. 112–117.

[19] S. Wolfram (1986). *Theory and Applications of Cellular Automata*. World Scientific Publishing, Singapore.

[20] C.S. Lent, B. Isaksen et M. Lieberman (2003). Molecular Quantum-Dot Cellular Automata. *J. Am. Chem. Soc.*, **125**, pp. 1056–1063.

[21] J.D. Brewster (2007). Lattice-Boltzmann Simulations of Three-Dimensional Fluid Flow on a Desktop Computer. *Anal. Chem.*, **79**, pp. 2965–2971.

[22] G. Wainer (2006). Applying Cell-DEVS Methodology for Modeling the Environment. *Simulation*, **82**, pp. 635–660.

[23] D.A. Wolf-Gladrow (2000). *Lattice-Gas Cellular Automata and Lattice Boltzmann Models*, Lecture Notes in Mathematics 1725. Springer-Verlag, Berlin, Germany.

[24] P.P. Chaudhuri, D.R. Chowdhury, S. Nandi , S. Chattopadhyay (1997). *Additive Cellular Automata: Theory and Applications Volume 1*. IEEE Computer Society, Los Alamitos, California.

[25] J. Walleczek et al. (2000). *Self-Organized Biological Dynamics & Nonlinear Control*. Cambridge University Press, Cambridge, UK.

[26] S. Camazine, J.-L. Deneubourg, N.R. Franks, J. Sneyd, G. Theraulaz, E. Bonabeau (2001). *Self-Organization in Biological Systems*. Princeton University Press, Princeton, New Jersey.

[27] F. Major et R. Griffey (2001). Computational methods for RNA structure determination. *Current Opinion in Structural Biology*, **11**, pp. 282–286.

[28] B.A. Shapiro et al. (2007). Bridging the gap in RNA structure prediction. *Cur. Op. Struct. Bio.*, **17**, pp. 157–165.

[29] F. Major, M. Turcotte, D. Gautheret, G. Lapalme, E. Fillion et R. Cedergren (1991). The Combination of Symbolic and Numerical Computation for Three-Dimensional Modeling of RNA. *Science*, **253**, pp. 1255–1260.

[30] D. Gautheret, F. Major et R. Cedergren (1993). Modeling the Three-dimensional Structure of RNA Using Discrete Nucleotide Conformational Sets. *Journal of Molecular Biology*, **229**, pp. 1049–1064.

[31] F. Major, R. Feldmann, G. Lapalme et R. Cedergren (1988). Reproducing the three-dimensional structure of tRNA molecule from structural constraints. *Proc. Natl. Acad. Sci. USA*, **90**, pp. 9408–9412.

[32] M. Zuker (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, **31**, pp. 3406–3415.

[33] S. Cao et S.-J. Chen (2007). Biphasic Folding Kinetics of RNA Pseudoknots and Telomerase RNA Activity. *J. Mol. Biol.*, **367**, pp. 909–924.

[34] A. Xayaphoummine, T. Bucher et H. Isambert (2005). Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Research*, **33**, pp. w605–w610.

[35] J.D. Madura et al. (1995). Electrostatics and diffusion of molecules in solutions: simulations with the University of Houston Brownian Dynamics program. *Computer Physics Communications*, **91**, pp. 57–95.

[36] J.A. McCammon, B.R. Gelin et M. Karplus (1977). Dynamics of folded proteins. *Nature*, **267**, pp. 585–590.

[37] O. Jardetzky et R. Holbrook (1989). *Protein Structure and Engineering*, NATO ASI Series A: Life Sciences Vol. 183. Plenum Press, New York, NY.

[38] W.F. Van Gunsteren, P.H. Hünenberger, A.E. Mark, P.E. Smith et I.G. Tironi (1995). Computer simulation of protein motion. *Computer Physics Communications*, **91**, pp. 305–319.

[39] R. Lavery, K. Zakrzewska et H. Sklenar (1995). JUMNA (junction minimisation of nucleic acids). *Computer Physics Communications*, **91**, pp. 135–158.

[40] D. Van Belle et S.J. Wodak (1995). Extended Lagrangian formalism applied to temperaure control and electronic polarization effects in molecular dynamics simulations. *Computer Physics Communications*, **91**, pp. 215–231.

[41] D.A. Pearlman et al. (1995). AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications*, **91**, pp. 1–41.

[42] H. Wako, S. Endo, K. Nagayama et N. Go (1995). FEDER/2: program for static and dynamic conformational energy analysis of macro-molecules in dihedral angle space. *Computer Physics Communications*, **91**, pp. 233–251.

[43] A.P. Lemon, P. Dauber-Osguthorpe et D.J. Osguthorpe (1995). FOCUS: a molecular dynamics analysis program. *Computer Physics Communications*, **91**, pp. 97–109.

[44] H.J.C. Berendsen, D. Van Der Spoel et R. Van Drunen (1995). GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, **91**, pp. 43–56.

[45] M. Nelson et al. (1995). MDScope - a visual computing environment for structural biology. *Computer Physics Communications*, **91**, pp. 111–133.

[46] R. Elber et al. (1995). MOIL: A program for simulations of macromolecules. *Computer Physics Communications*, **91**, pp. 159–189.

[47] G.R. Kneller, V. Keiner, M. Kneller et M. Schiller (1995). nMOLDYN: A program package for a neutron scattering oriented analysis of Molecular Dynamics simulations. *Computer Physics Communications*, **91**, pp. 233–251.

[48] F. Ding et N.V. Dokholyan. (2005). Simple but predictive protein models. *TRENDS in Biotechnology*, **23**, pp. 450–455.

[49] F. Ding et al. (2008). Ab initio RNA folding by discrete molecular dynamics: From structure prediction to folding mechanisms. *RNA*, **14**, pp. 1164–1173.

[50] L.B. Kier (2000). A Cellular Automata Model of Bond Interactions Among Molecules. *J. Chem. Inf. Comput. Sci.*, **40**, pp. 1285–1288.

[51] M. Zuker (2000). Calculating nucleic acid secondary structure. *Cur. Op. Struct. Bio.*, **10**, pp. 303–310.

[52] R.C. Penner et M.S. Waterman (1993). Spaces of RNA secondary structures. *Adv. Math.*, **101**, pp. 31–49.

[53] M.S. Waterman et T.F. Smith (1978). RNA secondary structure: A complete mathematical analysis. *Math. Biosc.*, **42**, pp. 257–266.

[54] M. Zuker et D. Sankoff (1984). RNA secondary structures and their prediction. *Bul. Math. Bio.*, **46**, pp. 591–621.

[55] J. Cupal, I.L. Hofacker et P.F. Stadler (1996). Dynamic programming algorithm for the density of states of RNA secondary structures. *Computer Science and Biology 96*, pp. 184–186.

[56] J. Cupal, C. Flamm, A. Renner et P.F. Stadler (1997). Density of States, Metastable States, and Saddle Points Exploring the Energy Landscape of RNA Molecule. *Proceedings of the ISMB-97*, pp. 88–91.

[57] I.L. Hofacker, P. Schuster et P.F. Stadler (1998). Combinatorics of RNA secondary structures. *Disc. App. Math.*, **88**, pp. 207–237.

[58] B. Liao et T.-M. Wang (2004). General combinatorics of RNA secondary structure. *Math. Biosc.*, **191**, pp. 69–81.

[59] X. Tang, B. Kirkpatrick, S. Thomas, G. Song et N.M. Amato (2005). Using Motion Planning to Study RNA Folding Kinetics. *Journal of Computational Biology*, **12**, pp. 862–881.

[60] R. Zwanzig, A. Szabo et B. Bagchi (1992). Levinthal's paradox. *Proc. Natl. Acad. Sci. USA*, **89**, pp. 20–22.

[61] B.N.M. van Buuren, T. Hermann, S.S. Wijmenga et E. Westhof (2002). Brownian-dynamics simulations of metal-ion binding to four-way junctions. *Nucleic Acids Research*, **30**, pp. 507–514.

[62] I. Chang, E.S. Gilbert, N. Eliashberg et J.D. Keasling (2003). A three-dimensional, stochastic simulation of biofilm growth and transport-related factors that affect structure. *Microbiology*, **149**, pp. 2859–2871.

[63] H. Soula et al. (2005). Modeling the emergence of multi-protein dynamic structures by principles of self-organization through the use of 3DSpi, a multi-agent-based software. *BMC Bioinformatics*, **6**, pp. 228.

[64] C. Hyeon et D. Thirumalai (2006). Forced-Unfolding and Force-Quench Refolding of RNA Hairpins. *Biophysical Journal*, **90**, pp. 3410–3427.

[65] C. Chen et al. (2007). Modeling of the Role of a Bax-Activation Switch in the Mitochondrial Apoptosis Decision. *Biophysical Journal*, **92**, pp. 4304–4315.

[66] C. Hyeon et D. Thirumalai (2007). Mechanical Unfolding of RNA: From Hairpins to Structures with Internal Multiloops. *Biophysical Journal*, **92**, pp. 731–743.

[67] M. Baiesi, E. Orlandini et A.L. Stella (2003). RNA Denaturation: Excluded Volume, Pseudoknots, and Transition Scenarios. *Phys. Rev. Lett.*, **91**, pp. 198102.1–198102.4.

[68] R. Backofen (2004). A polynomial time upper bound for the number of contacts in the HP-model on the face-centered-cubic lattice (FCC). *Journal of Discrete Algorithms*, **2**, pp. 161–206.

[69] J. Ruan, G.D. Stormo et W. Zhang (2004). An Iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, **20**, pp. 58–66.

[70] J. Ren, B. Rastegari, A. Condon et H.H. Hoos (2005). HotKnots: Heuristic prediction of RNA secondary structures including pseudoknots. *RNA*, **11**, pp. 1494–1504.

[71] Y.-J. Sheng, Y.-C. Mou et H.-K. Tsao (2006). Conformational entropy of a pseudoknot polymer. *J. Chem. Phys.*, **124**, 124904.

[72] J. Reeder, P. Steffen et R. Giegerich (2007). pknotsRG: RNA pseudoknot folding including near-optimal structures and sliding windows. *Nucleic Acids Research*, **35**, pp. w320–w324.

[73] E.Y. Jin et C.M. Reidys (2007). Asymptotic Enumeration of RNA Structures with Pseudoknots. *Bul. Math. Bio.*, 9265.

[74] J.D. Puglisi, J.R. Wyatt et I. Tinoco Jr (1990). Conformation of an RNA Pseudoknot. *J. Mol. Biol.*, **214**, pp. 437–453.

[75] W. Zhang et S.-J. Chen (2006). Exploring the Complex Folding Kinetics of RNA Hairpins: I. General Folding Kinetics Analysis. *Biophysical Journal*, **90**, pp. 765–777.

[76] J.R. Wyatt, J.D. Puglisi et I. Tinoco Jr (1990). RNA Pseudoknots Stability and Loop Size Requirements. *J. Mol. Biol.*, **214**, pp. 455–470.

Bibliographie

Afinsen C.B. (1973). Principles that Govern the Folding of Protein Chains. *Science*, **181**(4096), pp. 223–230.

Alberts B., Johnson A., Lewis J., Raff M. , Roberts K. et Walter P. (2002). *Molecular Biology of the Cell*, fourth edition. Garland Science, New York, NY.

Backofen R. (2004). A polynomial time upper bound for the number of contacts in the HP-model on the face-centered-cubic lattice (FCC). *Journal of Discrete Algorithms*, **2**, pp. 161–206.

Baiesi M., Orlandini E. et Stella A.L. (2003). RNA Denaturation: Excluded Volume, Pseudoknots, and Transition Scenarios. *Phys. Rev. Lett.*, **91**, pp. 198102.1–198102.4.

Berendsen H.J.C., Van Der Spoel D. et Van Drunen R. (1995). GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, **91**, pp. 43–56.

Berman H.M. et al. (2000). The Protein Data Bank. *Nucleic Acids Research*, **28**(1), pp. 235–242.

Bevilacqua P.C. et Blose J.M. (2008). Structures, Kinetics, Thermodynamics, and Biological Functions of RNA Hairpins. *Annu. Rev. Phys. Chem.*, **58**, pp. 79–103.

Bloomfield V.A., Crothers D.M. et Tinoco Jr I. (2000). *Nucleic Acids*: *Structures, Properties, and Functions*, University Science Books, California.

Brewster J.D. (2007). Lattice-Boltzmann Simulations of Three-Dimensional Fluid Flow on a Desktop Computer. *Anal. Chem.*, **79**, pp. 2965–2971.

Brion P. et Westhof E. (1997). Hierarchy and Dynamics of RNA Folding. *Annu. Rev. Biophys. Biomol. Struct.*, **26**, pp. 113–137.

Bundschuh R. et Hwa T. (2002). Statistical mechanics of secondary structures formed by random RNA sequences. *Phys. Rev. E*, **65**, 031903.

Burks A.W. (1970). *Essays on Cellular Automata*. University of Illinois Press, USA.

Bustamante C. (2005). Unfolding single RNA molecules: bridging the gap between equilibrium and non-equilibrium statistical thermodynamics. *Quartely Reviews of Biophysics*, **38**, pp. 291–301.

Camazine S., Deneubourg J.-L., Franks N.R., Sneyd J., Theraulaz G., Bonabeau E. (2001). *Self-Organization in Biological Systems*. Princeton University Press, Princeton, New Jersey.

Campbell N.A. (1995). *Biologie*, adaptation par R. Mathieu. Édition du Renouveau Pédagogique, Saint-Laurent, Québec.

Cao S. et Chen S.-J. (2005). Predicting RNA folding thermodynamics with a reduced chain representation model. *RNA*, **11**, pp. 1884–1897.

Cao S. et Chen S.-J. (2006). Predicting RNA pseudoknot folding thermo-dynamics. *Nucleic Acids Research*, **34**, pp. 2634–2652.

Cao S. et Chen S.-J. (2007). Biphasic Folding Kinetics of RNA Pseudoknots and Telomerase RNA Activity. *J. Mol. Biol.*, **367**, pp. 909–924.

Cedergren R., Gautheret D., Lapalme G. et Major F. (1988). A secondary and tertiary structure editor for nucleic acids. *Computer Applications in the BioSciences*, **4**(1), pp. 143–146.

Cedergren R. et Major F. (1998). Modeling the Tertiary Structure of RNA. In: *RNA Structure and Function*, Monograph 35, Simons et Grunberg-Manago ed.,Cold Spring Harbor Laboratory Press, New York, pp. 37—75.

Chang I., Gilbert E.S., Eliashberg N. et Keasling J.D. (2003). A three-dimensional, stochastic simulation of biofilm growth and transport-related factors that affect structure. *Microbiology*, **149**, pp. 2859–2871.

Chaudhuri P.P., Chowdhury D.R., Nandi S., Chattopadhyay S. (1997). *Additive Cellular Automata*: *Theory and Applications Volume 1*. IEEE Computer Society, Los Alamitos, California.

Chen C. et al. (2007). Modeling of the Role of a Bax-Activation Switch in the Mitochondrial Apoptosis Decision. *Biophysical Journal*, **92**, pp. 4304–4315.

Chen C. et Xiao Y. (2008). Observation of multiple folding pathways of β-hairpin trpzip2 from independent continuous folding trajectories. *Bioinformatics*, **24**, pp. 659–665.

Chen S.-J. et Dill K.A. (2000). RNA folding energy landscapes. *Proc. Natl. Acad. Sci. USA*, **97**, pp. 646–651.

Cheng C.-K. et Kier L.B. (1995). A Cellular Automata Model of Oil-Water Partitioning. *J. Chem. Inf. Comput. Sci.*, **35**, pp. 1054–1059.

Cheong C., Varani G. et Tinoco Jr I. (1990). Solution structure of an unusually stable RNA hairpin, 5'GGAC(UUCG)GUCC. *Nature*, **346**, pp. 680–682.

Chou H.-H. et Huang W. (2002). The Trend Cellular Automata Programming Environment. *Simulation*, **78**, pp. 59–75.

Cupal J., Hofacker I.L. et Stadler P.F. (1996). Dynamic programming algorithm for the density of states of RNA secondary structures. *Computer Science and Biology 96*, pp. 184–186.

Cupal J., Flamm C., Renner A. et Stadler P.F. (1997). Density of States, Metastable States, and Saddle Points Exploring the Energy Landscape of RNA Molecule. *Proceedings of the ISMB-97*, pp. 88–91.

Demongeot J., Golès E., Tchuente M. (1985). *Dynamical Systems and Cellular Automata*. Academic Press, Orlando, Florida.

Deng N.-J. et Cieplak P. (2007). Molecular Dynamics and Free Energy Study of the Conformational Equilibria in the UUUU RNA Hairpin. *J. Chem. Theory Comput.*, **3**, pp. 1435–1450.

Devlin T.M. (2002). *Textbook of Biochemistry with Clinical Correlations*, fifth edition. Wiley-Liss, New-York.

Diao Y., Ma D., Wen Z., Yin J., Xiang J. et Li M. (2008). Using pseudo amino acid composition to predict transmembrane regions in protein: cellular automata and Lempel-Ziv complexity. *Amino Acids*, **34**, pp. 111–117.

Dill K.A. et Chan H.S. (1997). From Levinthal to pathways to funnels. *Nature Structural Biology*, **4**, pp. 10–19.

Dimitrov R.A. et Zuker M. (2004). Prediction of Hybridization and Melting for Double-Stranded Nucleic Acids. *Biophysical Journal*, **87** pp. 215–226.

Ding F. et Dokholyan N.V. (2005). Simple but predictive protein models. *TRENDS in Biotechnology*, **23**, pp. 450–455.

Ding F. et al. (2008). Ab initio RNA folding by discrete molecular dynamics: From structure prediction to folding mechanisms. *RNA*, **14**, pp. 1164–1173.

Ding J., Carver T.J. et Windle A.H. (2001). Self-assembled structures of block copolymers in selective solvents reproduced by lattice Monte Carlo simulation. *Computational and Theoretical Polymer Science*, **11**, pp. 483–490.

Ding Y. et Lawrence C.E. (2003). A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, **31** pp. 7280–7301.

Ding Y., Chan C.Y. et Lawrence C.E. (2005). RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA*, **11** pp. 1157–1166.

Dugas H. (1996). *Principes de Base en Modélisation Moléculaire*, quatrième édition. La librairie de l'Université de Montréal, Canada.

Eddy S.R. (2004). How do RNA folding algorithms work ?. *Nature*, **22**, pp. 1457–1458.

Elber R. et al. (1995). MOIL: A program for simulations of macromolecules. *Computer Physics Communications*, **91**, pp. 159–189.

Finkelstein A.V. (1997). Protein structure: what is it possible to predict now? *Current Opinion in Structural Biology*, **7**, pp. 60–71.

Flamm C., Fontana W., Hofacker I.L. et Schuster P. (2000). RNA folding at elementary step resolution. *RNA*, **6**, pp. 325–338.

Galzitskaya O.V. et Finkelstein A.V. (1996). Computer simulation of secondary structure folding of random and ''edited'' RNA chains. *J. Chem. Phys.*, **105**, pp. 319–325.

Garcia A.E. et Paschek D. (2008). Simulation of the Pressure and Temperature Folding/Unfolding Equilibrium of a Small RNA Hairpin. *J. Am. Chem. Soc.*, **130**, pp. 815–817.

Gardner M. (1970). The fantastic combinations of John Conway's new solitaire game "Life". *Scientific American.*, **223**, pp. 120–123.

Gardner M. (1971). On cellular automata, self-reproduction, the Garden of Eden and the game of "Life". *Scientific American.*, **224**, pp. 112–117.

Gautheret D., Major F. et Cedergren R. (1993). Modeling the Three-dimensional Structure of RNA Using Discrete Nucleotide Conformational Sets. *Journal of Molecular Biology*, **229**, pp. 1049–1064.

Gierasch L.M. et King J. (1990). *Protein Folding*: *Deciphering the second half of the genetic code*. American Association for the Advancement of Science, Washington, DC.

Gillespie D.T. (1977). Exact Stochastic Simulation of Coupled Chemical Reactions. *J. Phys. Chem.*, **81**, pp. 2340–2361.

Gisiger T. (2001). Scale invariance in biology: coincidence or footprint of a universal mechanism. *Biol. Rev.*, **76**, pp. 161–209.

Gultyaev A.P., Van Batenburg F.H.D. et Pleij C.W.A. (1995). The Computer Simulation of RNA Folding Pathways Using a Genetic Algorithm. *Journal of Molecular Biology*, **250**, pp. 37–51.

Haire K.R., Carver T.J. et Windle A.H. (2001). A Monte Carlo lattice model for chain diffusion in dense polymer systems and its interlocking with molecular dynamics simulation. *Computer and Theoretical Polymer Science*, **11**, pp. 17–28.

Hilke J., Reggia J., Navarro-Gonzalez R. et Lohn J. (1995). A Modified Cellular Automata Model of Nucleotide Interactions and Non-Enzymatic Transcription of DNA. In: *Proceedings of the First International Symposium on Intelligence in Neural and Biological Systems*, IEEE, pp. 136—144.

Hofacker I.L. (1998). RNA Secondary Structures: A Tractable Model of Biopolymer Folding. *J. Theor. Biol.*, **212**, pp. 35–46.

Hofacker I.L., Schuster P. et Stadler P.F. (1998). Combinatorics of RNA secondary structures. *Disc. App. Math.*, **88**, pp. 207–237.

Horton H.R., Moran L.A., Ochs R.S., Rawn J.D. et Scrimgeour K.G. (1994). *Principes de Biochimie*, traduit par C. François. De Boeck-Wesmael, Bruxelles.

Hyeon C. et Thirumalai D. (2006). Forced-Unfolding and Force-Quench Refolding of RNA Hairpins. *Biophysical Journal*, **90**, pp. 3410–3427.

Hyeon C. et Thirumalai D. (2007). Mechanical Unfolding of RNA: From Hairpins to Structures with Internal Multiloops. *Biophysical Journal*, **92**, pp. 731–743.

Jacob C., Breton N. et Daegelen P. (1997). Stochastic theories of the activated complex and the activated collision: The RNA example. *J. Chem. Phys.*, **107**, pp. 2903–2912.

Jardetzky O. et Holbrook R. (1989). *Protein Structure and Engineering*, NATO ASI Series A: Life Sciences Vol. 183. Plenum Press, New York, NY.

Jin E.Y. et Reidys C.M. (2007). Asymptotic Enumeration of RNA Structures with Pseudoknots. *Bul. Math. Bio.*, 9265.

Kier L.B. et Cheng C.-K. (1994). A Cellular Automata Model of Water. *J. Chem. Inf. Comput. Sci.*, **34**, pp. 647–652.

Kier L.B. et Cheng C.-K. (1994). A Cellular Automata Model of an Aqueous Solution. *J. Chem. Inf. Comput. Sci.*, **34**, pp. 1334–1337.

Kier L.B., Cheng C.-K., Tute M. et Seybold P.G. (1998). A Cellular Automata Model of Acid Dissociation. *J. Chem. Inf. Comput. Sci.*, **38**, pp. 271–275.

Kier L.B., Cheng C.-K. et Testa B. (1999). A Cellular Automata Model of the Percolation Process. *J. Chem. Inf. Comput. Sci.*, **39**, pp. 326–332.

Kier L.B. (2000). A Cellular Automata Model of Bond Interactions Among Molecules. *J. Chem. Inf. Comput. Sci.*, **40**, pp. 1285–1288.

Kier L.B., Cheng C.-K. et Testa B. (2003). Studies of Ligand Diffusion Pathways over a Protein Surface. *J. Chem. Inf. Comput. Sci.*, **43**, pp. 255–258.

Kneller G.R., Keiner V., Kneller M. et Schiller M. (1995). nMOLDYN: A program package for a neutron scattering oriented analysis of Molecular Dynamics simulations. *Computer Physics Communications*, **91**, pp. 233–251.

Kopeikin Z. (2006). *Statistical thermodynamics for RNA structures with simple tertiary contacts and pseudoknots*. PhD thesis, University of Missouri, Columbia, USA.

Lathrop R.H. (1994). The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Engineering*, **7**(9), pp. 1059–1068.

Lavery R., Zakrzewska K. et Sklenar H. (1995). JUMNA (junction minimisation of nucleic acids). *Computer Physics Communications*, **91**, pp. 135–158.

Lemon A.P., Dauber-Osguthorpe P. et Osguthorpe D.J. (1995). FOCUS: a molecular dynamics analysis program. *Computer Physics Communications*, **91**, pp. 97–109.

Lent C.S., Isaksen B. et Lieberman M. (2003). Molecular Quantum-Dot Cellular Automata. *J. Am. Chem. Soc.*, **125**, pp. 1056–1063.

Leoni P. et Vanderzande C. (2003). Statistical mechanics of RNA folding: A lattice approach. *Physical Review E*, **68**, 051904.

Liao B. et Wang T.-M. (2004). General combinatorics of RNA secondary structure. *Math. Biosc.*, **191**, pp. 69–81.

Liu F. et Ou-Yang Z.C. (2004). Unfolding single RNA molecules by mechanical force: A stochastic kinetic method. *Physical Review E*, **70**, 040901.

Ma H., Proctor D.J., Kierzek E., Kierzek R., Bevilacqua P.C. et Gruebele M. (2006). Exploring the Energy Landscape of a small RNA Hairpin. *J. Am. Chem. Soc.*, **128**, pp. 1523–1530.

Madura J.D. et al. (1995). Electrostatics and diffusion of molecules in solutions: simulations with the University of Houston Brownian Dynamics program. *Computer Physics Communications*, **91**, pp. 57–95.

Major F., Feldmann R., Lapalme G. et Cedergren R. (1988). FUS: a system to simulate conformational changes in biological macromolecules. *Computer Applications in the BioSciences*, **4**(4), pp. 445–451.

Major F., Feldmann R., Lapalme G. et Cedergren R. (1988). Reproducing the three-dimensional structure of tRNA molecule from structural constraints. *Proc. Natl. Acad. Sci. USA*, **90**, pp. 9408–9412.

Major F. (1990). *La prédiction des structures macromoléculaires par une approche symbolique*. Doctorat en informatique, Université de Montréal, Canada.

Major F., Turcotte M., Gautheret D., Lapalme G., Fillion E. et Cedergren R. (1991). The Combination of Symbolic and Numerical Computation for Three-Dimensional Modeling of RNA. *Science*, **253**, pp. 1255–1260.

Major F. et Griffey R. (2001). Computational methods for RNA structure determination. *Current Opinion in Structural Biology*, **11**, pp. 282–286.

Markham N.R. et Zuker M. (2005). DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Research*, **33**, pp. w577–w581.

Mathews D.H. (2006). Revolutions in RNA Secondary Structure Prediction. *J. Mol. Biol.*, **359**, pp. 526–532.

McCammon J.A., Gelin B.R. et Karplus M. (1977). Dynamics of folded proteins. *Nature*, **267**, pp. 585–590.

McDowell S.E., Spackova N., Sponer J. et Walter N.G. (2006). Molecular Dynamics Simulations of RNA: An In Silico Single Molecule Approach. *Biopolymers*, **85**, pp. 169–184.

Mobley D.L., Cox D.L., Singh R.R.P., Kulkarni R.V. et Slepoy A. (2003). Simulations of Oligomeric Intermediates in Prion Diseases. *Biophysical Journal*, **85**, pp. 2213–2223.

Mukhopadhyay R., Emberly E., Tang C. et Wingreen N.S. (2003). Statistical mechanics of RNA folding: Importance of alphabet size. *Physical Review E*, **68**, 041904.

Nelson M. et al. (1995). MDScope - a visual computing environment for structural biology. *Computer Physics Communications*, **91**, pp. 111–133.

Onoa B. et Tinoco I. (2004). RNA folding and unfolding. *Cur. Op. Struct. Bio.*, **14**, pp. 374–379.

Ostrovsky B., Crooks G., Smith M.A. et Bar-Yam Y. (2001). Cellular automata for polymer simulation with application to polymer melts and polymer collapse including implications for protein folding. *Parallel Computing*, **27**, pp. 613–641.

Pearlman D.A. et al. (1995). AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications*, **91**, pp. 1–41.

Penner R.C. et Waterman M.S. (1993). Spaces of RNA secondary structures. *Adv. Math.*, **101**, pp. 31–49.

Pretti M. (2006). RNA-like polymer model: Exact calculation on the Bethe lattice. *Physical Review E.*, **74**, 051803.

Puglisi J.D., Wyatt J.R. et Tinoco Jr I. (1990). Conformation of an RNA Pseudoknot. *J. Mol. Biol.*, **214**, pp. 437–453.

Pyle A.M. et Green J.B. (1995). RNA folding. *Cur. Op. Struct. Bio.*, **5**, pp. 303–310.

Reeder J., Steffen P. et Giegerich R. (2007). pknotsRG: RNA pseudoknot folding including near-optimal structures and sliding windows. *Nucleic Acids Research*, **35**, pp. w320–w324.

Ren J., Rastegari B., Condon A. et Hoos H.H. (2005). HotKnots: Heuristic prediction of RNA secondary structures including pseudoknots. *RNA*, **11**, pp. 1494–1504.

Ruan J., Stormo G.D. et Zhang W. (2004). An Iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, **20**, pp. 58–66.

Russell R. et al. (2006). The Paradoxical Behavior of a Highly Structured Misfolded Intermediate in RNA Folding. *J. Mol. Bio.*, **363**, pp. 531–544.

Sarkar P. (2000). A Brief History of Cellular Automata. *ACM Computing Surveys*, **32**(1), pp. 80–107.

Schuster P. (2006). Prediction of RNA secondary structures: from theory to models and real molecules. *Rep. Prog. Phys.*, **69**, pp. 1419–1477.

Seibert M.M., Patriksson A., Hess B. et Van Der Spoel D. (2005). Reproducible Polypeptide Folding and Structure Prediction using Molecular Dynamics Simulations. *J. Mol. Biol.*, **354**, pp. 173–183.

Seybold P.G., Kier L.B. et Cheng C.-K. (1997). Simulation of First-Order Chemical Kinetics Using Cellular Automata. *J. Chem. Inf. Comput. Sci.*, **37**, pp. 386–391.

Seybold P.G., Kier L.B. et Cheng C.-K. (1998). Stochastic Cellular Automata Models of Molecular Excited-State Dynamics. *J. Phys. Chem. A*, **102**, pp. 886–891.

Shapiro B.A. et al. (2007). Bridging the gap in RNA structure prediction. *Cur. Op. Struct. Bio.*, **17**, pp. 157–165.

Sheng Y.-J., Mou Y.-C. et Tsao H.-K. (2006). Conformational entropy of a pseudoknot polymer. *J. Chem. Phys.*, **124**, 124904.

Shirvanyanz D.G., Pavlov A.S. et Khalatur P.G. (2000). Self-organization of comblike copolymers with end-functionalized side chains: A cellular-automaton-based simulation. *J. Chem. Phys.*, **112**, pp. 11069–11079.

Siegfried N.A., Metzger S.L. et Bevilacqua P.C. (2007). Folding Cooperativity in RNA and DNA Is Dependent on Position in the Helix. *Biochemistry*, **46**, pp. 172–181.

Silverman S.K. (2008). A Forced March accross an RNA Folding Landscape. *Chemistry & Biology*, **15**, pp. 211–213.

Sorin E.J., Engelhardt M.A., Herschlag D. et Pande V.S. (2002). RNA Simulations: Probing Hairpin Unfolding and the Dynamics of a GNRA Tetraloop. *J. Mol. Biol.*, **317**, pp. 493–506.

Sosnick T.R. et Pan T. (2003). RNA folding: models and perspectives. *Cur. Op. Struct. Bio.*, **13**, pp. 309–316.

Soula H. et al. (2005). Modeling the emergence of multi-protein dynamic structures by principles of self-organization through the use of 3DSpi, a multi-agent-based software. *BMC Bioinformatics*, **6**, pp. 228.

Tang X., Kirkpatrick B., Thomas S., Song G. et Amato N.M. (2005). Using Motion Planning to Study RNA Folding Kinetics. *Journal of Computational Biology*, **12**, pp. 862–881.

Tang X. et al. (2007). Tools for Simulating and Analyzing RNA Folding Kinetics. *RECOMB 2007*, pp. 268–282.

Testa B., Kier L.B. et Cheng C.-K. (2002). A Cellular Automata Model of Water Structuring by a Chiral Solute. *J. Chem. Inf. Comput. Sci.*, **42**, pp. 712–716.

Thirumalai D. et Hyeon C. (2005). RNA and Protein Folding: Common Themes and Variations. *Biochemistry*, **44**, pp. 4957–4970.

Tinoco Jr I. et Bustamante C. (1999). How RNA Folds. *J. Mol. Biol.*, **293**, pp. 271–281.

Treiber D.K. et al. (1998). Kinetic Intermediates Trapped by Native Interactions in RNA Folding. *Science*, **279**, pp. 1943–1946.

Treiber D.K. et Williamson J.R. (1999). Exposing the kinetic traps in RNA folding. *Cur. Op. Struct. Bio.*, **9**, pp. 339–345.

Treiber D.K. et Williamson J.R. (2001). Beyond kinetic traps in RNA folding. *Cur. Op. Struct. Bio.*, **11**, pp. 309–314.

Van Belle D. et Wodak S.J. (1995). Extended Lagrangian formalism applied to temperaure control and electronic polarization effects in molecular dynamics simulations. *Computer Physics Communications*, **91**, pp. 215–231.

Van Buuren B.N.M., Hermann T., Wijmenga S.S. et Westhof E. (2002). Brownian-dynamics simulations of metal-ion binding to four-way junctions. *Nucleic Acids Research*, **30**, pp. 507–514.

Van Gunsteren W.F., Hünenberger P.H., Mark A.E., Smith P.E. et Tironi I.G. (1995). Computer simulation of protein motion. *Computer Physics Communications*, **91**, pp. 305–319.

Vieregg J.R. et Tinoco I. (2006). Modelling RNA folding under mechanical tension. *Mol. Phys.*, **104**, pp. 1343–1352.

Villa A., Widjajakusuma E. et Stock G. (2008). Molecular Dynamics Simulation of the Structure, Dynamics, and Thermostability of the RNA Hairpins uCACCg and cUUCGg. *J. Phys. Chem. B*, **112**, pp. 134–142.

Wainer G. (2006). Applying Cell-DEVS Methodology for Modeling the Environment. *Simulation*, **82**, pp. 635–660.

Wako H., Endo S., Nagayama K. et Go N. (1995). FEDER/2: program for static and dynamic conformational energy analysis of macro-molecules in dihedral angle space. *Computer Physics Communications*, **91**, pp. 233–251.

Walleczek J. et al. (2000). *Self-Organized Biological Dynamics & Nonlinear Control.* Cambridge University Press, Cambridge, UK.

Waterman M.S. et Smith T.F. (1978). RNA secondary structure: A complete mathematical analysis. *Math. Biosc.*, **42**, pp. 257–266.

Watson J.D. et Crick F.H.C. (1953). A structure for Deoxyribose Nucleic Acid. *Nature*, **171**, p. 737.

Weiner S.J. et al. (1984). A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins. *J. Am. Chem. Soc.*, **106**, pp. 765–784.

Wolfinger M.T. et al. (2004). Efficient computation of RNA folding dynamics. *J. Phys. A: Math. Gen.*, **37**, pp. 4731–4741.

Wolfram S. (1986). *Theory and Applications of Cellular Automata*. World Scientific Publishing, Singapore.

Wolf-Gladrow D.A. (2000). *Lattice-Gas Cellular Automata and Lattice Boltzmann Models*, Lecture Notes in Mathematics 1725. Springer-Verlag, Berlin, Germany.

Wyatt J.R., Puglisi J.D. et Tinoco Jr I. (1990). RNA Pseudoknots Stability and Loop Size Requirements. *J. Mol. Biol.*, **214**, pp. 455–470.

Xayaphoummine A., Bucher T. et Isambert H. (2005). Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Research*, **33**, pp. w605–w610.

Xiao X. et al. (2005). Using cellular automata to generate image represen-tation for biological sequences. *Amino Acids*, **28**, pp. 29–35.

Xiao X., Shao S., Ding Y., Huang Z. et Chou K.-C. (2006). Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. *Amino Acids*, **30**, pp. 49–54.

Xu G. et Mattice W.L. (2001). Study on structure formation of short polyethylene chains via dynamic Monte Carlo simulation. *Computational and Theoretical Polymer Science*, **11**, pp. 405–413.

Zara R.A. et Pretti M. (2007). Exact solution of a RNA-like polymer model on the Husimi lattice. *J. Chem. Phys.*, **127**, 184902.

Zarrinkar P.P. et Williamson J.R. (1994). Kinetic Intermediates in RNA Folding. *Science*, **265**, pp. 918–924.

Zhang W. et Chen S.-J. (2001). A three-dimensional statistical mechanical model of folding double-stranded chain molecules. *J. Chem. Phys.*, **114**, pp. 7669–7681.

Zhang W. et Chen S.-J. (2002). RNA hairpin-folding kinetics. *Proc. Natl. Acad. Sci. USA*, **99**, pp. 1931–1936.

Zhang W. et Chen S.-J. (2003). Analysing the biopolymer folding rates and pathways using kinetic cluster method. *J. Chem. Phys.*, **119**, pp. 8716–8729.

Zhang W. et Chen S.-J. (2006). Exploring the Complex Folding Kinetics of RNA Hairpins: I. General Folding Kinetics Analysis. *Biophysical Journal*, **90**, pp. 765–777.

Zhang W. et Chen S.-J. (2006). Exploring the Complex Folding Kinetics of RNA Hairpins: II. Effect of Sequence, Length, and Misfolded States. *Biophysical Journal*, **90**, pp. 778–787.

Zhu H., YH Pang P., Sun Y. et Dhar P. (2004). Asynchronous adaptive time step in quantitative cellular automata modeling. *BMC Bioinformatics*, **5**, 85.

Zhu H. et al. (2005). Cellular Automata With Object-Oriented Features for Parallel Molecular Network Modeling. *IEEE Trans. on Nanobioscience*, **4**, pp. 141–148.

Zuker M. et Sankoff D. (1984). RNA secondary structures and their prediction. *Bul. Math. Bio.*, **46**, pp. 591–621.

Zuker M. (2000). Calculating nucleic acid secondary structure. *Cur. Op. Struct. Bio.*, **10**, pp. 303–310.

Zuker M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, **31**, pp. 3406–3415.

Zwanzig R., Szabo A. et Bagchi B. (1992). Levinthal's paradox. *Proc. Natl. Acad. Sci. USA*, **89**, pp. 20–22.

Annexe A Structures émergentes pour les molécules étudiées

Cette annexe présente les structures secondaires émergentes des 20 molécules étudiées. Les figures sont des impressions d'écran de CA-RNA. L'affichage sur écran se fait en trois dimensions. Il peut arriver que l'effet de profondeur ne soit pas bien rendu à l'impression.

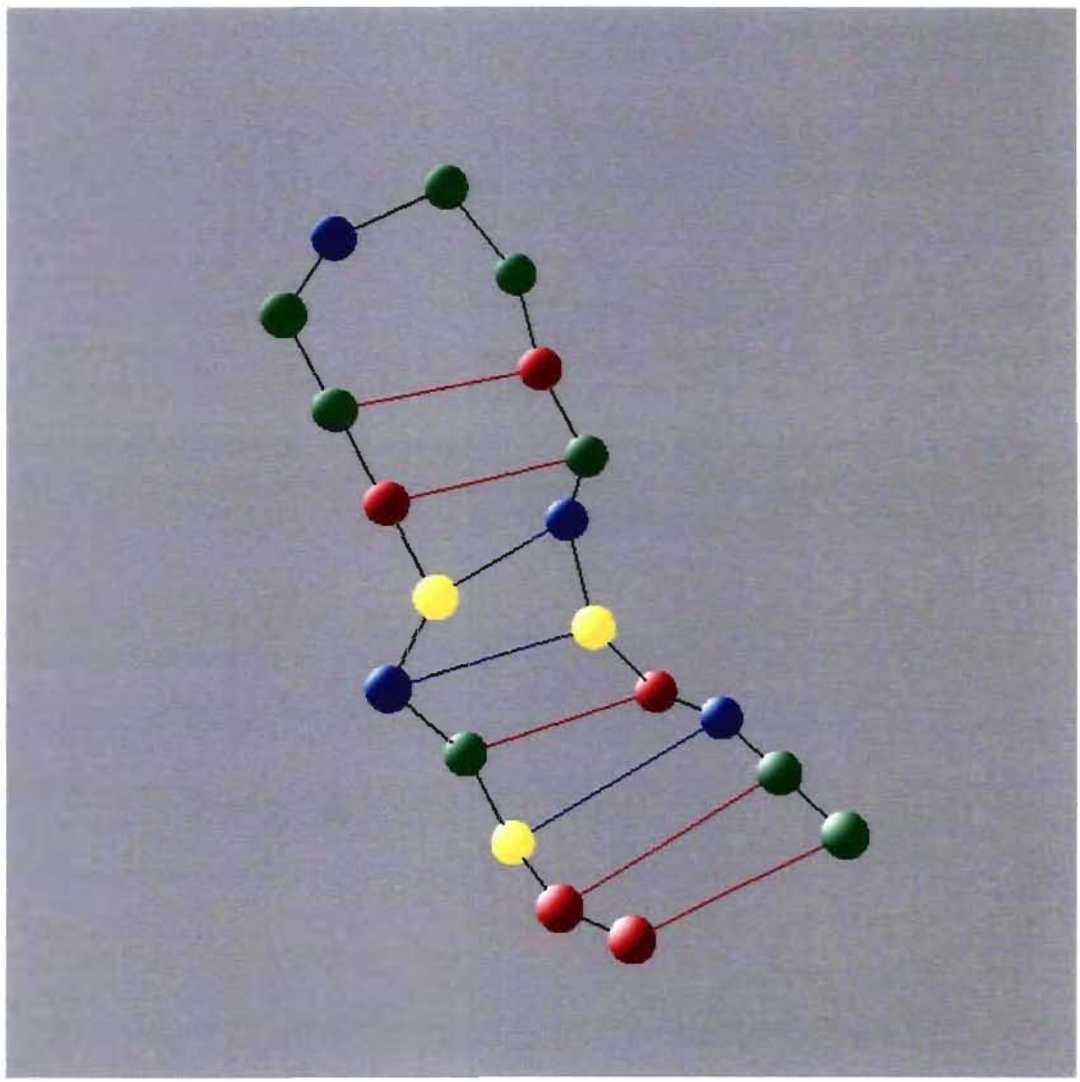

Figure A.1 Structure secondaire émergente avec 1BN0
Structure de 20 nucléotides avec cinq liens GC et trois liens AU.
(Code couleur : G = vert, C = rouge, A = bleu, U = jaune).

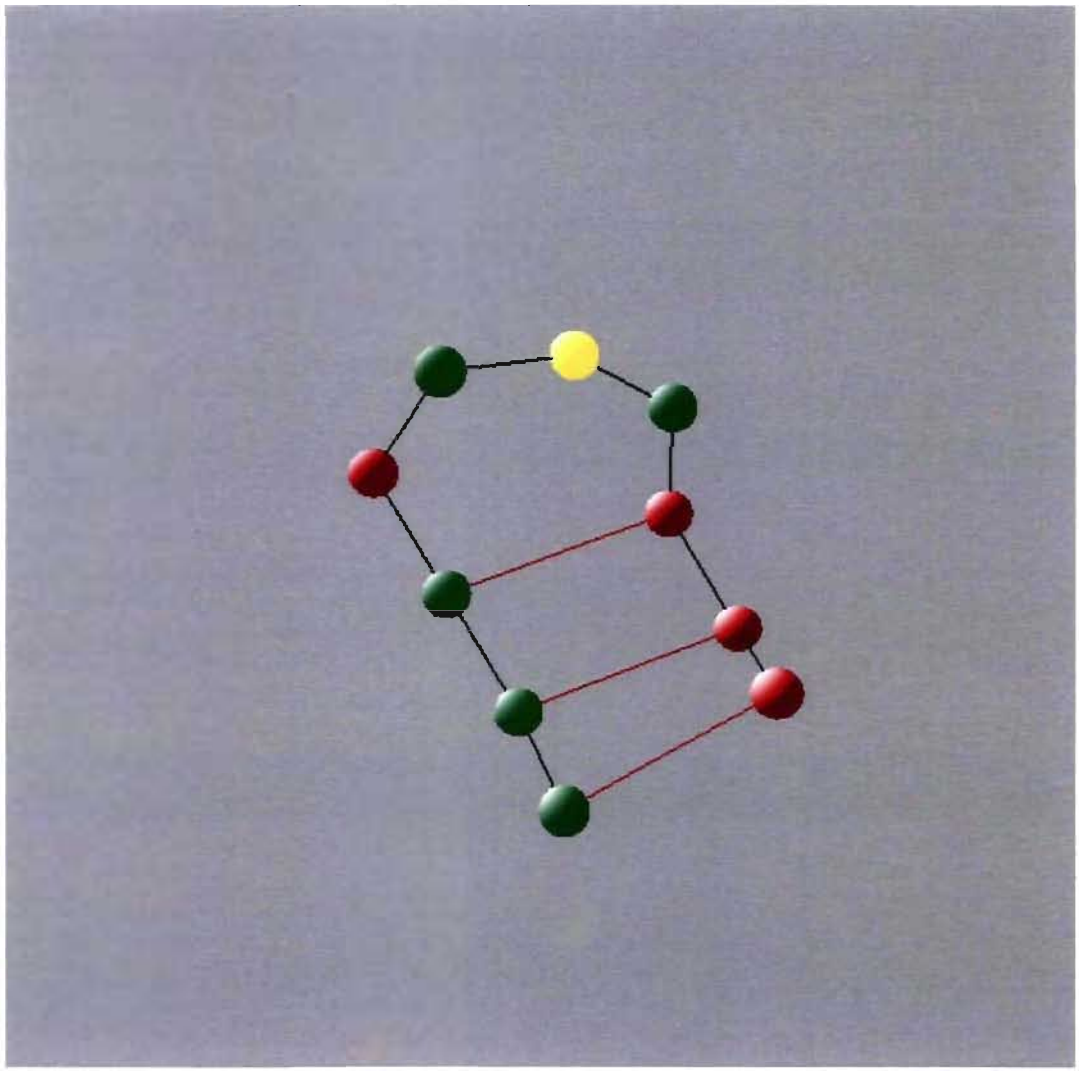

Figure A.2 Structure secondaire émergente avec 1IDV
Structure de 10 nucléotides avec trois liens GC.
(Code couleur : G = vert, C = rouge, A = bleu, U = jaune).

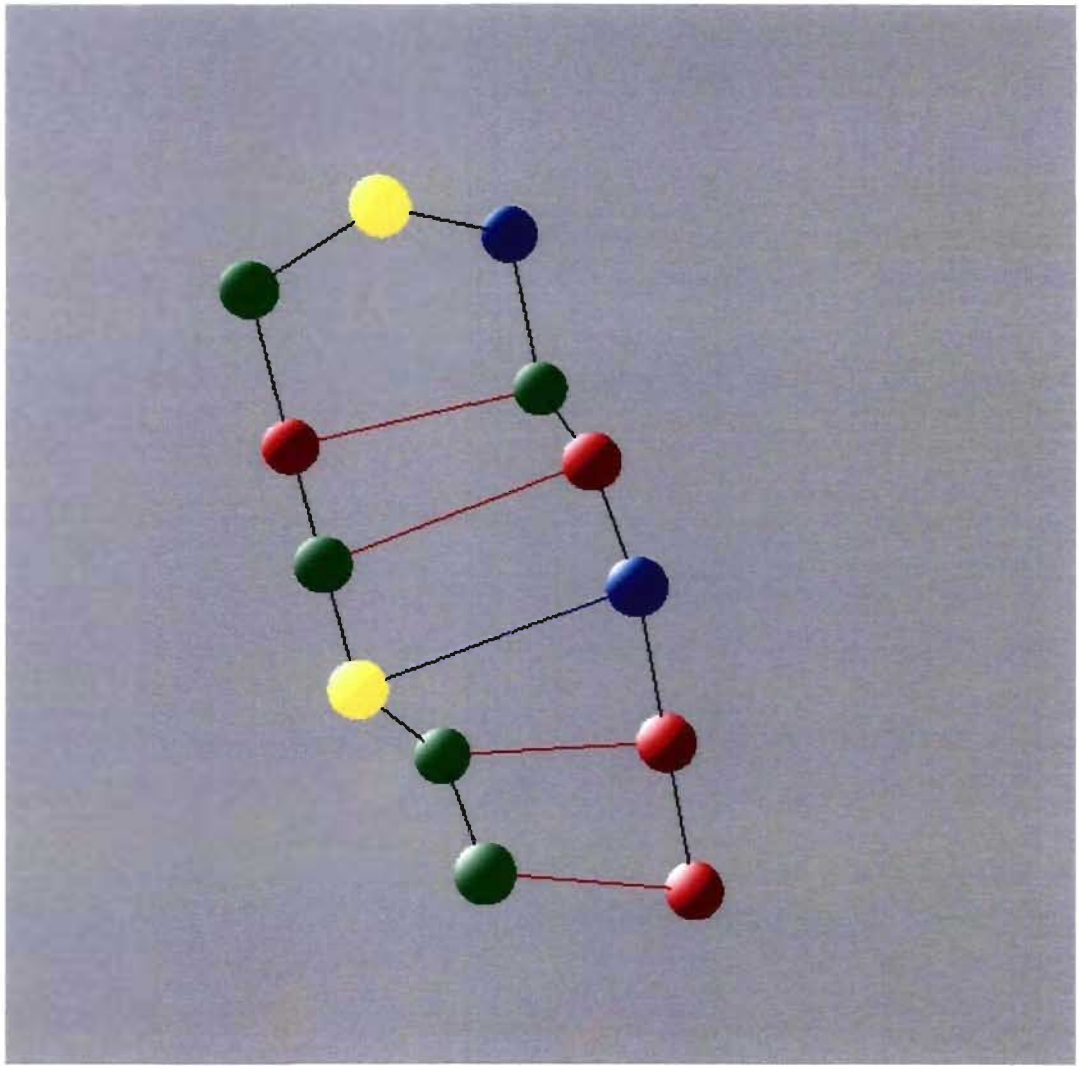

Figure A.3 Structure secondaire émergente avec 1I46
Structure de 13 nucléotides avec quatre liens GC et un lien AU.
(Code couleur : G = vert, C = rouge, A = bleu, U = jaune).

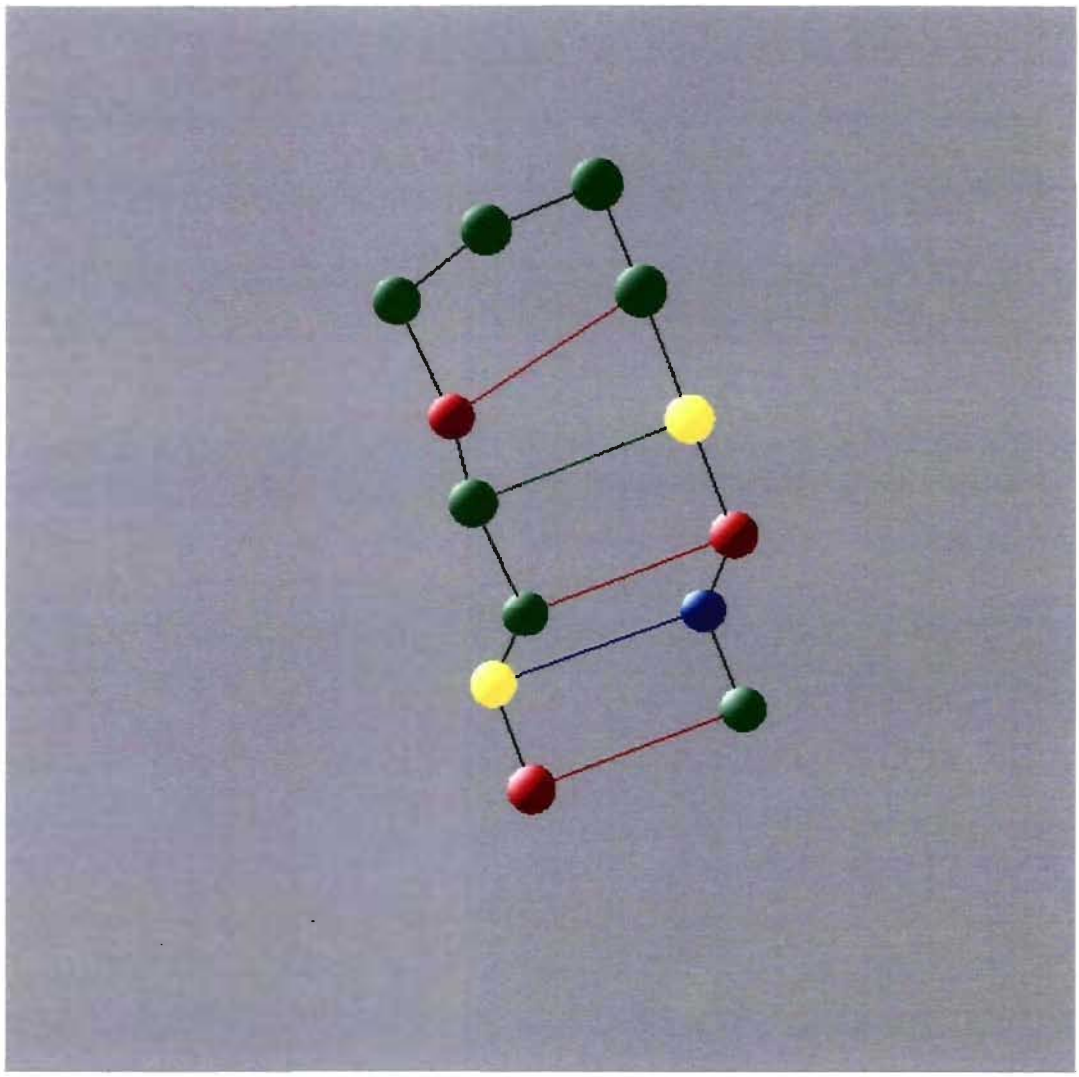

Figure A.4 Structure secondaire émergente avec 1VOP
Structure de 13 nucléotides avec trois liens GC, un lien AU et un lien GU.
(Code couleur : G = vert, C = rouge, A = bleu, U = jaune).

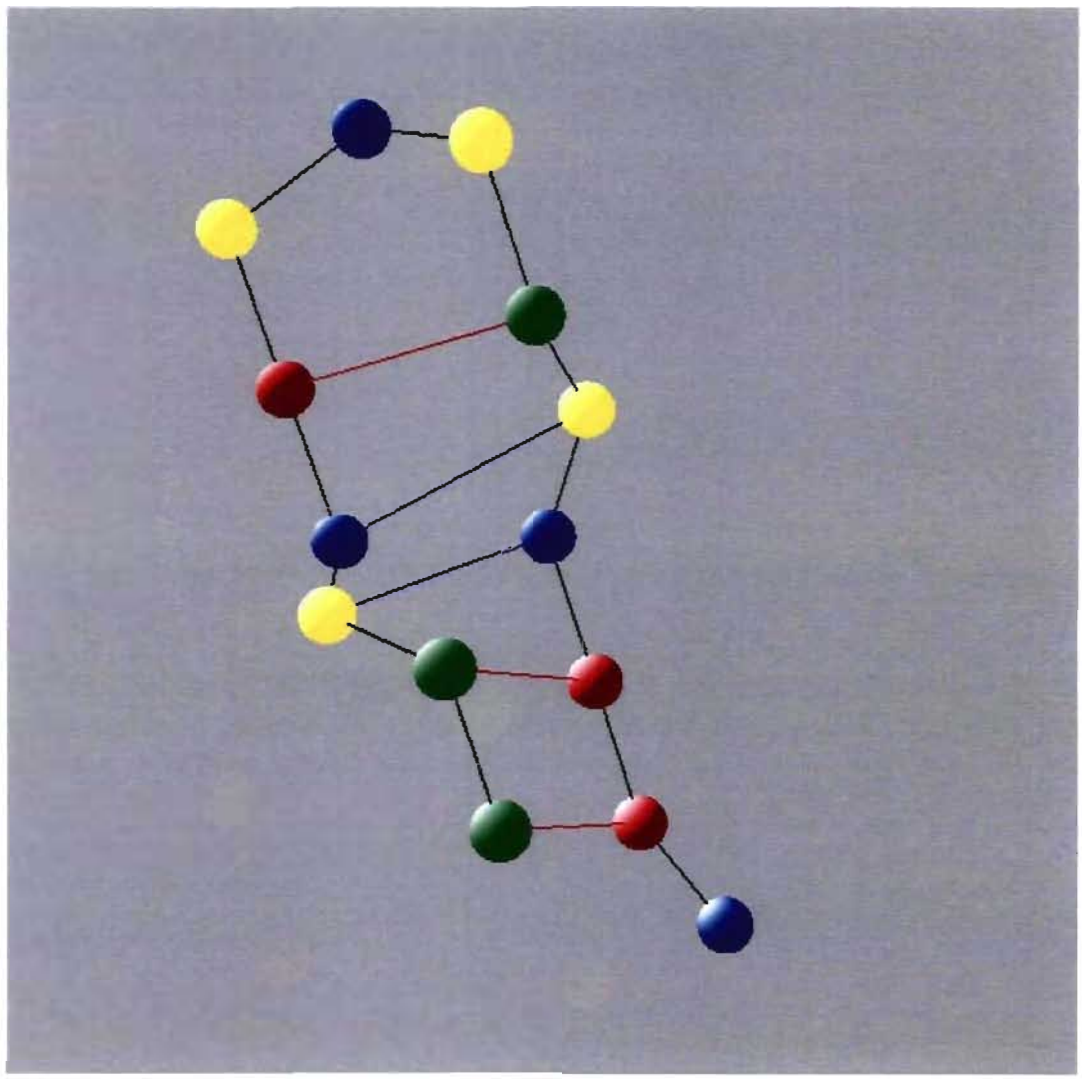

Figure A.5 Structure secondaire émergente avec 1IK1
Structure de 14 nucléotides avec trois liens GC et deux liens AU.
(Code couleur : G = vert, C = rouge, A = bleu, U = jaune).

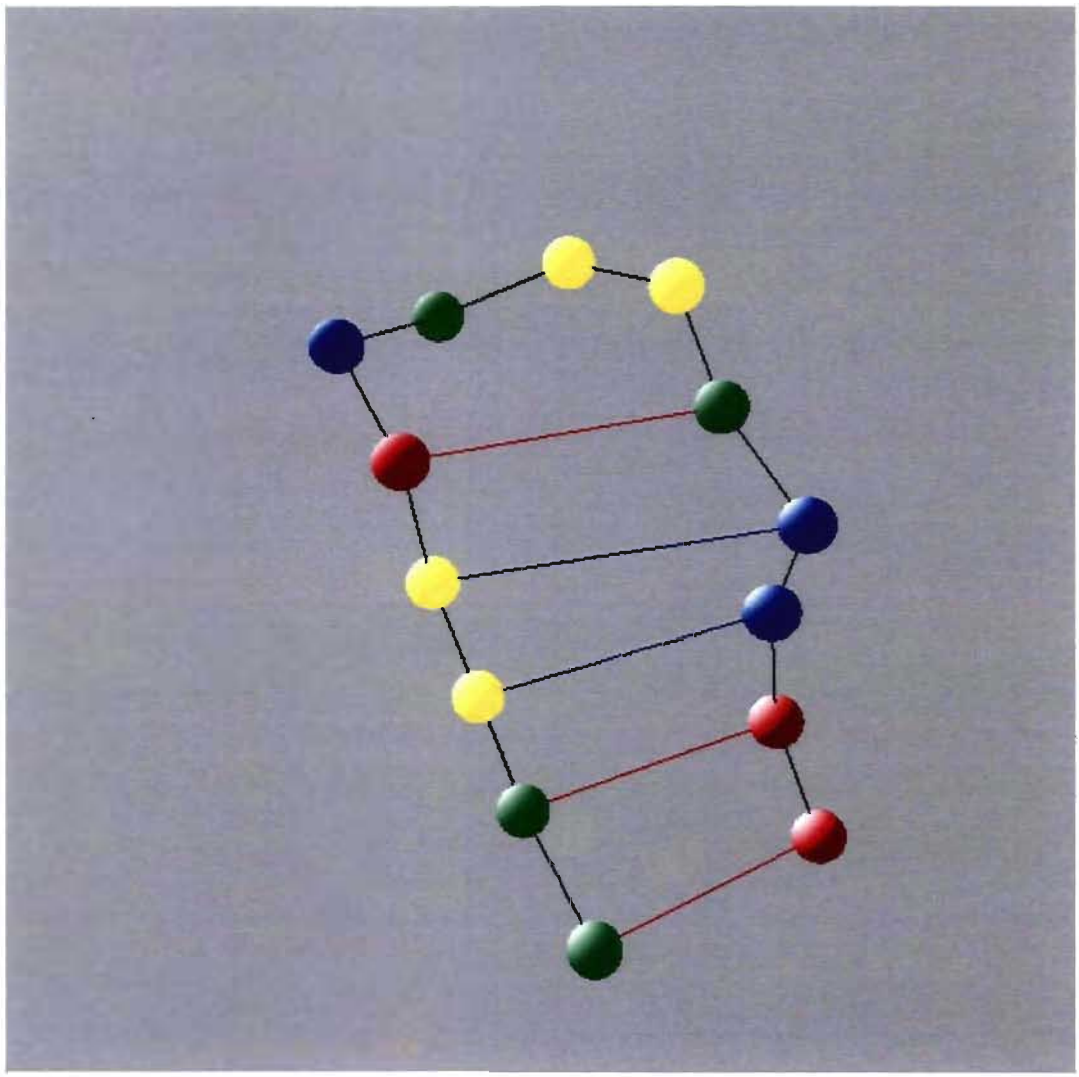

Figure A.6 Structure secondaire émergente avec 1K4B
Structure de 14 nucléotides avec trois liens GC et deux liens AU.
(Code couleur : G = vert, C = rouge, A = bleu, U = jaune).

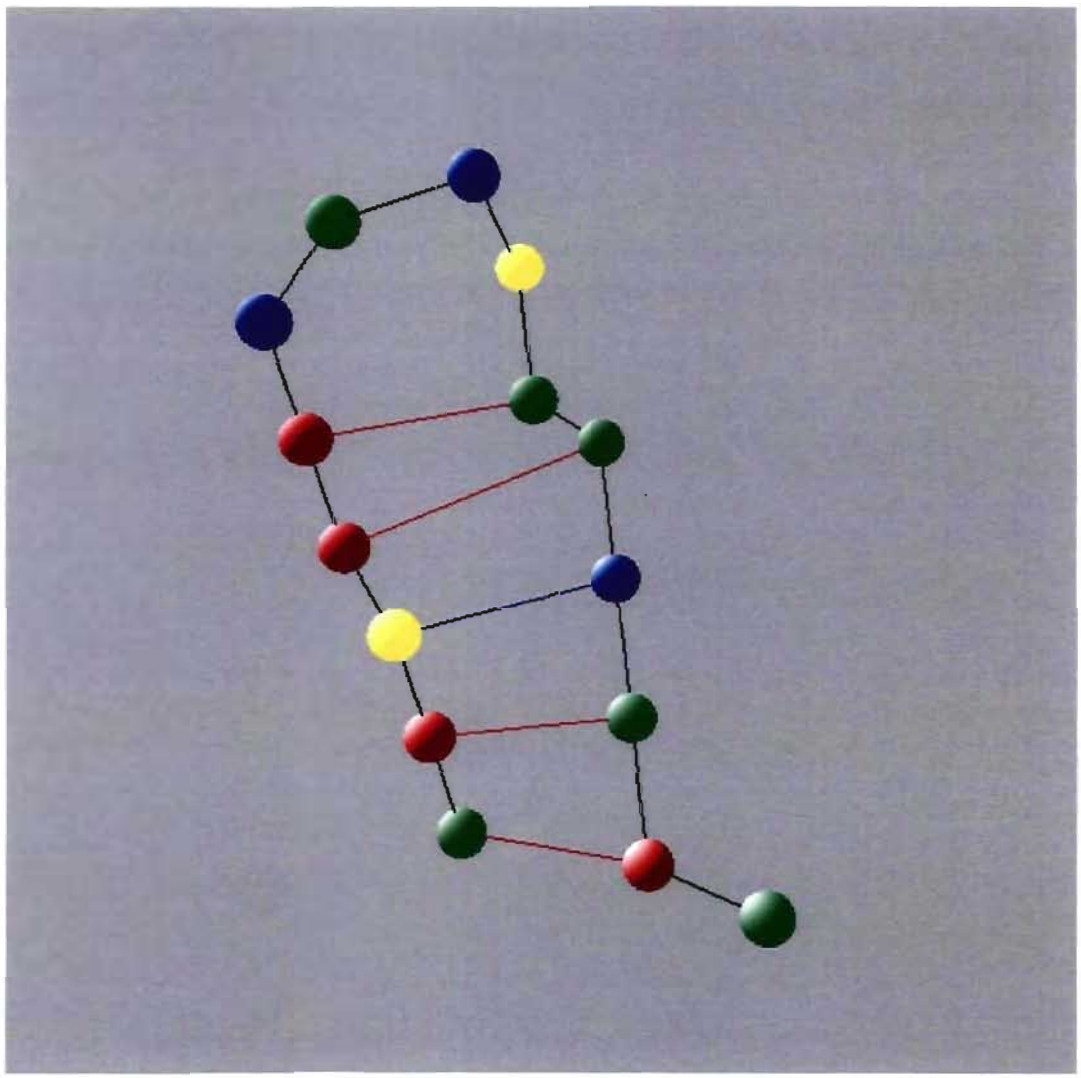

Figure A.7 Structure secondaire émergente avec 1ATW
Structure de 15 nucléotides avec quatre liens GC et un lien AU.
(Code couleur : G = vert, C = rouge, A = bleu, U = jaune).

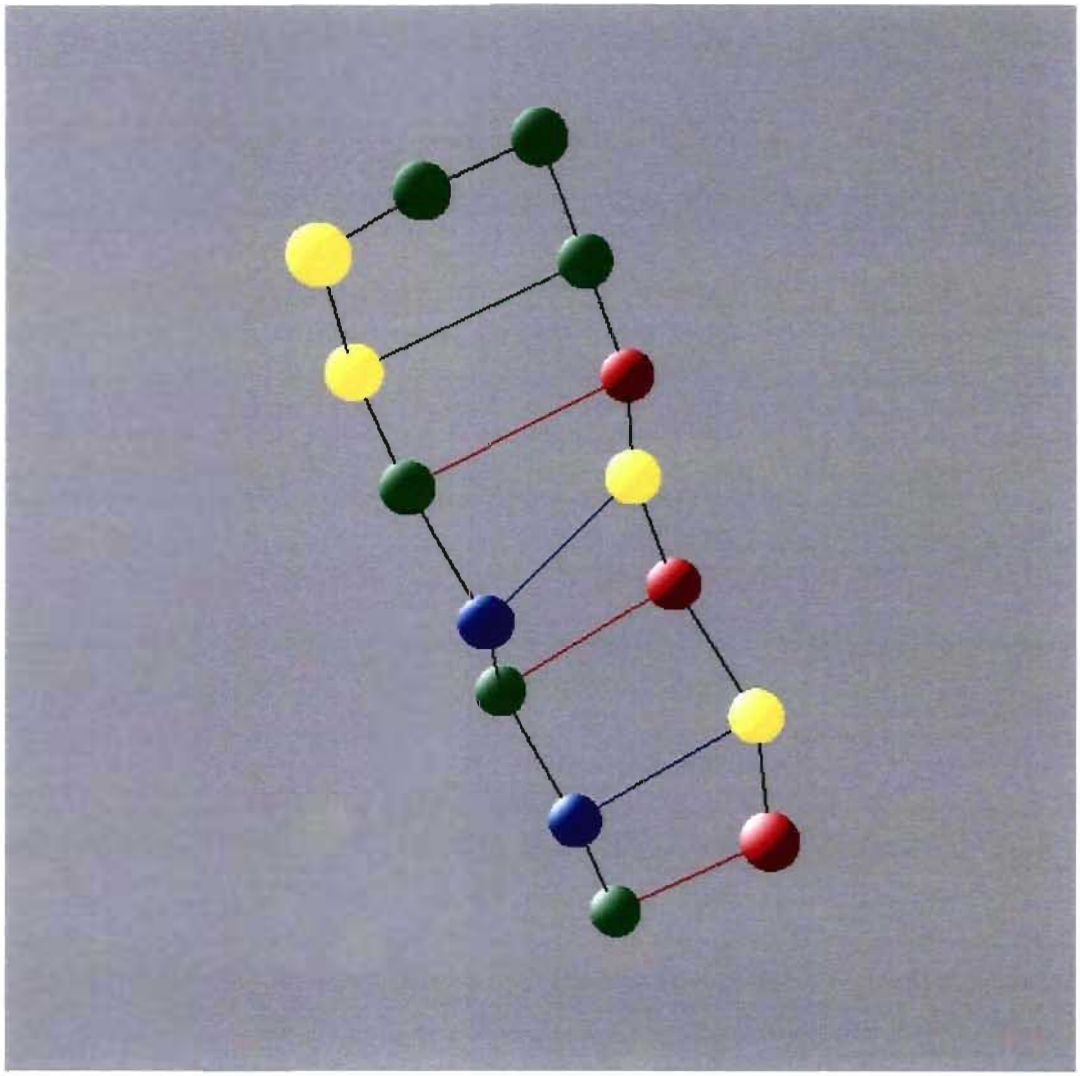

Figure A.8 Structure secondaire émergente avec 1OQ0
Structure de 15 nucléotides avec trois liens GC, deux liens AU et un lien GU.
(Code couleur : G = vert, C = rouge, A = bleu, U = jaune).

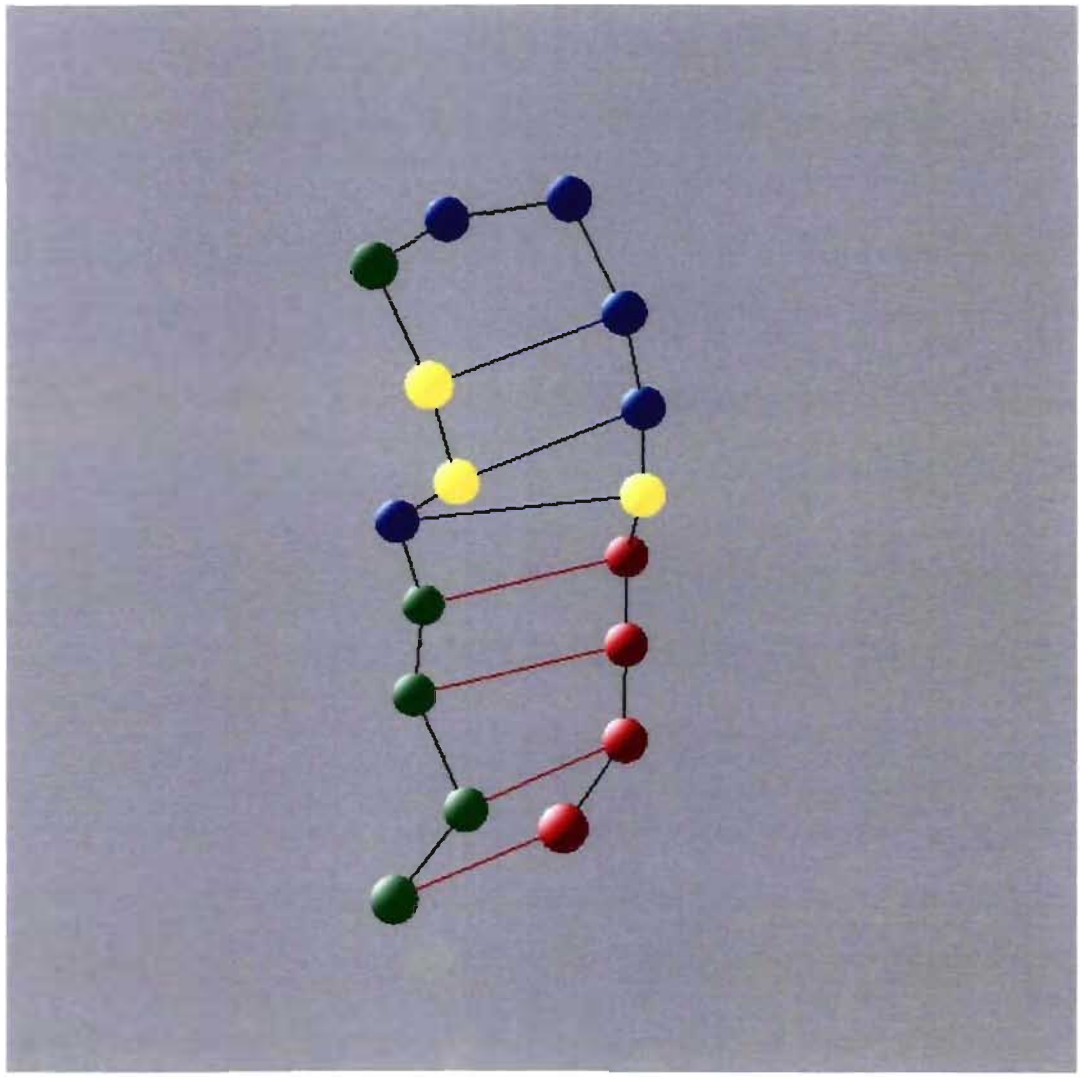

Figure A.9 Structure secondaire émergente avec 1J4Y
Structure de 17 nucléotides avec quatre liens GC et trois liens AU.
(Code couleur : G = vert, C = rouge, A = bleu, U = jaune).

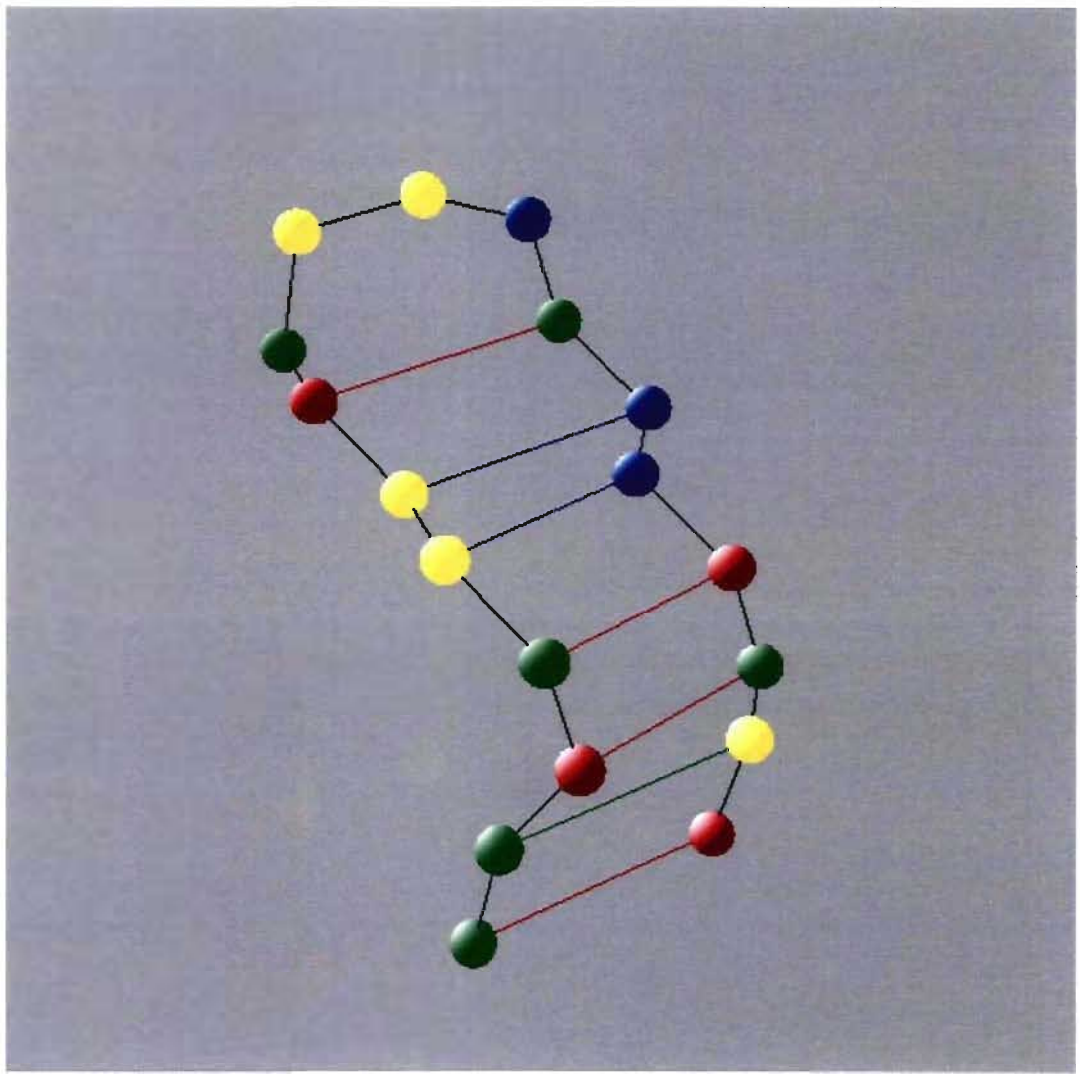

Figure A.10 Structure secondaire émergente avec 1Z30
Structure de 18 nucléotides avec quatre liens GC, deux liens AU et un lien GU.
(Code couleur : G = vert, C = rouge, A = bleu, U = jaune).

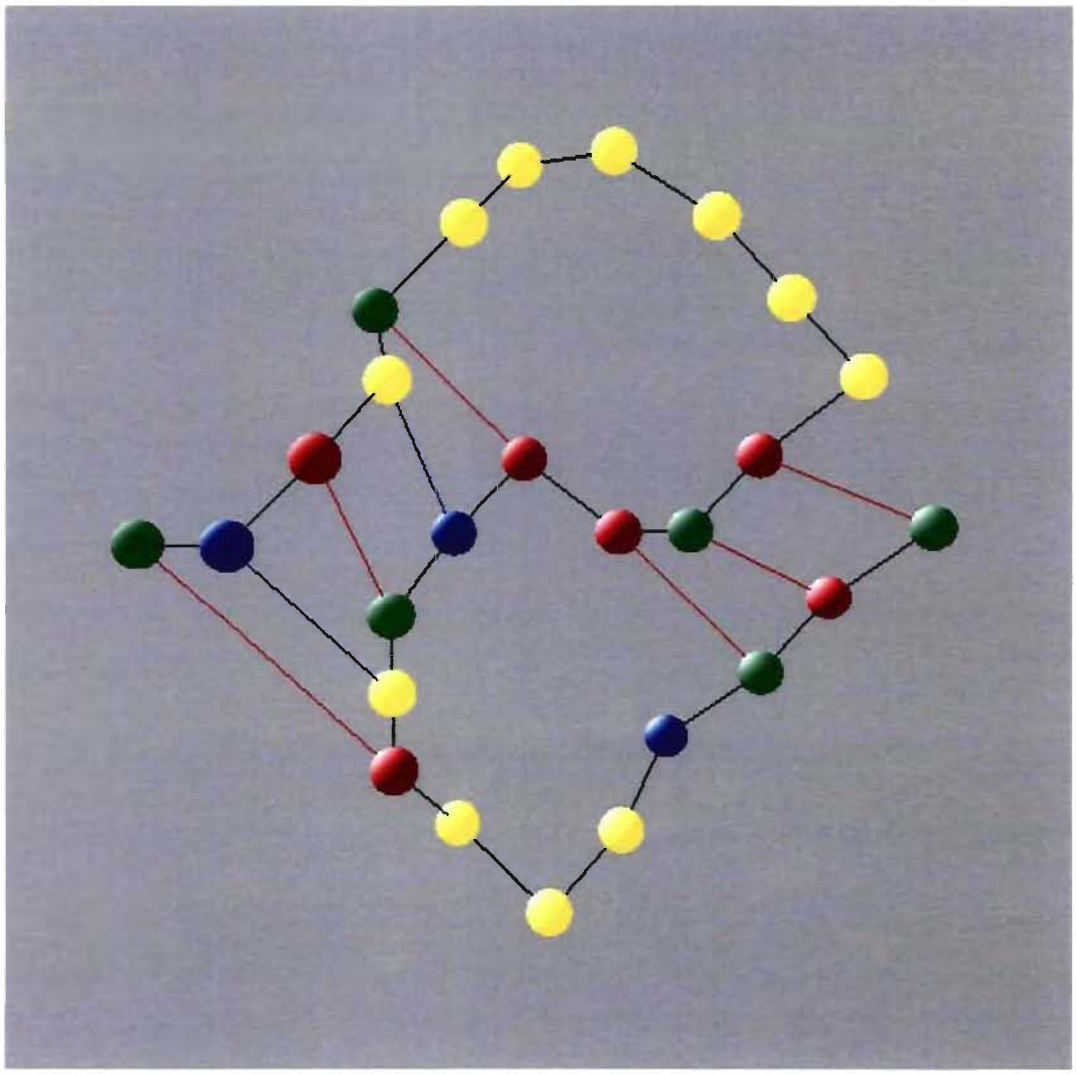

Figure A.11 Structure secondaire émergente avec PK5
Structure de 26 nucléotides avec des pseudonoeuds.
(Code couleur : G = vert, C = rouge, A = bleu, U = jaune).

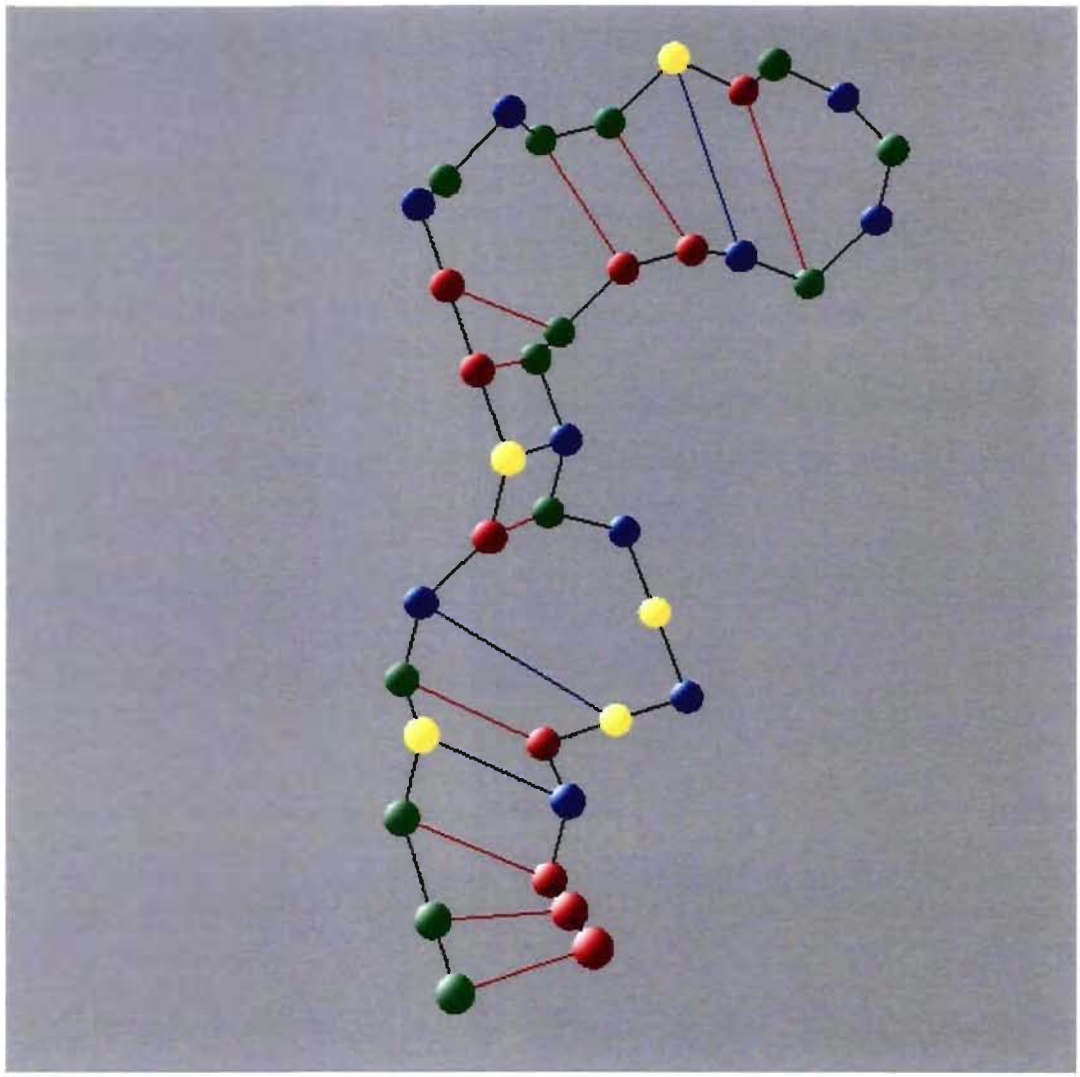

Figure A.12 Structure secondaire émergente avec 1A9L
Structure de 38 nucléotides avec dix liens GC et quatre liens AU.
(Code couleur : G = vert, C = rouge, A = bleu, U = jaune).

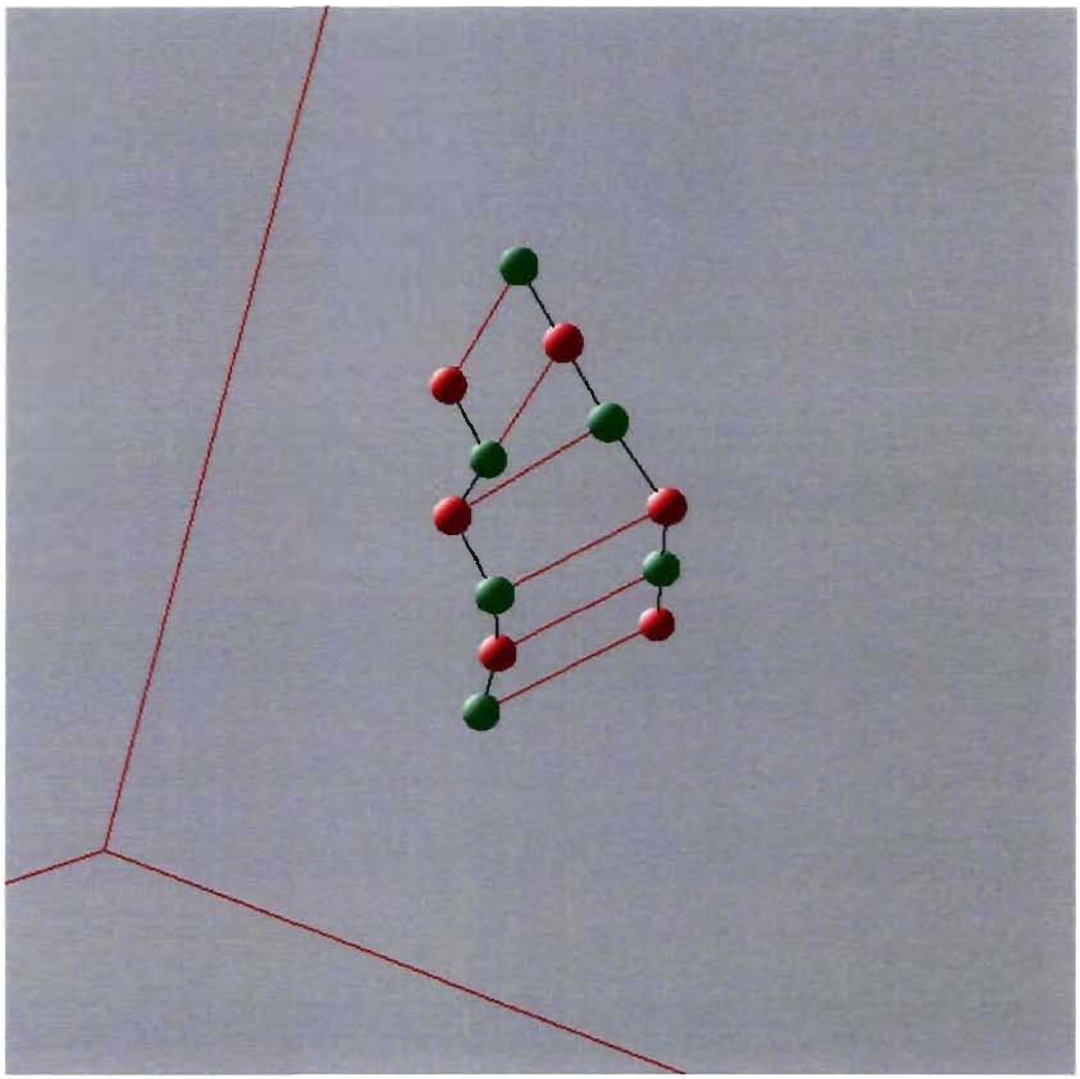

Figure A.13 Structure secondaire émergente avec 1PBM
Structure de 12 nucléotides formée à partir de 2 brins de 6 nucléotides.
(Code couleur : G = vert, C = rouge, A = bleu, U = jaune).

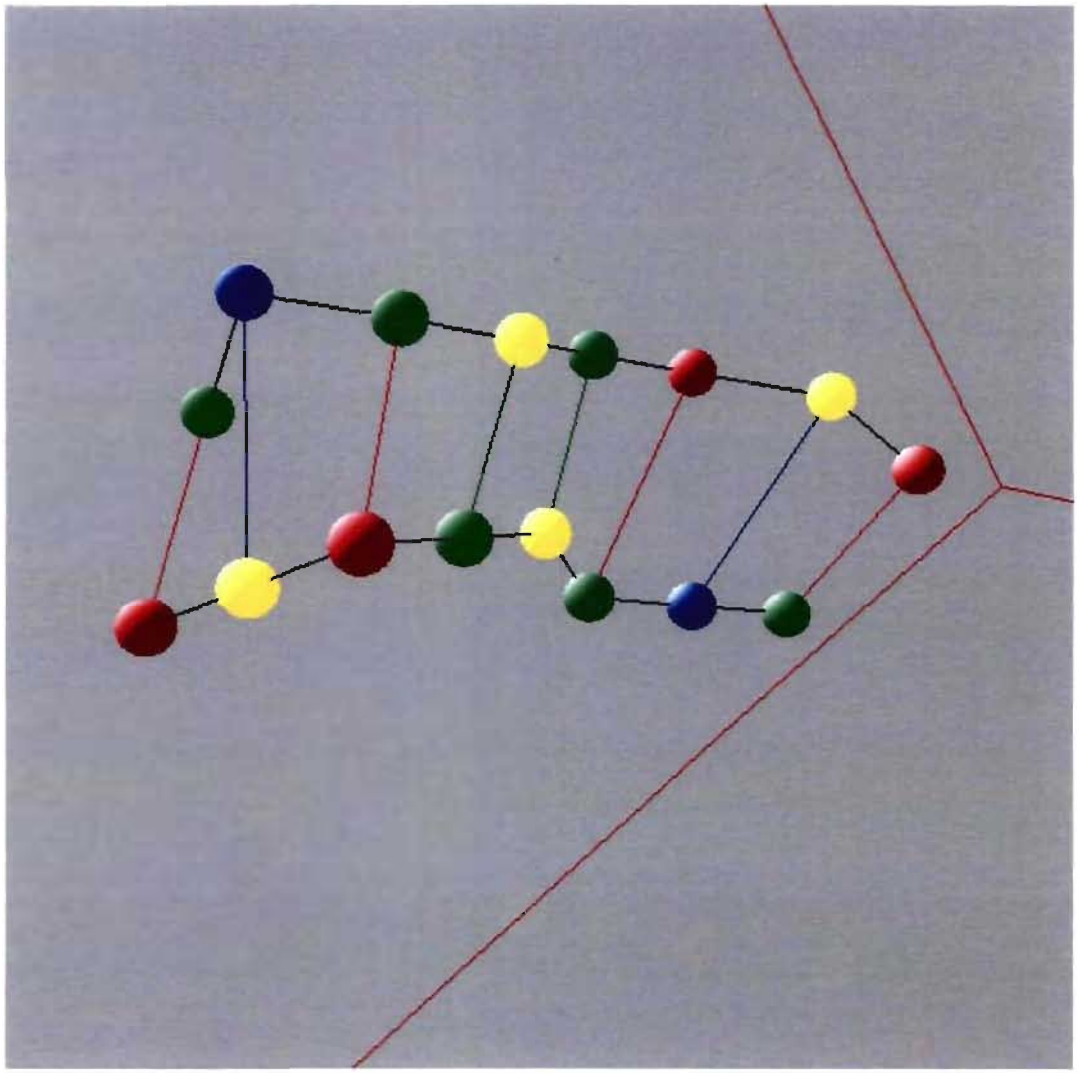

Figure A.14 Structure secondaire émergente avec 1EKA
Structure de 16 nucléotides formée à partir de 2 brins de 8 nucléotides.
(Code couleur : G = vert, C = rouge, A = bleu, U = jaune).

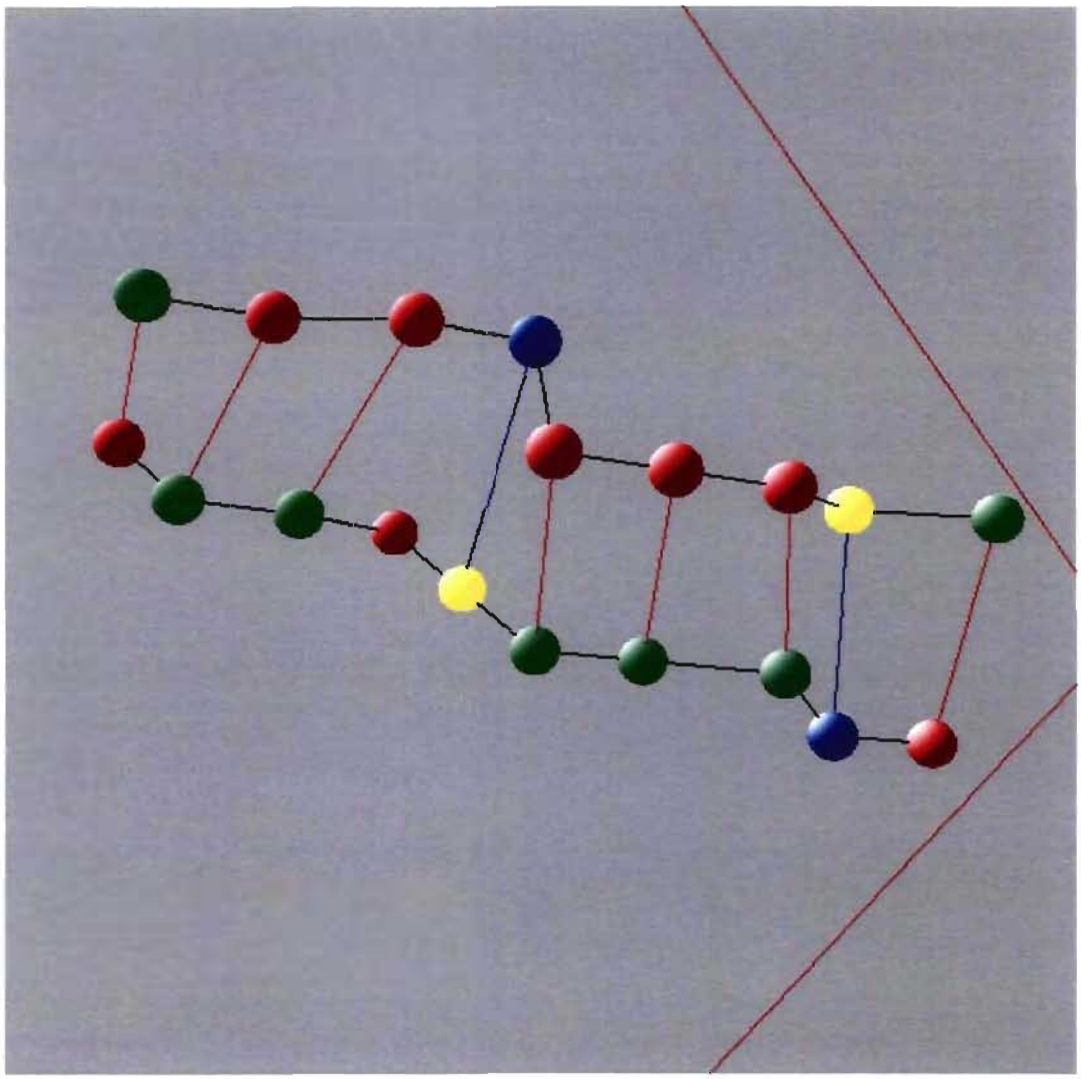

Figure A.15 Structure secondaire émergente avec 1DQF
Structure de 19 nucléotides formée à partir de 2 brins de 9 et 10 nucléotides.
(Code couleur : G = vert, C = rouge, A = bleu, U = jaune).

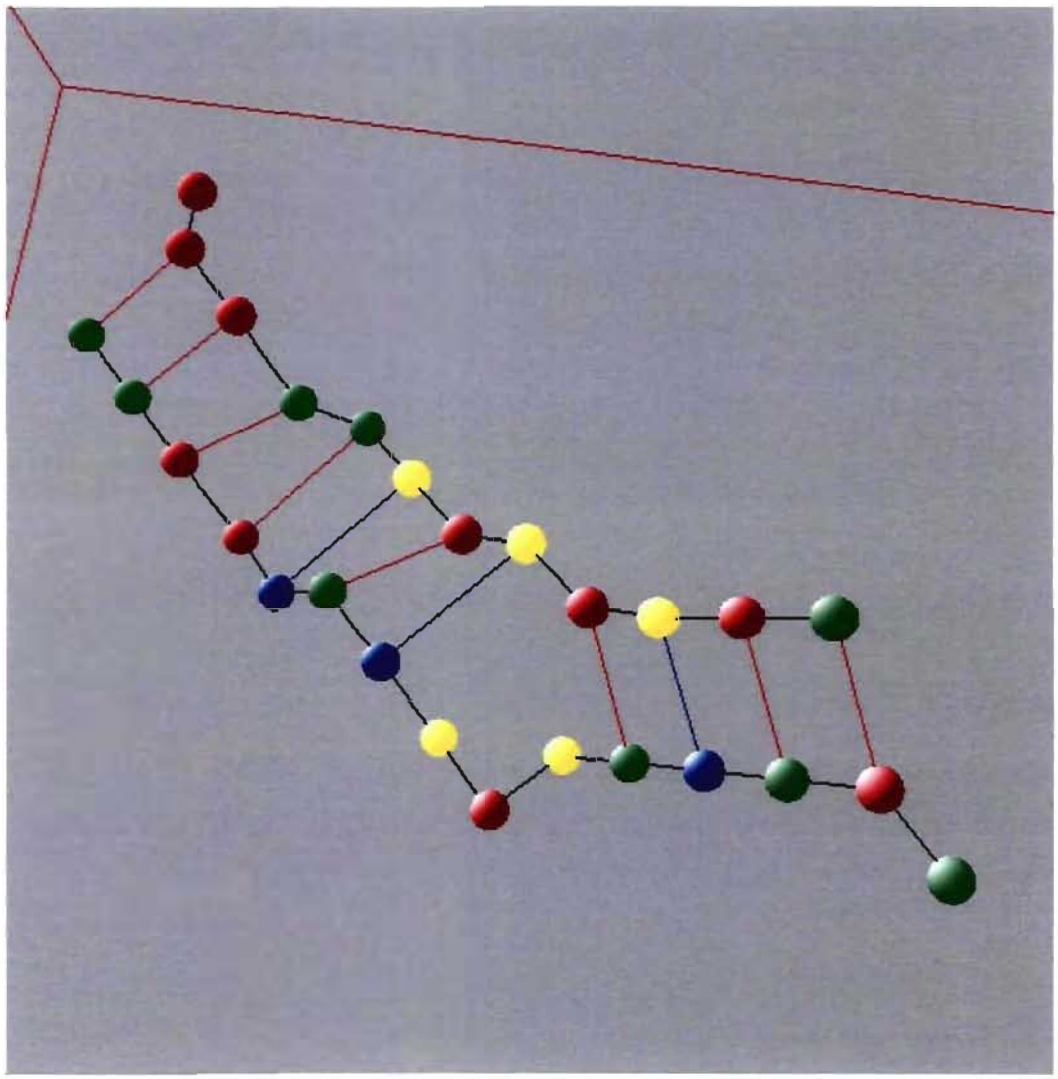

Figure A.16 Structure secondaire émergente avec 397D
Structure de 27 nucléotides formée à partir de 2 brins de 12 et 15 nucléotides.
(Code couleur : G = vert, C = rouge, A = bleu, U = jaune).

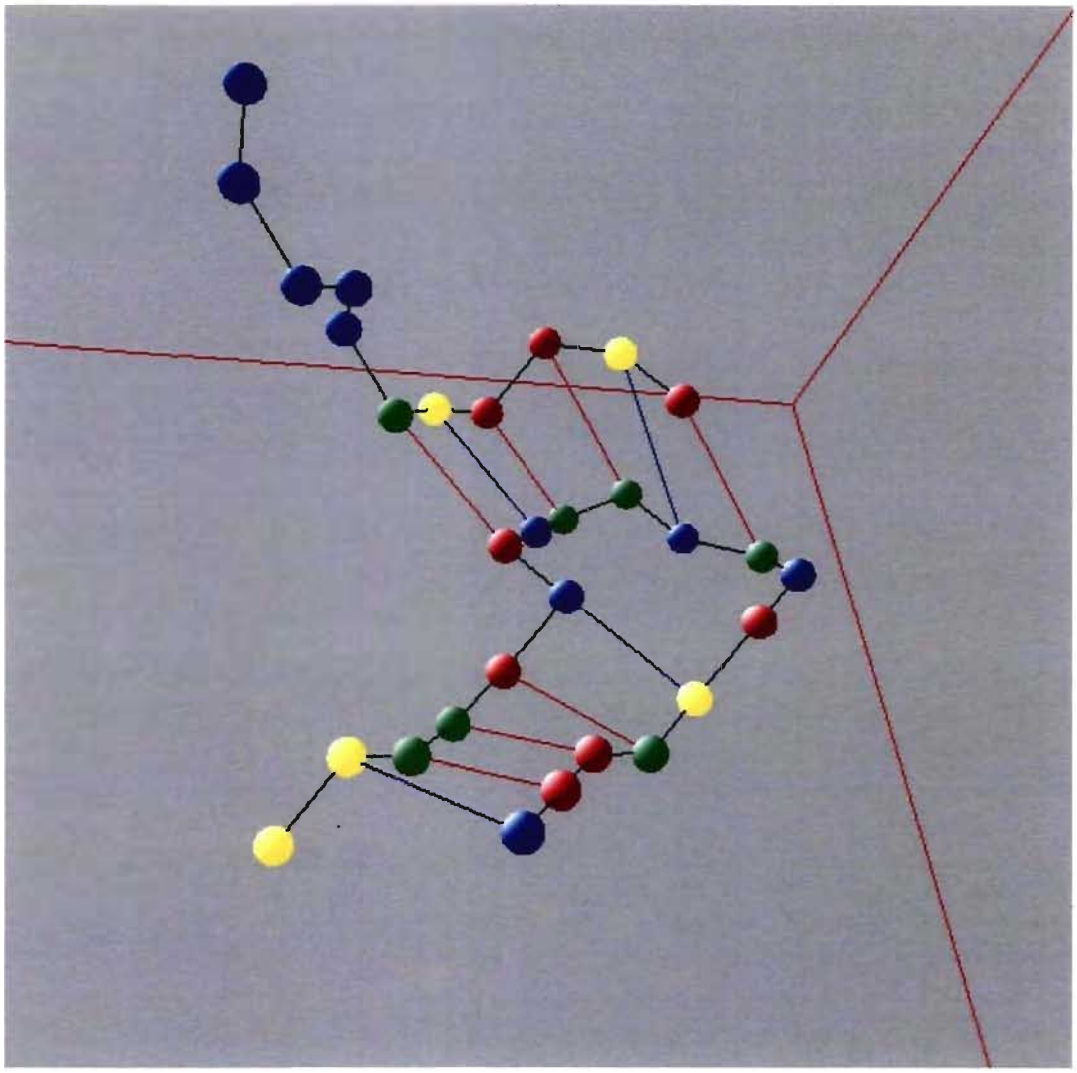

Figure A.17 Structure secondaire émergente avec 1F27
Structure de 30 nucléotides formée à partir de 2 brins de 11 et 19 nucléotides. La structure présente des pseudonoeuds.
(Code couleur : G = vert, C = rouge, A = bleu, U = jaune).

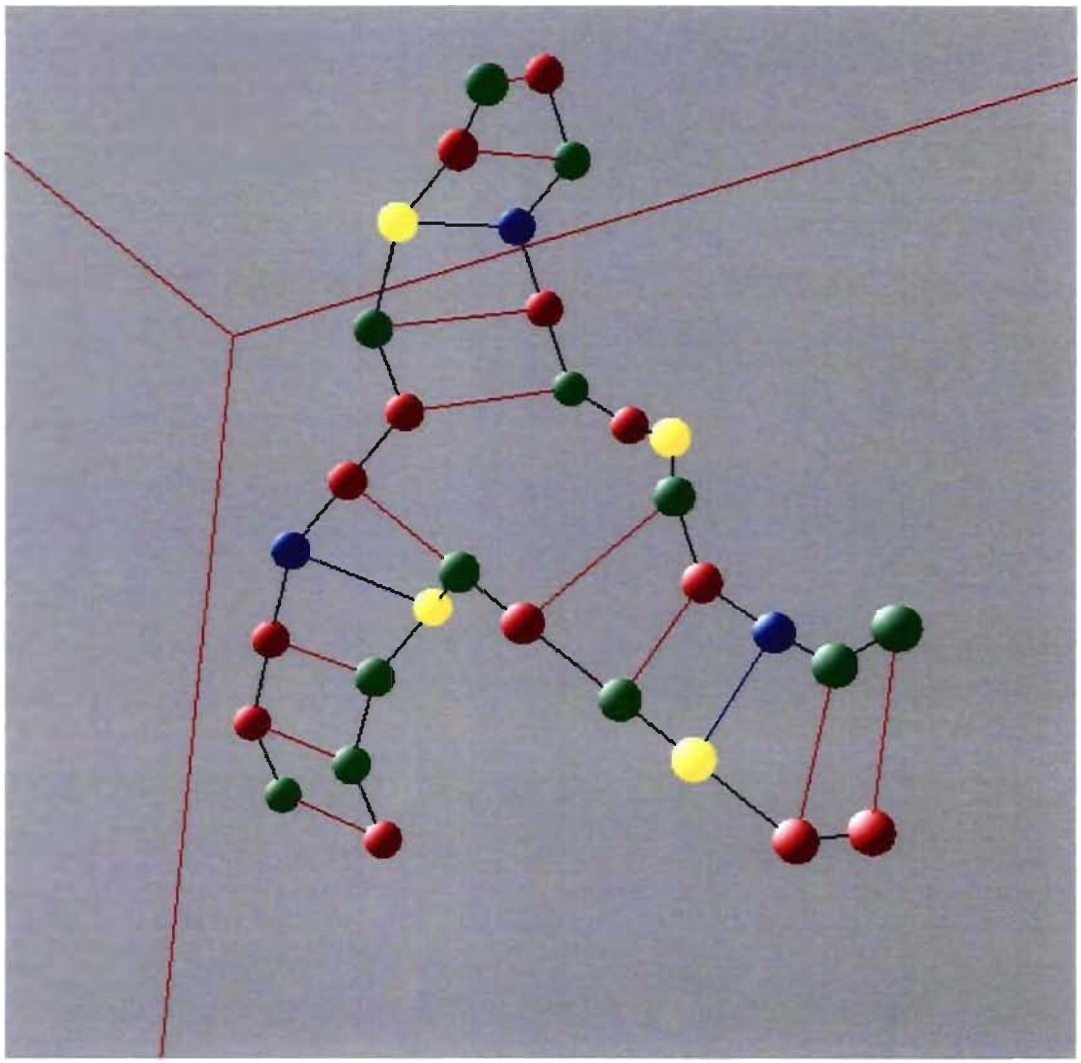

Figure A.18 Structure secondaire émergente avec 1EKW
Structure d'ADN de 32 nucléotides formée à partir de 3 brins de 10, 12 et 10 nucléotides. La structure présente une forme générale en Y.
(Code couleur : G = vert, C = rouge, A = bleu, T = jaune).

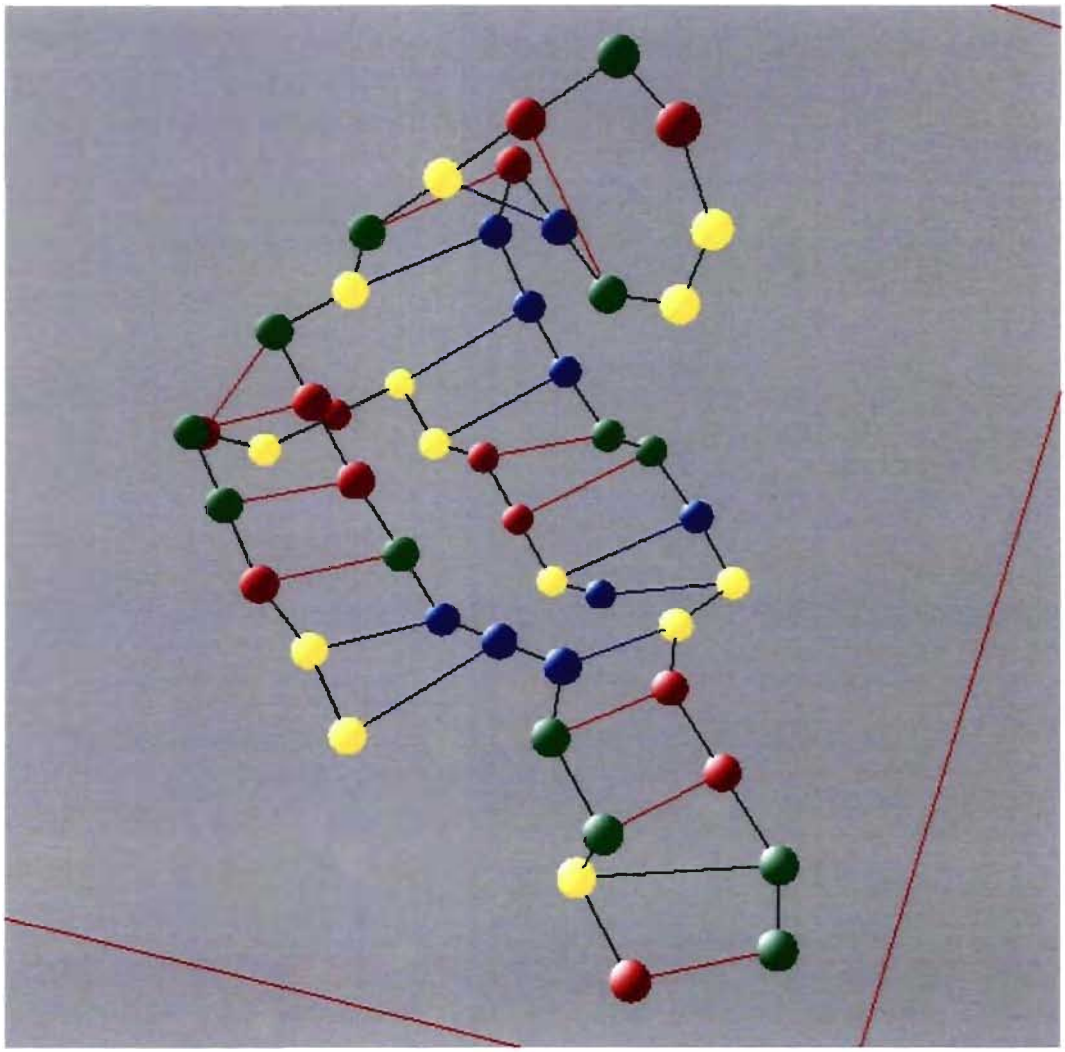

Figure A.19 Structure secondaire émergente avec 2P89
Structure de 48 nucléotides formée à partir de 2 brins de 34 et 14 nucléotides. La structure présente une forme générale complexe avec des pseudonoeuds.
(Code couleur : G = vert, C = rouge, A = bleu, T = jaune).

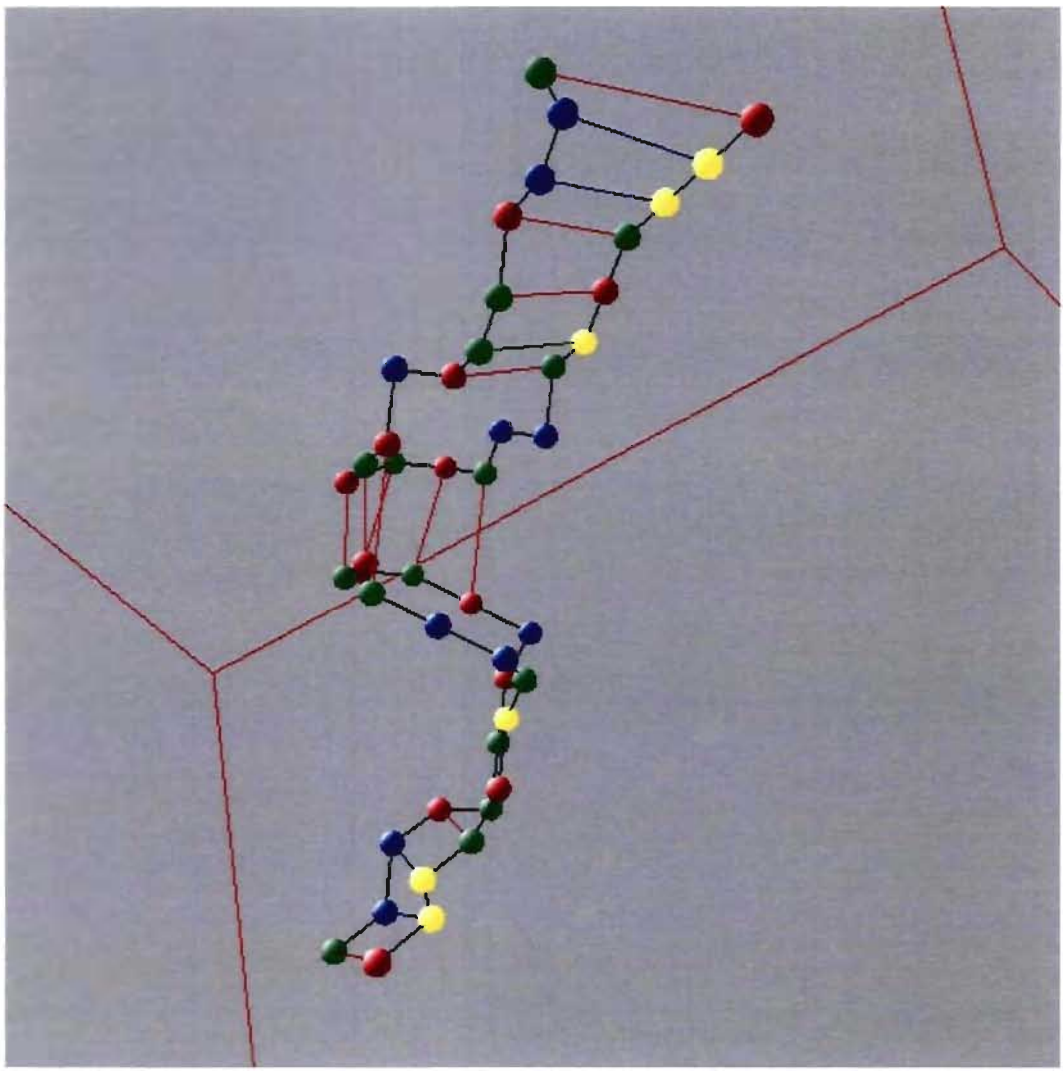

Figure A.20 Première structure secondaire émergente avec 2B8R
Structure de 46 nucléotides formée à partir de 2 brins de 23 nucléotides. La structure présente une forme générale de deux têtes d'épingle attachées par les têtes. Cette structure est la $S_{hybride}$ et elle contient des pseudonoeuds.
(Code couleur : G = vert, C = rouge, A = bleu, T = jaune).

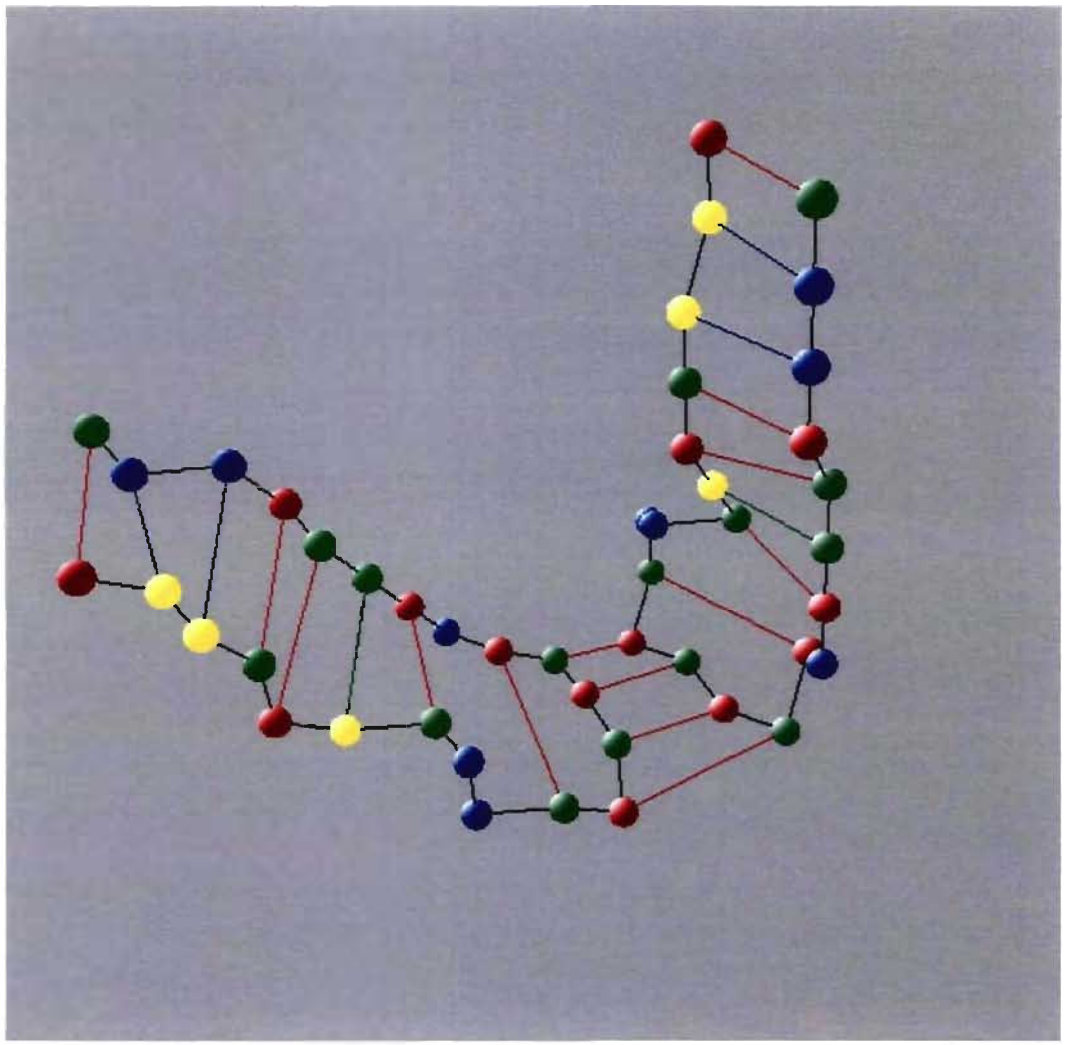

Figure A.21 Deuxième structure secondaire émergente avec 2B8R
Structure de 46 nucléotides formée à partir de 2 brins de 23 nucléotides. La structure présente une forme générale de deux têtes d'épingle attachées en partie par les têtes. Cette structure est dans le voisinage de la S_{hybride} et elle contient des pseudonoeuds.
(Code couleur : G = vert, C = rouge, A = bleu, T = jaune).

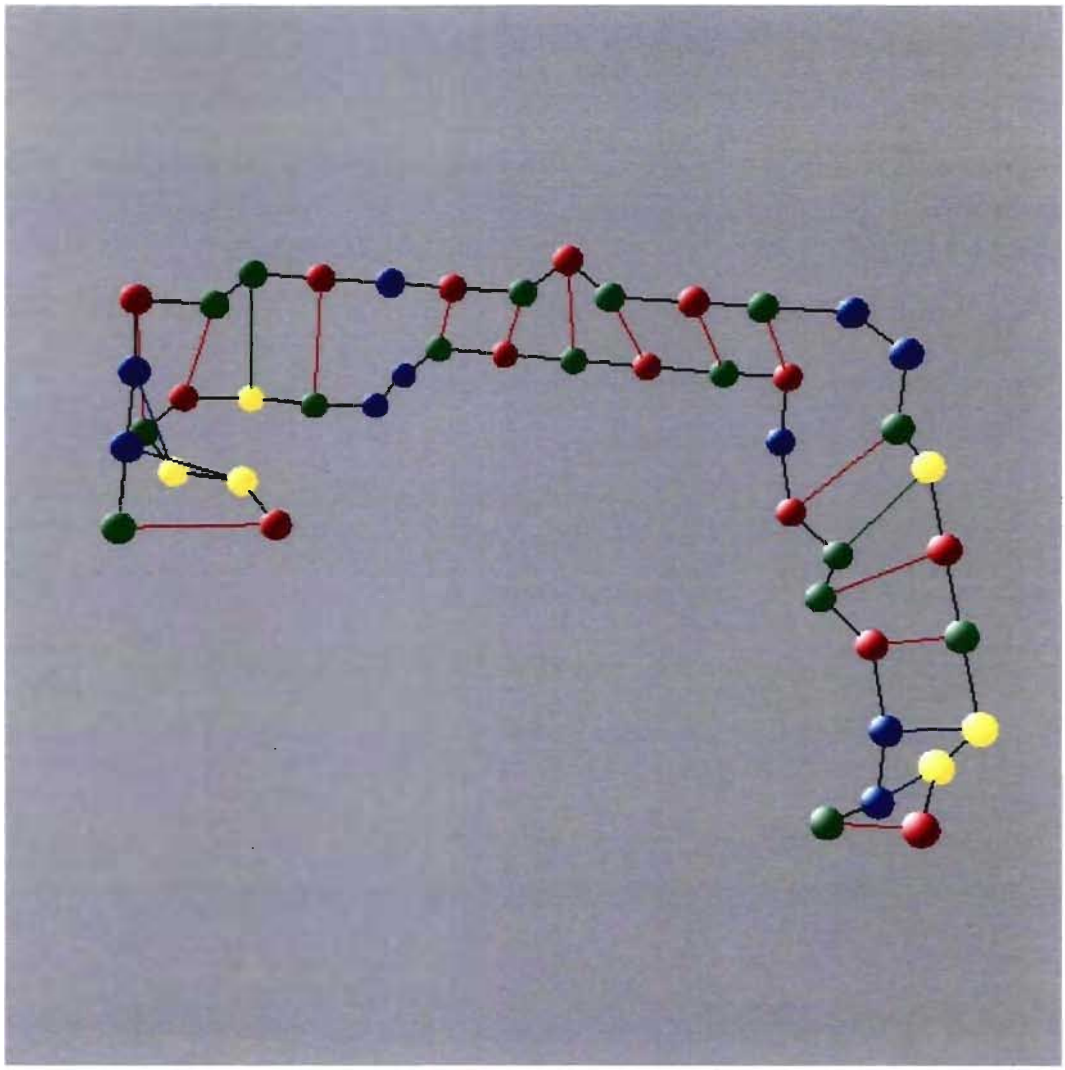

Figure A.22 Troisième structure secondaire émergente avec 2B8R
Structure de 46 nucléotides formée à partir de 2 brins de 23 nucléotides. La structure présente la forme générale d'une échelle. Cette structure est totalement différente de la $S_{hybride}$. Par contre, elle correspond à la structure de minimum d'énergie calculée par le programme Mfold [32].
(Code couleur : G = vert, C = rouge, A = bleu, T = jaune).