N° d'ordre : 4357

Classification d'ARN codants et d'ARN non-codants

THÈSE

présentée et soutenue publiquement le 31 mars 2009

pour l'obtention du

Doctorat de l'Université des Sciences et Technologies de Lille

(spécialité informatique)

par

Arnaud Fontaine

Composition du jury

Rapporteurs :	Thomas Schiex, D.R. INRA	INRA – Unité de Toulouse
	Claude Thermes, D.R. CNRS	Centre de Génétique Moléculaire – Gif-sur-Yvette
Examinateurs :	Fabrice Leclerc, C.R. CNRS	MAEM, Université Henri Poincaré – Nancy 1
	Nouredine Melab, Professeur	LIFL, Université des Sciences et Technologies de Lille
	Fariza Tahi, Maître de conférences	IBISC, Université d'Evry-Val d'Essonne
Directeur :	Hélène Touzet, C.R. CNRS	LIFL, Université des Sciences et Technologies de Lille

UNIVERSITÉ DES SCIENCES ET TECHNOLOGIES DE LILLE – LILLE 1
ÉCOLE DOCTORALE SCIENCES POUR L'INGÉNIEUR
Laboratoire d'Informatique Fondamentale de Lille — UMR 8022
U.F.R. d'I.E.E.A. – Bât. M3 – 59655 VILLENEUVE D'ASCQ CEDEX
Tél. : +33 (0)3 28 77 85 41 – Télécopie : +33 (0)3 28 77 85 37 – email : direction@lifl.fr

Table des matières

Introduction

Tous les organismes vivants, des plus simples aux plus complexes, sont composés de cellules qui présentent des caractéristiques communes en terme de structure mais également de fonctionnement. Trois types de macromolécules fondamentales sont impliquées dans cette unité cellulaire et moléculaire du vivant : les ADN, les ARN et les protéines. Schématiquement, la *séquence* des éléments qui composent ces molécules constitue la représentation minimale permettant de décrire l'information qu'elles contiennent.

La séquence des ADN est responsable du stockage de l'information génétique qui détermine le patrimoine génétique d'un organisme. L'information génétique est segmentée en gènes dont l'expression est à l'origine de la synthèse d'ARN puis de protéines. La séquence d'une protéine ne constitue qu'un premier niveau dans sa description. Les protéines se replient en effet dans l'espace pour former une structure tridimensionnelle qui détermine leur fonction. Les ARN sont quant à eux des acteurs plus polyvalents. Les ARN codants contiennent l'information nécessaire à la synthèse d'une protéine, tandis que les ARN non-codants se comportent sommairement comme des protéines en se repliant sur eux-mêmes pour adopter une conformation spatiale qui détermine leur fonction.

Alors que l'on dispose de plus en plus de séquences d'ADN et d'ARN provenant de la génomique et de la transcriptomique, la signification de la plupart de ces séquences reste encore à élucider. Les premiers travaux d'annotation automatique de séquences remontent au début des années 80, suite au premier séquençage complet d'un génome. A l'heure actuelle, près de 1 000 génomes d'organismes différents sont complètement séquencés et disponibles publiquement, et plus de 4 000 projets de séquençage sont en cours [1]. De prime abord, l'analyse systématique de ces séquences par des techniques expérimentales en vue de leur annotation n'est plus envisagée à cause de leurs coûts humain et financier trop importants. L'analyse automatique de séquences par des moyens informatiques est donc plus que jamais un challenge majeur.

L'aboutissement de nombreux projets de séquençage ces dernières années a toutefois quelque peu changé la donne en contribuant à l'émergence de la génomique comparative. En effet, bien que portée par les mêmes supports et exprimée selon des mécanismes communs, l'information génétique est également la source de la diversité des organismes vivants. L'unicité cellulaire du vivant suggère ainsi l'évolution à partir d'ancêtres communs plus ou moins éloignés durant laquelle les séquences génomiques se transforment, tout en préservant certaines fonctions.

La génomique comparative consiste à étudier et analyser les ressemblances et les différences apparues durant l'évolution entre des séquences génomiques. Ces séquences ne sont ainsi plus considérées de manière individuelle mais reliées entre elles par l'évolution. Dans ce contexte,

[1] `http://www.genomesonline.org`

la ressemblance significative entre des séquences d'ADN ou d'ARN constitue un indicateur sur une éventuelle fonction commune. Toutefois, la proposition réciproque est fausse. Une absence de similarité significative entre plusieurs séquences d'ADN ou d'ARN n'implique pas nécessairement l'absence d'une fonction partagée. Des séquences différentes d'ARN codants peuvent en effet être à l'origine de protéines homologues, et des séquences différentes d'ARN non-codants peuvent former des structures communes. En travaillant sur des ensembles de séquences plutôt que sur des séquences isolées, les approches modernes de génomique comparative qui prennent part à l'annotation de séquences fonctionnelles s'avèrent particulièrement fécondes.

Les travaux que nous présentons dans ce manuscrit s'inscrivent dans ce contexte. D'un côté, Protea, dédié à la prédiction de séquences codantes homologues, et de l'autre, caRNAc, dédié à la prédiction de structures secondaires conservées. Il existe plusieurs méthodes fructueuses pour traiter ces deux problématiques par analyse comparative. Ces méthodes s'appuient quasi systématiquement sur un alignement préalable des séquences supposé fiable. Il est cependant difficile de produire un alignement de qualité sur des séquences faiblement conservées. Les méthodes actuelles ne sont donc pas bien adaptées pour traiter ce type de séquences. Nos méthodes, Protea et caRNAc, ont été conçues pour compléter l'arsenal des méthodes existantes en palliant à ce problème avec un traitement adapté aux séquences faiblement conservées. Toutes deux acceptent en entrée une famille de quelques séquences non alignées, de moins d'une dizaine jusqu'à plusieurs dizaines, dont la longueur doit être globalement homogène, mais peut atteindre jusqu'à plusieurs milliers de bases.

Plan de lecture

Ce document est organisé en quatre chapitres.

Le premier chapitre est consacré à l'introduction des notions biologiques nécessaires à la bonne compréhension des méthodes présentées dans ce document. Ce chapitre débute par la présentation des mécanismes en lien direct avec le stockage et l'expression de l'information génétique communs à tous les organismes vivants. Ensuite, nous détaillons la *transcription*, première étape de l'expression de l'information génétique, qui permet de synthétiser les ARN, et la *maturation* des transcrits. Puis, nous nous intéressons plus en détails aux deux types d'ARN existants : les ARN codants et les ARN non-codants. Enfin, la dernière partie de ce chapitre est consacrée à la mise en place de notre fil conducteur : l'analyse comparative de séquences avec notamment la définition de *méta-séquence* qui revient de manière récurrente dans les chapitres suivants.

Le second chapitre porte sur la détection de séquences codantes. Tout d'abord, nous présentons trois types de méthodes existantes : l'approche *ab initio*, l'approche par similarité de séquence et enfin l'approche comparative. La littérature foisonne de méthodes dédiées à ce problème. Nous n'en présentons qu'une sélection éclairée représentative de la diversité des méthodes existantes. Ensuite, nous présentons Protea, notre contribution au problème. Partant d'observations concrètes sur deux séquences, nous formalisons dans un premier temps la détection d'une séquence d'acides aminés conservée sur deux séquences par la comparaison de leurs traductions potentielles. Puis, ce principe est étendu à des ensembles de taille quelconque de séquences. Enfin, la dernière partie de ce chapitre est consacrée aux résultats expérimentaux de Protea. Les performances de Protea sont évaluées sur un large éventail de séquences, puis nous passons à un cas pratique en appliquant Protea à l'annotation de

nouvelles séquences codantes sur le génome humain.

Le troisième chapitre porte sur la prédiction de structures d'ARN et la prédiction de gènes à ARN. Nous commençons par poser le problème de la prédiction de structures en introduisant les deux grandes familles d'approches : l'approche thermodynamique et l'approche comparative. Nous refermons ce rapide état de l'art sur la prédiction de structures par l'évaluation de référence des méthodes existantes, BRAliBase I proposée par Gardner en 2004 [GG04]. Ensuite, nous nous intéressons à un problème étroitement lié à la prédiction de structures d'ARN, la prédiction de gènes à ARN. Pour présenter ce problème, nous dressons un parallèle avec les approches mises en œuvre pour la prédiction de gènes codants en partant des biais de composition en séquence pour terminer par l'approche comparative. La troisième partie qui compose ce chapitre décrit notre contribution au problème de la prédiction de structures. Nous y présentons ainsi les évolutions apportées au logiciel caRNAc, dédié à la prédiction de structures secondaires conservées, notamment l'introduction des méta-séquences. Enfin, les résultats de caRNAc par rapport à ceux des méthodes existantes sur le jeu de données de référence, BRAliBase I, sont exposés dans la dernière partie de ce chapitre. Avant de refermer ce chapitre, nous décrivons les travaux que nous avons menés dans l'optique de définir une méthode de prédiction de gènes à ARN basée sur l'existence de structure conservée significative prédite par caRNAc.

Le quatrième et dernier chapitre de ce document est dédié à la présentation de deux travaux collaboratifs menés au sein de l'équipe qui intègrent Protea et caRNAc. La première partie de ce chapitre est consacrée à l'alignement multiple de séquences codantes homologues et de séquences qui partagent une structure commune, avec le logiciel Magnolia. Magnolia est un logiciel d'alignement multiple qui résulte de la combinaison de Protea, caRNAc et Gardenia, un logiciel d'alignement de structures d'ARN développé dans l'équipe. La seconde partie de ce chapitre est dédiée à l'application de Protea et caRNAc pour l'annotation de séquences génomiques. Nous présentons le pipeline logiciel développé dans l'équipe avant de fournir quelques résultats expérimentaux.

Chapitre 1

Les Acides RiboNucléiques

Les travaux décrits dans ce manuscrit portent sur l'analyse des acides ribonucléiques (ARN) codants et non-codants par des approches de bio-informatique. Dans ce premier chapitre, nous présentons le contexte biologique général et les mécanismes moléculaires sur lesquels s'appuient nos travaux.

Nous commençons par situer les ARN, leur synthèse et leurs fonctions au sein de la cellule, en section 1.1. Nous nous intéressons ensuite plus précisément aux caractéristiques des ARN codants dans la section 1.2, des ARN non-codants dans la section 1.3 et à leur évolution dans la section 1.4. Etant informaticien, toute cette présentation n'est pas rédigée par un spécialiste, et s'adresse en priorité à des non-spécialistes. Le but est de fournir les notions de base en biologie moléculaire nécessaires à la compréhension du document et de légitimer les choix de modèles que nous avons faits dans la suite du travail.

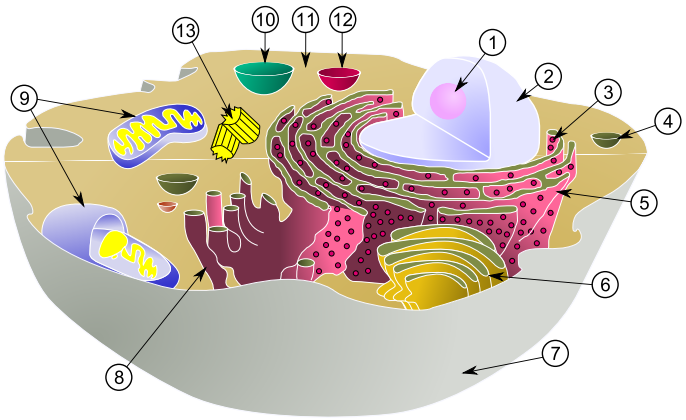
Enfin, dans la dernière section du chapitre, nous abordons les premiers formalismes et méthodes de bio-informatique avec la génomique comparative, l'alignement de séquences et l'introduction du concept de *méta-séquence*. Les méta-séquences et les ensembles de méta-séquences nous serviront tout au long de ce document pour représenter des ensembles de séquences d'ARN de distance évolutive hétérogène.

1.1 L'ARN au sein de la cellule

Les ARN font partie des molécules essentielles au bon fonctionnement d'un organisme vivant, et plus précisément à celui de ses cellules. Nous nous intéressons donc en premier lieu aux cellules des organismes vivants dans la section 1.1.1. Puis, dans la section 1.1.2, nous nous intéressons au dogme central qui décrit le cycle de vie des ARN.

1.1.1 Les organismes vivants

Un organisme vivant est un être issu de l'assemblage d'une ou plusieures entités microscopiques : les cellules. Les organismes vivants font l'objet d'une classification selon une multitude de critères dont le premier porte sur la structure de leurs cellules. Ce critère fait apparaître deux *domaines* distincts : les eucaryotes et les procaryotes. Les cellules eucaryotes comportent un noyau et plusieurs compartiments spécialisés alors que les cellules procaryotes n'ont pas de noyau. La figure 1.1 présente de manière schématique la structure d'une cellule eucaryote animale.



1. Nucléole
2. Noyau
3. Ribosome
4. Vésicule
5. Réticulum endoplasmique rugueux
6. Appareil de Golgi
7. Microtubule
8. Réticulum endoplasmique lisse
9. Mitochondrie
10. Vacuole
11. Cytoplasme
12. Lysosome
13. Centrosome

Source http://en.wikipedia.org/wiki/File:Biological_cell.svg

FIG. 1.1 – Schéma d'une cellule eucaryote animale.

Parmi les eucaryotes, on retrouve aussi bien des organismes unicellulaires, tels que les levures, que des organismes pluricellulaires tels que les plantes et les animaux. La majorité des procaryotes sont quant à eux des organismes unicellulaires microscopiques. Les procaryotes sont divisés en deux *règnes* : les bactéries et les archées. Bien que la taille et la forme des archées sont similaires à celles des bactéries, les archées s'en distinguent par des caractères plus similaires à ceux des eucaryotes tels que la structure des gènes et les mécanismes relatifs à leur expression.

Toute cellule est régie essentiellement par cinq types de macromolécules : les lipides, les glucides, les acides désoxyribonucléiques (ADN), les acides ribonucléiques (ARN) et les protéines. L'ADN assure le stockage de l'information génétique et sa transmission au fil des générations. L'ensemble de l'ADN qui définit un organisme est appelé son génome. Chez les eucaryotes, l'ADN est stocké dans le noyau. Les lipides sont les principaux constituants des membranes cellulaires. Les glucides sont le soutien de la vie : ils servent de source et de stockage d'énergie, ils participent aux paroies cellulaires, ... Les protéines sont des molécules indispensables à la structuration et au fonctionnement des cellules, et résultent de l'expression de l'information génétique. Les acides ribonucléiques (ARN) sont quant à eux des protagonistes plus ambigus qui peuvent endosser plusieurs rôles auxquels nous nous intéressons par la suite.

1.1.2 Le dogme central de la biologie moléculaire

Les mécanismes d'expression de l'information génétique sont formalisés dans le dogme central, proposé par Francis Crick à la fin des années 50, puis repris dans *Nature* en 70 [Cri70]. Le dogme central définit deux principes fondamentaux, la *transcription* et la *traduction* dont l'enchaînement est illustré sur le schéma de la figure 1.2.

L'information génétique contenue dans l'ADN est organisée en segments appelés les gènes. La première étape de l'expression de l'information génétique consiste à transcrire un gène en un ARN. Cet ARN est ensuite traduit en protéine. Ce rôle de médiateur de l'information

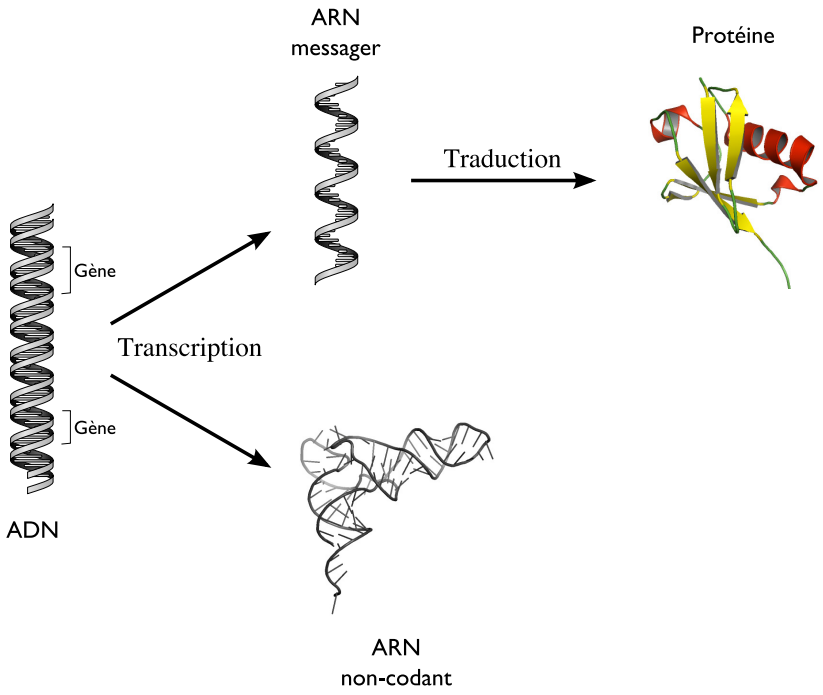


FIG. 1.2 – Le dogme central présentant la transcription de l'ADN en ARN messagers traduits par la suite en protéine, et la transcription de l'ADN en ARN non-codants.

génétique constitue le premier rôle de l'ARN que l'on nomme alors ARN messager.

Le dogme central mentionne également deux autres types d'ARN : les ARN ribosomiques et les ARN de transfert. Contrairement aux ARN messagers, ces ARN sont des molécules fonctionnelles non traduites en protéine et que l'on regroupe sous le terme d'ARN non-codants. Au moment de leurs découvertes, ces deux types d'ARN non-codants apparaissent comme des exceptions au dogme central. Depuis, de nombreux autres ARN non-codants ont été découverts, portant à plus de 600 le nombre de familles d'ARN non-codants connues à ce jour. Ces ARN sont impliqués dans de nombreux processus essentiels des cellules tels que la synthèse des protéines, la maturation des ARN messagers, les processus de régulation pré- et post-transcriptionnelle, ...

Afin de clarifier le discours, les gènes à l'origine d'ARN messagers seront par la suite appelés des gènes codants, et par opposition, les gènes à l'origine d'ARN non-codants, des gènes à ARN.

1.1.3 Les acides nucléiques

L'ADN et l'ARN sont tous deux des acides nucléiques, c'est-à-dire des chaînes plus ou moins longues de nucléotides. Chaque nucléotide est composé de trois substances fondamentales : un sucre, un groupe phosphate et une base azotée. La composition du groupe phosphate est constante pour tous les acides nucléiques, tandis que celle du sucre varie en fonction du type d'acide nucléique : le désoxyribose pour les nucléotides de l'ADN, le ribose pour ceux de l'ARN. Il existe en tout cinq types de nucléotides, induits par cinq bases azotées différentes : l'adénine (`A`), la cytosine (`C`), la guanine (`G`), la thymine (`T`) et l'uracile (`U`). La thymine et l'uracile sont très semblables, mais on ne rencontre la thymine que dans l'ADN, et l'uracile que

dans l'ARN. La structure chimique des bases azotées permet de distinguer deux groupes : les purines, constituées de l'adénine et de la guanine, et les pyrimidines, constituées de la cytosine, de la thymine et de l'uracile. L'alternance des phosphates et des sucres produit le squelette des acides nucléiques sur lequel s'attachent les bases azotées. La molécule ainsi formée est souvent appelée brin. Elle possède des extrémités différentes, notées $5'$ et $3'$ en raison de notations relatives à la géométrie des sucres.

Au sein d'un brin ou entre deux brins différents les bases peuvent s'apparier au moyen de liaisons hydrogène. La quantité de liaisons qui se forment entre deux bases détermine la stabilité de leur appariement. Les appariements de type Watson-Crick désignent ainsi les deux appariements les plus stables qui se forment entre l'adénine et la thymine (l'uracile pour l'ARN) reliées par deux liaisons hydrogène, et la cytosine et la guanine reliées par trois liaisons hydrogène. Pour faire référence à ces appariements, on parle également d'appariements canoniques ou encore de complémentarité entre les bases : l'adénine et la thymine (l'uracile pour l'ARN) sont complémentaires, de même que la cytosine et la guanine.

La composition des nucléotides n'est pas le seul élément qui diffère entre l'ADN et l'ARN. L'ADN est en fait composé de deux brins reliés et stabilisés par les appariements qui se forment entre leurs nucléotides respectifs. La figure 1.3 présente de manière schématique la structure chimique de l'ADN. Les brins d'ADN sont antiparallèles, c'est-à-dire que les extrémités chargées $5'$ et $3'$ de chacun des brins se font face sous la contrainte de leur polarité. Ils sont également complémentaires car ils ne sont reliés que par des appariements de type Watson-Crick. Ainsi assemblés, les deux brins se vrillent pour former une double hélice comme illustré en figure 1.4. Contrairement à l'ADN, l'ARN est généralement simple brin, sauf chez quelques organismes tels que les rétrovirus. Sous sa forme simple brin, l'ARN est plus malléable, ce qui lui permet de se replier sur lui-même et aux bases des nucléotides qui le composent de s'apparier.

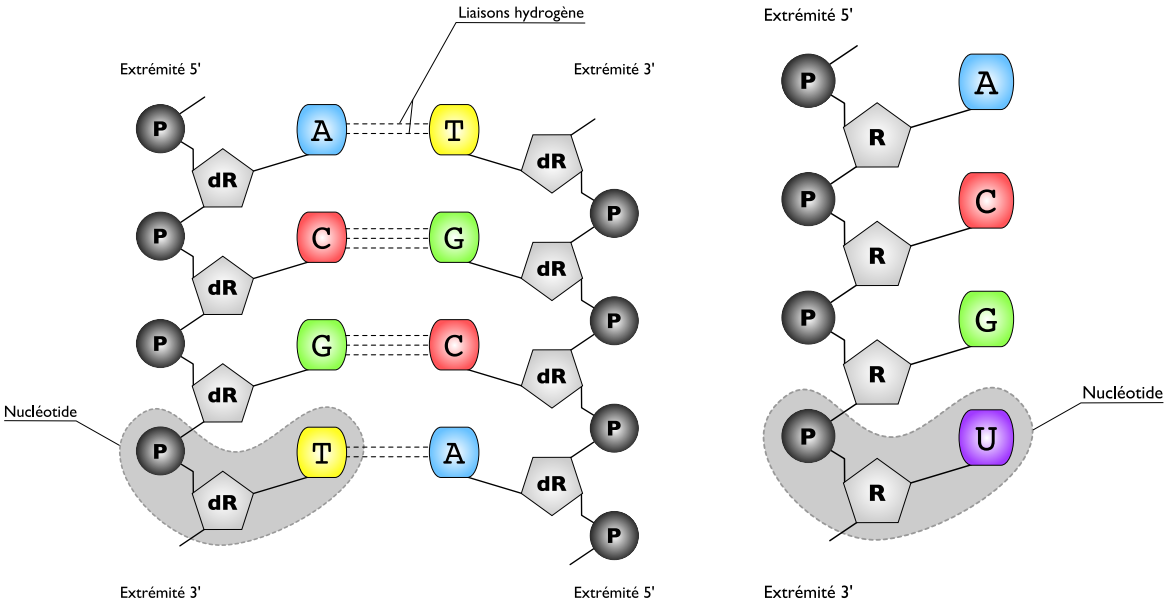
L'organisation de l'information génétique stockée dans l'ADN varie selon les organismes. Le génome des eucaryotes est généralement organisé en plusieurs molécules d'ADN empaquetées, les chromosomes. Le génome des procaryotes et des archées n'est en général constitué que d'un seul chromosome qui se présente sous forme circulaire, c'est-à-dire que les extrémités $5'$ et $3'$ de la molécule d'ADN sont liées ce qui a pour effet de fermer la molécule.

1.1.4 La transcription d'un gène en un ARN

La transcription est le processus qui synthétise un ARN en recopiant la séquence d'un gène. Ce processus se décompose en trois étapes schématisées sur la figure 1.5 : l'*initiation*, l'*élongation* et la *terminaison*.

Durant la phase d'initiation de la transcription, l'ARN polymérase, un complexe protéique, se fixe sur une région particulière de l'ADN, située en amont du gène à transcrire : le site promoteur. La liaison entre l'ADN et l'ARN polymérase permet d'une part d'ouvrir la double hélice et d'autre part de catalyser l'insertion des ribonucléotides pour former un brin d'ARN. Contrairement aux procaryotes, les eucaryotes et les archées disposent de quatre types d'ARN polymérases recrutées en fonction du type d'ARN à synthétiser et/ou du compartiment cellulaire de destination de l'ARN néo-synthétisé.

Le site promoteur diffère quelque peu selon les gènes et les organismes. Ce site comporte deux boîtes, c'est-à-dire deux séquences spécifiques. Chez les bactéries, la boîte de PRIBNOW, dont la séquence canonique est `TATAAT`, marque le début de transcription et se situe à une dizaine de bases en amont du gène. Chez les eucaryotes et les archées, la boîte de PRIB-



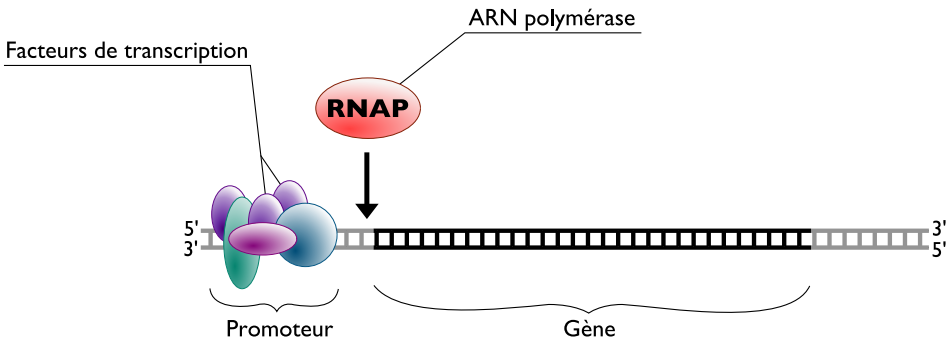
(a) Schéma de l'Acide DésoxyriboNucléique.

(b) Schéma de l'Acide RiboNucléique.

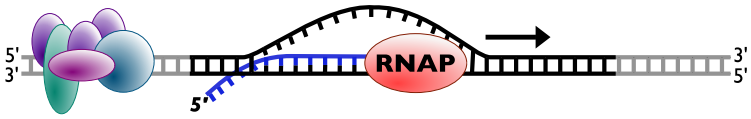
FIG. 1.3 – Structure des acides nucléiques. Chaque groupe phosphate (P) est lié à un sucre (dR ou R) lui-même lié à une base azotée (`A`, `C`, `G`, `T` ou `U`).

Source `http://openclipart.org/media/files/hs/1771`

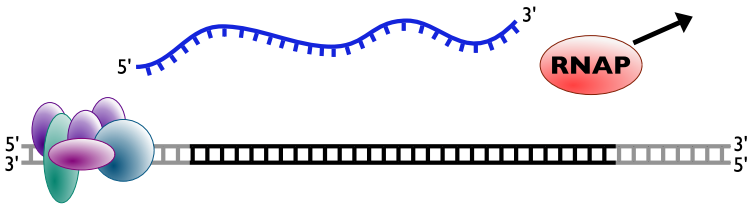
FIG. 1.4 – Schéma de la structure en double hélice de l'ADN.



(a) Initiation.



(b) Elongation.



(c) Terminaison.

Source http://en.wikipedia.org/wiki/File:Simple_transcription_initiation1.svg

FIG. 1.5 – Les trois étapes successives de la transcription.

NOW est l'équivalente de la boîte TATA, dont la séquence canonique est `TATAAAA`, située une vingtaine de bases en amont du gène. Pour tous les organismes, il existe la boîte CAAT située 70 à 80 nucléotides en amont du gène qui sert à la régulation de la vitesse de transcription du gène. Lorsque l'ARN polymérase se fixe sur la boîte TATA, elle s'associe avec différentes protéines, les facteurs de transcription, pour former une particule d'initiation. L'élongation de la transcription correspond à l'incorporation des nucléotides sur le brin d'ARN. Durant cette phase, l'ARN polymérase progresse de manière séquentielle de l'extrémité $3'$ vers l'extrémité $5'$ du brin d'ADN codant, c'est-à-dire le brin complémentaire du brin contenant le fragment à recopier. L'incorporation des nucléotides se faisant par complémentarité entre nucléotides, l'ARN synthétisé est une copie conforme de la région à transcrire. La terminaison de la transcription intervient lorsque l'ARN polymérase rencontre un terminateur. Chez les procaryotes, ce terminateur est le plus souvent une région riche en `G` et en `C` qui contient une petite structure en tige-boucle (section 1.3.1), suivie d'une série de `A` sur l'ADN. Chez les eucaryotes, les mécanismes de terminaison de la transcription sont moins connus.

Chez les procaryotes, des groupes de gènes contiguës, appelés des opérons, peuvent partager un même promoteur et se retrouver ainsi transcrits simultanément en un seul ARN. Cette organisation "optimisée" permet l'expression et la régulation simultanée de plusieurs gènes impliqués dans un même processus cellulaire.

1.1.5 La maturation de l'ARN

Un ARN nouvellement transcrit est appelé *transcrit primaire*. Chez les eucaryotes et les archées les transcrits primaires d'ARN subissent quelques transformations post-transcriptionnelles. Cette phase dite de maturation des transcrits comporte trois étapes schématisées sur la figure 1.6 : l'addition d'une coiffe en $5'$ du transcrit, l'addition d'une queue poly `A` en $3'$ et enfin l'épissage du transcrit primaire. L'épissage est le changement le plus marquant au cours duquel des fragments de l'ARN sont excisés, et les fragments restants sont raboutés. Les fragments excisés sont nommés des introns, les fragments conservés des exons. Les jonctions intron/exon sont délimitées par deux sites, le site donneur `GU` qui marque le début d'un intron, et le site accepteur `AG` qui en marque la fin. L'ablation des introns est réalisée par des ribonucléoprotéines, complexes composés de protéines et de petits ARN, des snRNA. Le découpage en exons et en introns n'est pas nécessairement unique. Ainsi, un même transcrit primaire peut donner lieu à différents transcrits matures de longueurs différentes issus d'épissages alternatifs. Aujourd'hui, il est admis que près de 60% des gènes chez l'être humain subissent l'épissage alternatif. Quelques cas extrêmes d'épissage alternatif sont connus, comme par exemple le gène *Dscam* de la Drosophile pour lequel il existe 38 016 ARN matures différents [WFM$^+$04, BK06, SC09].

La maturation concerne tous les transcrits issus de gènes codants chez les eucaryotes et les archées.

1.2 Les ARN codants

Dans le cas des gènes codants, l'ARN messager issu de la transcription, puis de la maturation éventuelle, contient toute l'information nécessaire à la production d'une protéine. La traduction d'un ARN messager mature en une protéine est un processus complexe qui repose sur une cascade d'assemblages de molécules en interaction avec l'ARN messager à traduire.

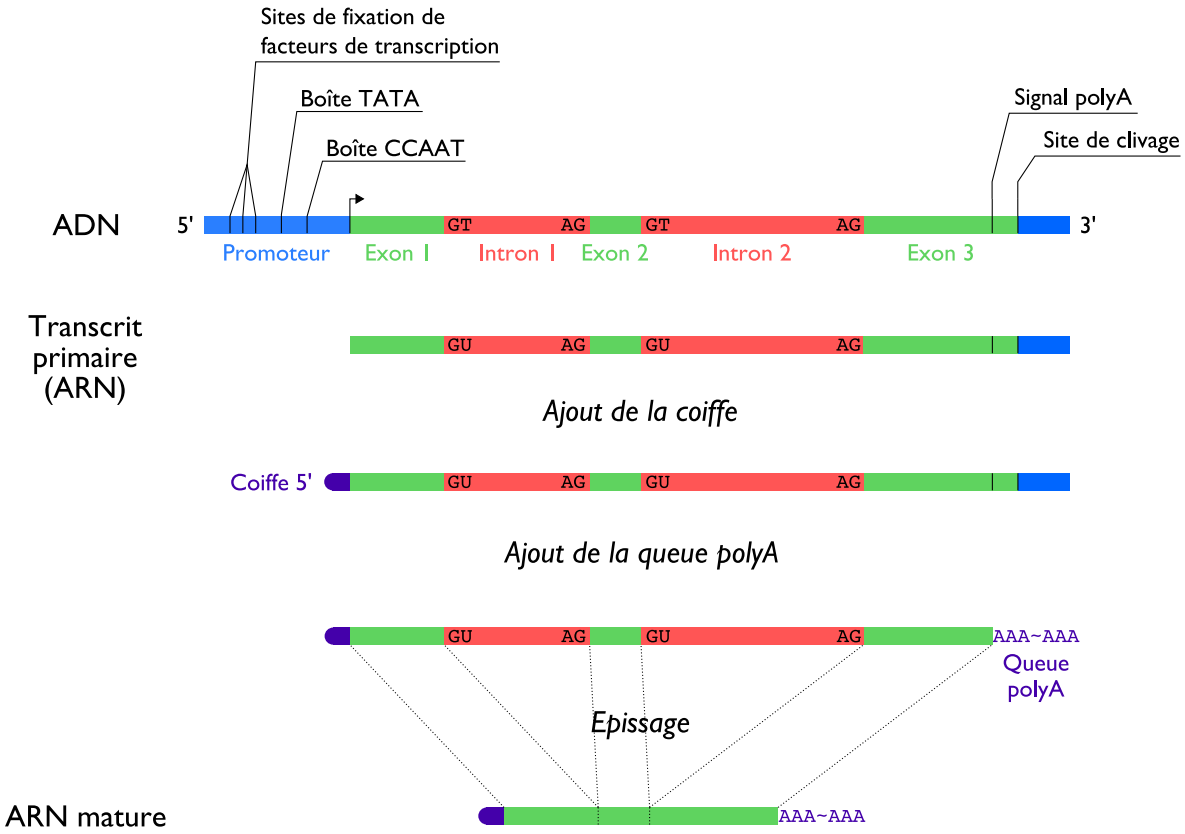
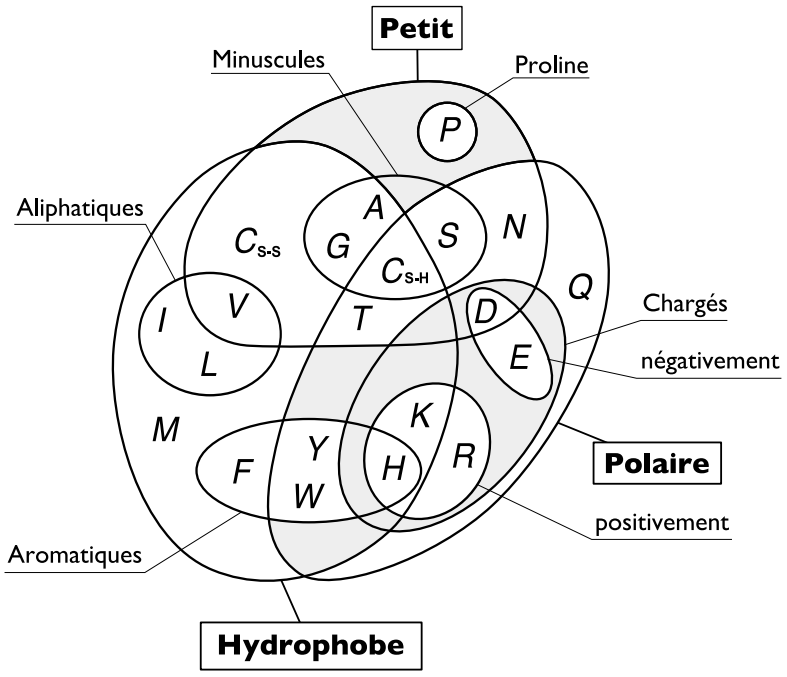


FIG. 1.6 – Processus de maturation des transcrits primaires.



Source http://fr.wikipedia.org/wiki/Fichier:Acides_amin%E9s_propri%E9t%E9s_diagramme_Venn.svg

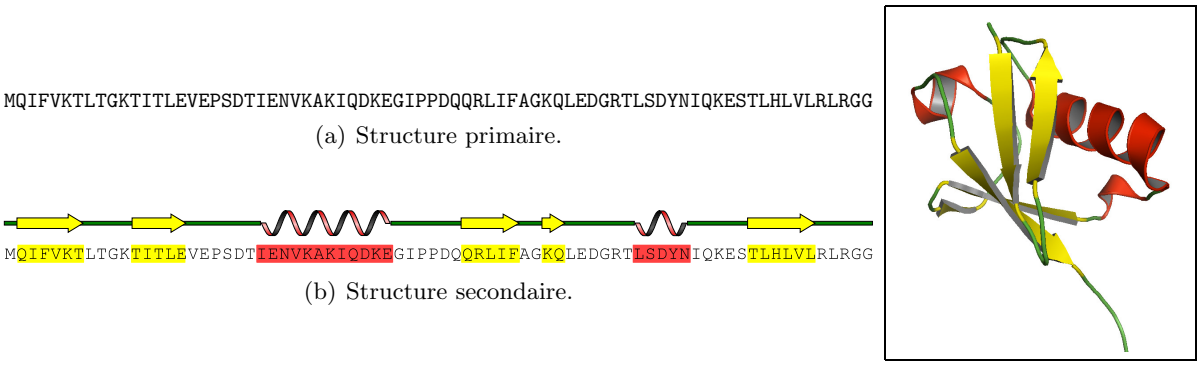
FIG. 1.7 – Diagramme de Venn des acides aminés selon leur propriétés.

1.2.1 Les protéines

La traduction est un processus qui, comme son nom l'indique, traduit l'information portée par un ARN messager en une protéine. Une protéine est une molécule formée par l'enchaînement d'acides aminés liés entre eux par des liaisons peptidiques. Il existe plus d'une centaine d'acides aminés présents dans la nature [CPL+07], cependant, seuls vingt deux d'entre eux peuvent être intégrés dans les protéines synthétisées par la traduction d'un ARN. La figure 1.7 montre une classification de ces acides aminés selon leurs propriétés.

Les biochimistes distinguent quatre niveaux pour la structure d'une protéine, illustrés en figure 1.8 :

- la structure primaire : la séquence d'acides aminés ;
- la structure secondaire : des éléments de structure locaux, stabilisés par des liens hydrogène. Les éléments les plus fréquents sont les hélices alpha et les feuillets bêta ;
- la structure tertiaire : la structure tridimensionnelle de la protéine, où les éléments de la structure secondaire sont en interaction. La structure tertiaire peut être stabilisée par la formation de quelques liaisons hydrogènes, ponts disulfides, ... ;
- la structure quaternaire : la structure tertiaire d'une protéine dans une conformation particulière, souvent en interaction avec une ou plusieurs autres molécules, notamment lors d'assemblages plus conséquents ou lors de la formation d'un complexe de protéines.



(a) Structure primaire.

(b) Structure secondaire.

(c) Structure tertiaire.

FIG. 1.8 – Structure de l'ubiquitine d'*Homo sapiens*. Les hélices alpha sont en rouge, les feuillets bêta en jaune.

1.2.2 La traduction en protéine

La traduction est un processus séquentiel et linéaire qui consiste à décoder l'information contenue dans un ARN messager mature pour obtenir la protéine correspondante. Durant la traduction, un complexe nommé le ribosome progresse le long de l'ARN messager à traduire en lisant les triplets successifs de nucléotides appelés codons. A chaque lecture de chaque triplet, l'acide aminé correspondant est ajouté à la protéine en cours d'assemblage. La correspondance entre codons et acides aminés est presque universelle et régie par le code génétique donné en figure 1.9. Dans la cellule, ce sont les ARN de transfert, des ARN non-codants, qui sont garants de cette correspondance entre codons et acides aminés. La figure 1.10 présente un exemple de traduction partielle d'une séquence codante.

UUU Phe F	UCU Ser S	UAU Tyr Y	UGU Cys C
UUC	UCC	UAC	UGC
UUA Leu L	UCA	UAA STOP *	UGA STOP *
UUG	UCG	UAG STOP *	UGG Trp W
CUU Leu L	CCU Pro P	CAU His H	CGU Arg R
CUC	CCC	CAC	CGC
CUA	CCA	CAA Gln Q	CGA
CUG	CCG	CAG	CGG
AUU Ile I	ACU Thr T	AAU Asn N	AGU Ser S
AUC	ACC	AAC	AGC
AUA	ACA	AAA Lys K	AGA Arg R
AUG Met M	ACG	AAG	AGG
GUU Val V	GCU Ala A	GAU Asp D	GGU Gly G
GUC	GCC	GAC	GGC
GUA	GCA	GAA Glu E	GGA
GUG	GCG	GAG	GGG

FIG. 1.9 – Le code génétique universel. Couleurs inspirées de RASMOL.

Etant donné une séquence nucléique à traduire, il existe trois cadres de lecture potentiels

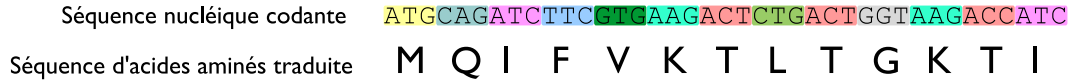


FIG. 1.10 – Traduction du début d'un gène codant pour la poly-ubiquitine C. Source : REFSEQ, *Homo sapiens*, NM_021009.

selon la position à laquelle débute la lecture des codons, auxquels s'ajoutent trois autres cadres de lecture lorsque l'orientation de la molécule est indéterminée. Toutefois, un seul cadre de lecture parmi les six permet d'obtenir la séquence de codons codant la protéine synthétisée. Certains codons particuliers, les codons START et STOP, marquent respectivement le début et la fin de la séquence de codons à traduire d'un ARN messager. Mises à part quelques exceptions, les codons START et STOP sont bien déterminés : `AUG` pour le codon START, `UAA`, `UAG` et `UGA` pour le codon STOP. L'enchaînement ininterrompu de codons effectivement traduits correpond à la séquence codante d'un ARN messager et est appelé le *cadre ouvert de lecture*.

Même si l'immense majorité des organismes vivants utilisent le code génétique standard, on note toutefois quelques exceptions à cette règle chez certains organismes pour lesquels les acides aminés codés ne sont pas les mêmes. Par exemple, chez les champignons *Candida*, le codon `CUG` habituellement traduit par la leucine correspond à la sérine, ou encore chez certains procaryotes où le codon STOP `UAG` code parfois pour un acide aminé supplémentaire, la pyrrolysine.

1.2.3 La régulation de la transcription

L'orchestration de la traduction est avant tout conduite par la présence de signaux organisés dans l'ARN messager à traduire. La plupart de ces signaux sont des motifs caractéristiques qui permettent la fixation d'autres molécules. Parmi ces molécules, une partie sont requises pour la traduction effective en protéine, alors que d'autres participent à sa régulation.

La figure 1.11 représente de manière schématique l'organisation d'un ARN messager mature codant pour une protéine. Le cadre ouvert de lecture est flanqué de deux régions qui ne sont pas traduites mais qui contiennent des signaux nécessaires à la machinerie traductionnelle. Certains sont communs à tous les ARN messagers, d'autres signaux sont variables d'un ARN à un autre et sont le plus souvent impliqués dans la régulation de la traduction. La région 5′ non traduite contient notamment le site de fixation du complexe responsable de la traduction nommé le *ribosome*. Chez les procaryotes, la séquence du site de fixation du ribosome a pour forme canonique `AGGAGGU`, et est également connue sous le nom de séquence de Shine Dalgarno.

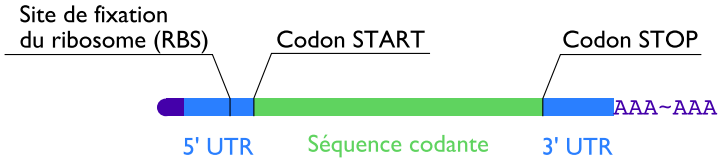


FIG. 1.11 – Organisation d'un ARN messager mature.

La région 3′ non traduite d'un ARN messager contient le signal de poly-adénylation utilisé lors de la maturation, mais également des sites de fixation pour des protéines dont le rôle est d'orienter vers sa destination finale dans la cellule l'ARN messager puis la protéine produite. Les régions 5′ et 3′ non traduites peuvent en sus comporter d'autres signaux pour la régulation traductionnelle, c'est-à-dire des sites destinées à la fixation d'autres molécules venant activer ou au contraire éteindre la traduction.

La traduction est un processus séquentiel et linéaire. Certains éléments dans l'ARN messager peuvent parfois perturber la lecture des triplets par le ribosome allant jusqu'à entraîner un glissement du ribosome d'une ou plusieurs bases "en avant" ou "en arrière". Le glissement du ribosome induit un changement de cadre de lecture, également appelé frameshift. Ce processus est induit par certaines répétitions ou motifs dans la séquence souvent accompagnés d'une petite structure locale formée par l'ARN messager. Bien qu'il s'agisse le plus souvent d'une erreur, le glissement du ribosome peut être une action programmée. Plusieurs études mettent en évidence des changements de cadre de lecture programmés chez les virus, les éléments transposables (section 1.4.2) [Jac88, GA96].

1.3 Les ARN non-codants

A l'inverse des ARN messagers issus de la transcription de gènes codants, les ARN non-codants issus de la transcription de gènes à ARN n'ont pas vocation à coder pour des protéines. Les ARN non-codants assurent diverses fonctions pour la plupart déterminées par la structure spatiale qu'adopte la molécule d'ARN en se repliant sur elle-même. Les ARN non-codants qui ne se replient pas de manière spécifique s'associent le plus souvent à d'autres molécules telles ques des protéines pour former des complexes. Les fonctions de ces ARN "non-structurés" sont alors en lien étroit avec leur séquence. Dans la suite de cet ouvrage, nous nous intéressons essentiellement aux ARN non-codants qui adoptent une structure caractéristique que nous tentons de prédire par des moyens informatiques.

1.3.1 La structure de l'ARN

Contrairement à l'ADN, l'ARN est une molécule simple brin qui a la capacité de se replier sur lui-même permettant ainsi à ses bases de s'apparier entre elles. Ces appariements se font de manière contiguë pour former des *tiges*. Les régions non appariées forment alors des *boucles* (figure 1.12).

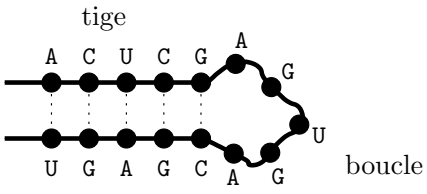
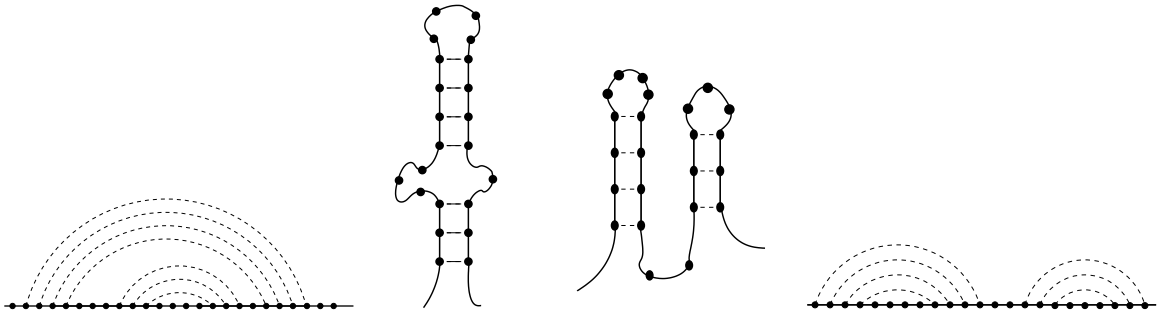


FIG. 1.12 – Exemple de formation d'une tige-boucle.

Une structure est décrite par une classification en quatre niveaux hiérarchiques :
– la *structure primaire* est simplement la séquence, orientée de 5′ en 3′, des bases qui composent la molécule ;



(a) Tiges emboîtées. (b) Tiges juxtaposées.

FIG. 1.13 – Conformations possibles des tiges des structures secondaires.

- la *structure secondaire* est l'ensemble des appariements sans croisement, formant des tiges emboîtées ou juxtaposées comme illustré sur les schémas de la figure 1.13 ;
- la *structure tertiaire* est l'ensemble de tous les appariements. En plus des appariements de la structure secondaire, les appariements suivants sont donc autorisés : les pseudo-nœuds (appariements chevauchants), les triplets (appariements à trois), les quadruplets (à quatre) et les appariements isolés ;
- la *structure spatiale* désigne la configuration de la molécule dans l'espace, généralement en interaction avec d'autres molécules.

La figure 1.14 présente les différentes manières usuelles de représenter les structures d'ARN selon cette hiérarchie. La figure 1.15 montre la structure tertiaire d'un ARN ribosomique 18S de levure.

La stabilité d'une molécule d'ARN est mesurée par son *énergie libre* qui est issue des principes de la thermodynamique. Plus l'énergie libre d'une structure est faible, plus celle-ci est stable. Les tiges stabilisent une structure, tandis que les boucles la déstabilisent. La stabilité apportée par une tige est fonction de sa longueur et de la nature de ses appariements : les appariements canoniques (`G`≡`C`, `A`=`U` et `G`=`U`) sont plus stables que les appariements non canoniques (`G`−`A`, `C`−`U`, ...). Toutes ces caractéristiques sont reprises dans le modèle d'énergie libre de Turner [TSF88, MSZT99].

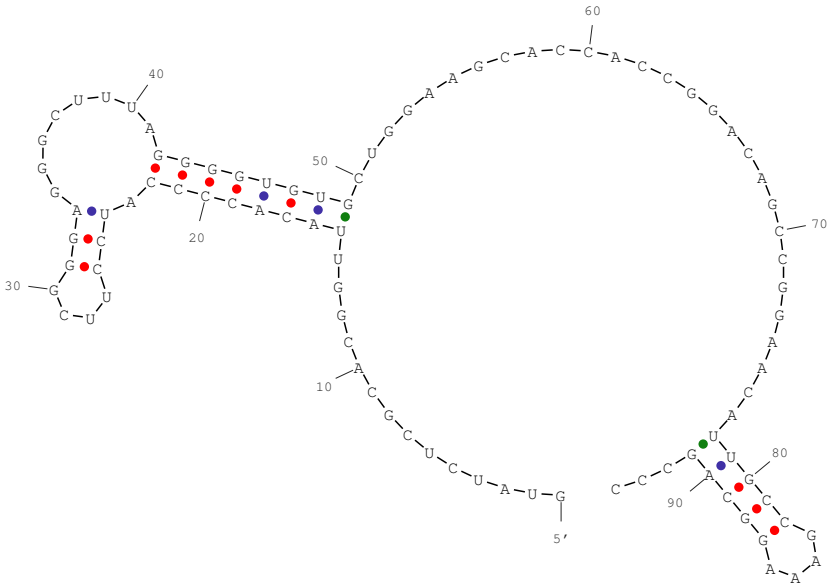
La description d'une structure secondaire ou tertiaire d'un ARN par la simple énumération de ses appariements peut s'avérer insuffisante pour décrire les interactions avec d'autres molécules. Ces interactions font intervenir des *motifs* particuliers dont la description requiert le plus souvent l'utilisation d'une représentation plus fine. A cet effet, la classification de Leontis-Westhof [LW01] a largement été adoptée par la communauté pour la description d'intéractions tridimensionnelles caractéristiques. L'observation de structures tridimensionnelles réelles a notamment permis l'élaboration de la base de données SCOR [KTHB02] contenant plus de 8 000 motifs récurrents.

1.3.2 Les familles d'ARN non-codants

Actuellement, plus de 600 familles d'ARN non-codants sont connues et répertoriées dans des banques de données publiques généralistes comme RFAM [GJBM+03, GJMM+05], NONCODE [CBG+05] et FRNADB [KYT+07, MYH+08], ou dans des banques plus spécifiques dédiées à certaines familles d'ARN non-codants ou d'organismes. Les ARN non-codants interviennent dans de nombreux processus essentiels de la cellule. En 1978, le premier ARN

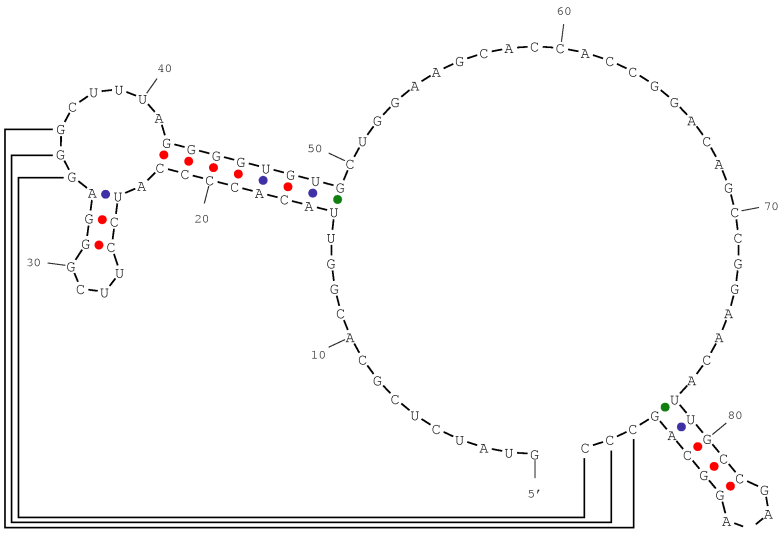
GUAUCUCGCACGGUUACACCCCAUCCUUCGGGAGGGCUUUAGGGGUGUGCUG
GAAGCACCACCGGACAGCCGGAACAUUGCCGAAAGGCAGCCC

(a) Structure primaire.



(b) Structure secondaire.

(c) Structure secondaire sous forme arc-annotée.



(d) Structure tertiaire.

FIG. 1.14 – Structures de l'ARN. Exemple d'un élément non traduit structuré de l'ARN génomique du Tombusvirus. Source : RFAM, RF00176.

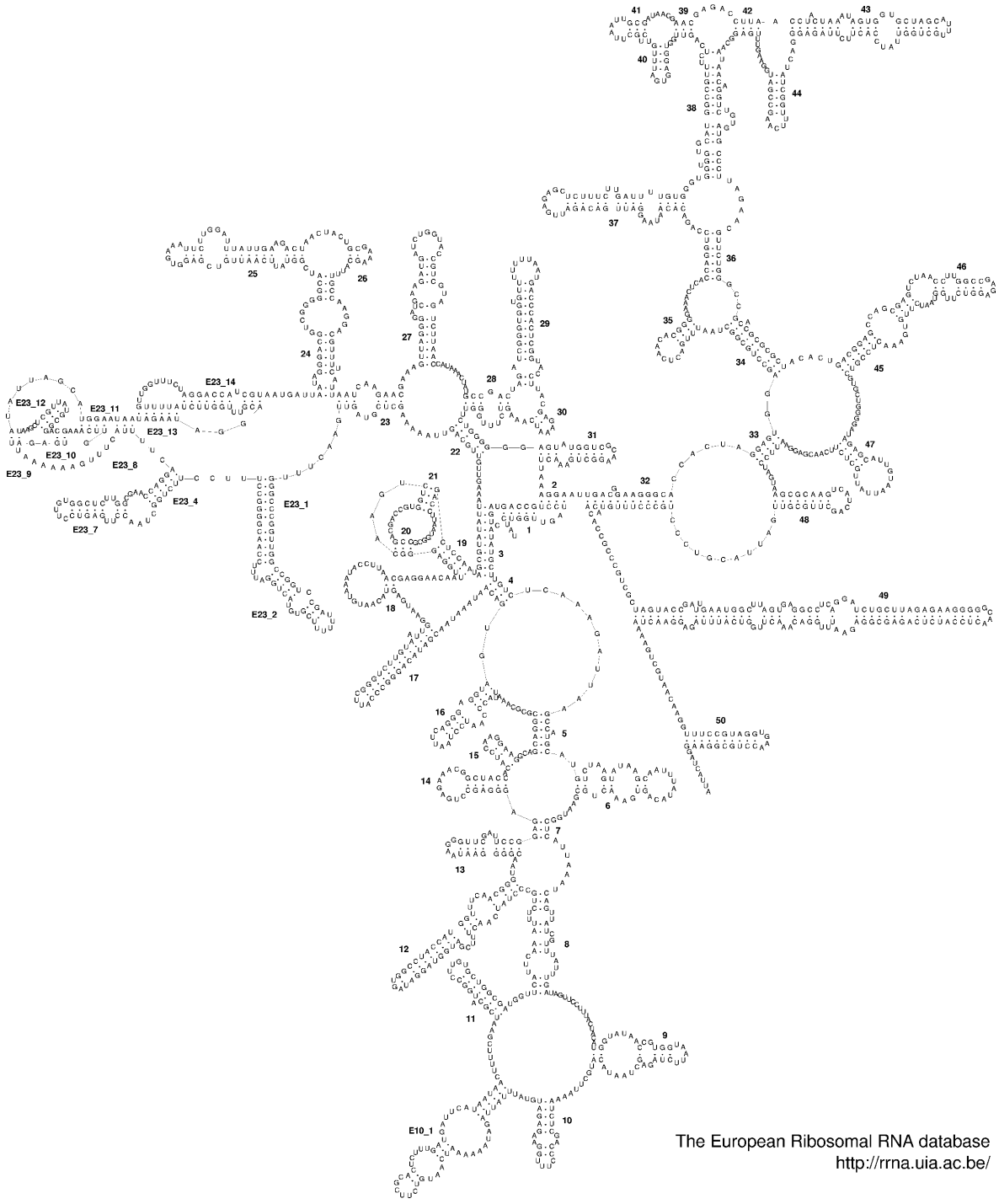


FIG. 1.15 – L'ARN ribosomique 18S de *Saccharomyces cerevisiae*.

présentant des propriétés catalytiques a été découvert, couplé à une protéine, la ribonucléase P (RNase P). Cette découverte décisive d'ARN aux propriétés catalytiques, les *ribozymes*, a été couronnée par un prix Nobel de chimie en 1989 [SKBA78]. Plus récemment, plusieurs études ont révélé l'existence de nombreux petits ARN [Edd01, HSP05, WPK$^+$07]. Déjà en 1979, de petits ARN inconnus avaient été isolés [LS79]. Ces ARN forment un complexe avec une protéine, le complexe ribonucléoprotéique (RNP), responsable de l'altération de certains ARN messagers. Ces ARN ont par la suite été nommés petits ARN nucléaires (snRNA) car on les trouve exclusivement dans le noyau des cellules, lieu de la transcription. En 1997, d'autres petits ARN ont été trouvés dans le nucléole, un pseudo-compartiment du noyau [MH97]. Ces petits ARN nucléolaires (snoRNA), servent à guider une enzyme vers une base précise d'un ARN ribosomique à modifier. A ce jour, deux sous-familles sont connues : les petits ARN nucléolaires à boîte C/D et à boîte H/ACA. Chacune est caractérisée par un motif qui guide avec précision l'enzyme vers la base qu'elle doit modifier. En 2001, une famille d'ARN impliquée dans la régulation de la traduction, les micro ARN, a été découverte grâce à une approche bio-informatique, l'analyse comparative de génomes, confirmée par des méthodes plus classiques de biochimie [Ruv01].

1.4 L'évolution des acides nucléiques

Qu'il soit codant ou non-codant, la fonction d'un ARN est déterminée par la séquence de ses bases, identique, au moins par morceaux, à celle du gène dont il est issu. Néanmoins, pour un ARN non-codant ou une protéine donnée, la séquence du gène correspondant n'est pas nécessairement identique d'un organisme à un autre, d'un individu à un autre. A l'origine de cette diversité : l'évolution. Au fil des générations d'une population d'individus, et sous l'effet de facteurs extérieurs, les séquences des gènes évoluent. L'évolution des séquences nucléiques est ainsi le moteur de l'évolution des espèces.

1.4.1 Généralités

L'évolution est causée par la présence de variations parmi les traits héréditaires d'une population, et par divers mécanismes qui favorisent la propagation de certains traits plutôt que d'autres. Par *la sélection naturelle*, les traits héréditaires favorisant la survie et la reproduction des individus d'une population voient leurs fréquences croître d'une génération à l'autre. La *pression de sélection* correspond à un ensemble de contraintes environnementales auxquelles est assujettie une population d'individus, comme par exemple la composition chimique d'un milieu.

D'un point de vue génétique, l'évolution des espèces, guidée par la sélection naturelle et la pression de sélection, est engendrée par une évolution des séquences génomiques. Plusieurs mécanismes sont impliqués dans l'apparition de mutations dans les séquences génomiques, c'est-à-dire des altérations induites ou spontanées des acides nucléiques.

1.4.2 Les mécanismes de l'évolution

Quelque soit l'origine des mutations, les conséquences varient en fonction non seulement de leur nombre, mais également des positions altérées et de la nature même des changements opérés. Nous allons distinguer deux classes de transformations : les transformations macro-

scopiques "visibles" à l'échelle génomique, et les transformations microscopiques qui altèrent le contenu même des séquences.

A l'échelle génomique

Essentiellement trois familles de mécanismes interviennent dans l'évolution des acides nucléiques à l'échelle génomique : les transferts horizontaux, les éléments transposables et les recombinaisons chromosomiques. Ces mécanismes sont étroitement liés à la plasticité du génome et produisent des insertions et/ou des délétions de séquences plus ou moins longues de nucléotides dans l'ADN, soit délibérément, soit par erreur.

Un *transfert horizontal* est une intégration d'un fragment d'ADN au sein du matériel génétique d'un organisme. Il peut prendre trois formes : la transformation, la conjugaison et la transduction. La *transformation* consiste pour un organisme à insérer dans son génome du matériel génétique présent dans son environnement proche. Ce matériel peut, par exemple, provenir d'un autre organisme mort dont l'ADN s'est retrouvé "libéré". Les transferts horizontaux concernent principalement les organismes procaryotes mais ont déjà été observés chez des eucaryotes unicellulaires comme les levures et les champignons. Comme son nom l'indique, la *conjugaison bactérienne* ne s'applique qu'aux bactéries. Ce processus est un échange unidirectionnel de matériel génétique d'une bactérie vers une autre bactérie. Après avoir recopié une portion de son ADN, appelée plasmide, une bactérie transfère cette copie à une autre bactérie par le biais d'un canal chimique. Une fois transmise, cette copie peut ou non être intégrée dans le génome de la bactérie receveuse. Le dernier type de transfert horizontal est la *transduction*. L'intégration de matériel génétique étranger est ici réalisée par un virus. Tous les virus ont un cycle lytique, ou infectieux, pendant lequel ils injectent leur matériel génétique. La machinerie cellulaire de l'hôte est alors détournée pour produire des copies du virus jusqu'à ce que la cellule hôte éclate. Chaque copie produite peut alors infecter une autre cellule et soit suivre un nouveau cycle lytique, soit suivre un cycle lysogénique. Lors d'un cycle lysogénique, le matériel génétique du virus s'intègre au matériel génétique de l'hôte qui le transmet à ses descendants. Durant le cycle lytique du virus, une partie de l'ADN de l'hôte peut être "emportée" avec une copie de l'ADN du virus. S'il s'en suit un cycle lysogénique, alors le fragment d'ADN emporté est intégré au matériel génétique de l'hôte en même temps que l'ADN du virus, ce qui constitue la transduction.

Un *élément transposable* est une séquence capable de se déplacer et/ou de se multiplier de manière autonome dans un génome. On distingue couramment deux types d'éléments transposables selon leur mode de transposition : les éléments à ARN, ou de classe I, et les éléments à ADN, ou transposons de classe II. Les éléments de classe I, présents uniquement chez les eucaryotes, fonctionnent selon le principe du "copier-coller", c'est-à-dire qu'ils ne se déplacent pas dans un génome mais qu'une copie de l'élément est insérée ailleurs dans le génome. Ce processus, appelé transposition réplicative, fonctionne en trois temps : la séquence génomique de l'élément est transcrite en ARN, puis l'ARN synthétisé est rétro-transcrit en ADN, et ce fragment d'ADN est finalement inséré dans le génome. Les éléments de classe I sont ainsi appelés rétro-transposons ou rétro-posons. Les éléments de classe II, également appelés transposons, peuvent être sujets à une transposition réplicative ou conservative. La transposition conservative suit alors le principe du "couper-coller", c'est-à-dire que l'élément transposable est excisé du génome puis réinséré ailleurs. Quelque soit la nature de la transposition, celle-ci ne se fait pas de manière aléatoire, mais au niveau d'un court motif spécifique dans la séquence génomique. Toutefois, l'identification du site peut être approximative provoquant

l'excision ou l'insertion d'un fragment de séquence voisin de l'élément transposable lors d'une transposition.

Les *recombinaisons chromosomiques* interviennent exclusivement chez les organismes eucaryotes durant la reproduction sexuée des espèces, et plus particulièrement au cours de la méïose, c'est-à-dire la division d'une cellule permettant d'obtenir quatre cellules sexuelles. Les recombinaisons assurent le brassage génétique par la formation de nouvelles combinaisons génétiques, essentielles à la diversité d'une population et à l'évolution des espèces. Il existe deux principes complémentaires de recombinaisons chromosomiques : les recombinaisons inter-chromosomiques et les recombinaisons intra-chromosomiques. La *recombinaison inter-chromosomique* désigne la séparation aléatoire des chromosomes homologues d'une cellule au cours de sa méïose. Etant donné une cellule comportant n paires de chromosomes homologues, la méïose de cette cellule produit quatre cellules comportant chacune un exemplaire de chaque paire. Le nombre de combinaisons de chromosomes possibles est donc de 2^n ce qui, pour une cellule humaine composée de 23 paires de chromosomes, représente plus de huit millions de combinaisons. La *recombinaison intra-chromosomique*, également appelée enjambement, est un échange de segments entre deux chromosomes homologues au niveau de sites précis des chromosomes, appelés chiasmas. On dénombre en moyenne entre un et cinq sites possibles entre deux chromosomes homologues. La recombinaison intra-chromosomique peut-être déséquilibrée, c'est-à-dire qu'un fragment d'ADN peut être inséré ou au contraire délété dans les chromosomes. Dans ce cas, les conséquences varient selon la longueur du fragment inséré ou délété, mais surtout de la région affectée.

A l'échelle nucléotidique

Les mécanismes décrits précédemment ont des conséquences à l'échelle des génomes en provoquant des insertions ou des délétions de fragments de séquences entiers relativement longs. Les mécanismes auxquels nous allons maintenant nous intéressés sont plus discrets et portent sur l'insertion, la délétion et la substitution, c'est-à-dire le remplacement, de quelques nucléotides.

Les insertions de nucléotides peuvent être le résultat de duplications de certains fragments, à l'image des transpositions réplicatives à l'échelle du génome. Certains facteurs extérieurs peuvent également à l'origine d'insertion ou la délétion de nucléotides, telle que l'exposition à des rayons ultraviolets par exemple.

Les substitutions de nucléotides peuvent être classées en deux groupes : les *transitions* et les *transversions*. Une transition est une substitution d'une purine par une autre purine, ou d'une pyrimidine par une autre pyrimidine, tandis qu'une transversion est une substitution d'une pyrimidine par une purine ou inversement. La transition d'un `C` en `U` est un phénomène fréquent résultant de la dégradation spontanée par désamination de la cytosine. Cette modification est réversible grâce à un processus qui détecte l'uracile dans l'ADN. Cependant si la réparation n'est pas effectuée avant la prochaine réplication de l'ADN, la guanine appariée à la cytosine d'origine sur le brin opposé est substituée par une adénine lors de la réplication, et l'uracile remplacée par une thymine. Le second type de mutation spontanée est lié au dinucléotide $5'$–`CG`, appelé un "point chaud", car il est l'objet de fréquentes mutations lorsque la cytosine en $5'$ est sous sa forme méthylée. La désamination de la méthylcytosine produit en effet une thymine, qui ne peut alors être reconnue par aucun mécanisme de réparation. Le troisième type de mutation spontanée est l'oxydation des nucléotides par les radicaux libres de l'oxygène, sous-produits du métabolisme oxydatif normal des cellules. Une guanine oxydée,

par exemple, s'apparie à tort avec une adénosine induisant une transversion de G-C en T-A.

Les probabilités d'apparition, de conservation et de transmission d'une mutation sont les objets de nombreuses études [Rus93, Cro97, DCCC98]. Dans le cadre de nos travaux, nous nous sommes plus particulièrement intéressés aux mutations dans les séquences des gènes codants et des gènes à ARN car sous la pression de sélection, ces mutations suivent certains schémas à l'origine de biais locaux.

1.4.3 L'évolution des gènes codants

Au cours de l'évolution, les séquences fonctionnelles d'un génome sont soumises à la pression de sélection. Pour les séquences codantes, cette pression de sélection porte sur la fonctionnalité de la protéine produite. Les mutations dans une séquence codante ont en particulier des effets très variables sur la protéine codée.

Les substitutions dans une séquence codante sont classées en trois catégories en fonction de leur impact sur le codon modifié :

- une mutation faux-sens : le codon affecté ne code plus pour le même acide aminé. L'impact de ce type de mutation sur la protéine produite dépend du rôle de l'acide aminé original dans la protéine et de l'acide aminé qui lui a été substitué. On parle parfois de mutation synonyme lorsque le nouvel acide aminé codé a des propriétés physico-chimiques proches de l'acide aminé codé avant la mutation (section 1.2.1) ;
- une mutation non-sens : le codon affecté ne code plus pour un acide aminé mais pour un codon STOP. La protéine produite est alors tronquée comme cela peut se produire avec une insertion ou une délétion décalante ;
- une mutation silencieuse : l'acide aminé codé reste le même, donc cette mutation n'a aucune conséquence sur la protéine codée. Ce type de substitution est rendu possible grâce aux nombreuses redondances dans le code génétique. Cette propriété est également appelée dégénérescence du code génétique.

Une insertion ou une délétion dans une séquence codante peut allonger ou réduire la longueur de la protéine codée. En particulier, on parle de mutation décalante si la séquence insérée ou supprimée provoque un changement de cadre de lecture, c'est-à-dire lorsque la longueur de la séquence en question n'est pas multiple de trois. Le plus souvent, une mutation décalante entraîne l'apparition d'un codon STOP prématuré dans le cadre de lecture, ce qui a pour effet de produire une protéine tronquée lors de la traduction. Dans la plupart des cas, la protéine ainsi tronquée n'est plus capable d'assurer sa fonction. Il existe des exemples où ce type de mutation n'est pas létal et peut même conférer un avantage significatif aux individus qui en sont porteurs, comme par exemple le variant CCR5-Δ32 du gène CCR5 [GS03]. Le gène CCR5 code pour une protéine qui se trouve à la surface de certaines cellules, notamment des cellules immunitaires, servant de récepteur à chemokine. Le variant Δ32 du gène CCR5 est issu de la délétion décalante de 32 nucléotides dans le gène et dont la traduction génère une protéine plus courte incapable d'assurer son rôle. Cependant, cette mutation prodigue aux individus porteurs une immunité naturelle à certains virus comme la petite variole et le HIV car ces virus sont alors incapables d'infecter les cellules immunitaires dont les récepteurs à chemokine sont absents ou non fonctionnels. Le variant CCR5-Δ32 est largement répandu de nos jours en Europe où il touche entre 5 et 14% de la population, mais est beaucoup plus rare en Afrique ou en Asie. L'hypothèse la plus vraissemblable est que ce variant a fait l'objet d'une sélection naturelle au cours de la pandémie de peste noire qui a décimé plus d'un tiers de la population européenne au XIVème siècle, les porteurs du variant CCR5-Δ32 étant alors

plus résistants que les autres à la maladie.

1.4.4 L'évolution des gènes à ARN

A l'instar des séquences codantes, les séquences non-codantes fonctionnelles sont soumises à la pression de sélection. Pour les séquences d'ARN non-codants, la pression s'exerce à plusieurs niveaux : la conservation de la structure pour les régions appariées, et la conservation de la séquence pour les régions non appariées correspond à des sites d'interaction avec d'autres molécules.

Pour les bases appariées, la substitution d'une base par une autre peut avoir deux conséquences : soit l'appariement est préservé, soit il est rompu. Lorsque deux bases appariées sont individuellement substituées, mais que ces substitutions préservent l'appariement, on parle de mutations compensées, ou mutations compensatoires.

Toutefois, contrairement aux mutations silencieuses dans les régions codantes qui n'ont aucun effet sur la protéine codée, les substitutions qui préservent les appariements modifient la stabilité locale et globale de la molécule. Par exemple, pour deux bases appariées `G` et `C`, la transition de `C` en `T` peut préserver l'appariement puisque `G` et `T` peuvent s'apparier mais ce nouvel appariement est moins stable que l'appariement original. Au delà de la stabilité des structures d'ARN c'est la conformation spatiale locale et globale de la structure de l'ARN qui est soumise à la pression de sélection. Au voisinage de certains motifs structuraux, certains nucléotides, appariés ou non, sont en effet exposés d'une manière particulière. La conformation de ces nucléotides est le plus souvent primordiale à la fonction de la molécule. Les séquences de ces motifs structuraux, au même titre que les régions non appariées, sont donc contraintes d'être conservées.

1.5 L'analyse comparative de séquences nucléiques

Au cours de l'évolution, les séquences nucléiques subissent des transformations et des remaniements à différentes échelles. L'étude *a posteriori* de ces événements évolutifs fait appel à l'analyse comparative de séquences dont le principe est d'extraire de l'information des ressemblances et des différences observables entre plusieurs séquences. L'analyse comparative est largement plébiscitée pour établir des phylogénies, c'est-à-dire le parcours évolutif et les liens de parenté entre des espèces, pour transférer des informations connues d'une espèce à une autre ou encore pour inférer de l'information à un ensemble de séquences supposées partager une fonction commune.

L'analyse comparative de séquences s'appuie presque systématiquement sur un *alignement* des séquences à analyser. Dans la section 1.5.1, nous introduisons les bases en matière d'alignement de séquences nécessaires à la bonne compréhension de la suite de cet ouvrage. Par la suite, nous nous intéressons plus spécifiquement à l'utilisation d'alignements dans le cadre d'une analyse comparative de séquences codantes et de séquences partageant une structure commune (section 1.5.2). Enfin, nous présentons dans la section 1.5.3 un nouvel objet pour l'analyse comparative, les méta-séquences, que nous avons introduit dans le but d'améliorer l'analyse comparative de séquences hétérogènes en terme de conservation.

1.5.1 L'alignement de séquences comme support de l'analyse comparative

L'alignement de séquences consiste à établir une correspondance maximale entre les éléments qui les composent. Les algorithmes à même d'accomplir cette tâche de manière optimale sont issus de la communauté de l'algorithmique du texte. Ces algorithmes ne sont pas l'objet de cette section centrée sur l'utilisation des alignements dans le cadre de l'analyse comparative. Pour plus de détails sur ces algorithmes, on pourra se référer au chapitre 4 où plusieurs sections leurs sont consacrées.

La représentation la plus courante d'un alignement, quelque soit la méthode de construction utilisée, est une matrice où les bases alignées sont empilées et les insertions/délétions marquées par un tiret. Parfois, les séquences alignées sont séparées par des symboles qui facilitent la lecture de l'alignement : l'identité entre deux éléments est marquée par une barre verticale, la substitution d'un élément par un autre est marquée par un point. La figure 1.16 présente un alignement semi-global de deux séquences nucléiques.

```
AAN33049          1 CGAATGCCAGGCCCAGCCCTCA---CCTCTCGCTCCGCAGGGGGGAGTCG     47
                      |||        ||||..||   ..|||||||||        ||.||
AAA31576          1    ATG--------AGCCGGCAGAGTATCTCGCTCC--------GATTC-     30

AAN33049         48 CCTGCACCGGTGGCCGCTGCTCCTGCTGCTGCTGCTGCTGC-TCCC----     92
                                ||||||||.||.|||||||||..||  | ||||
AAA31576         31 -------------CCGCTGCTTCTCCTGCTGCTGTCGC--CATCCCCCGT     65

AAN33049         93 -------GCCGCCCCCGGTCCTGCCCGCG-----GAAGCC              120
                          ||.|.|||||..| |||||||     |||.||
AAA31576         66 CTTCTCAGCGGACCCCGGGGC-GCCCGCGCCAGTGAACCCCTGCTGTTAC    114
```

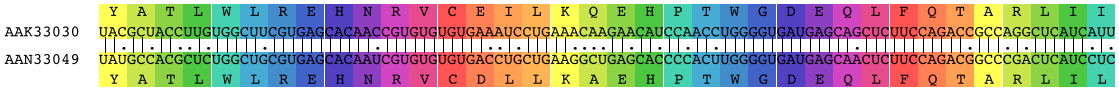
FIG. 1.16 – Exemple d'alignement semi-global entre deux fragments de séquences homologues de gènes codants pour la prostaglandine dont le pourcentage d'identité est de 44,9%.

Toute une terminologie permet de décrire les ressemblances entre séquences selon différents points de vue : la similarité, l'identité et l'homologie. L'identité désigne la proportion de nucléotides ou d'acides aminés identiques entre deux séquences. Elle est souvent exprimée en pourcentage et s'obtient en calculant le ratio entre le nombre de nucléotides ou d'acides aminés identiques et la longueur de l'alignement. La similarité désigne la proportion de substitutions, identités incluses, entre deux séquences alignées par rapport à la longueur de l'alignement. L'homologie a une connotation évolutive : deux séquences sont dites homologues si elles sont issues d'un même ancêtre commun et partagent une même fonction. La similarité est un indicateur d'homologie : on considère qu'une similarité significative est signe d'homologie. L'inverse n'est cependant pas vrai : une absence de similarité significative entre deux séquences n'implique pas nécessairement que ces séquences ne soient pas homologues.

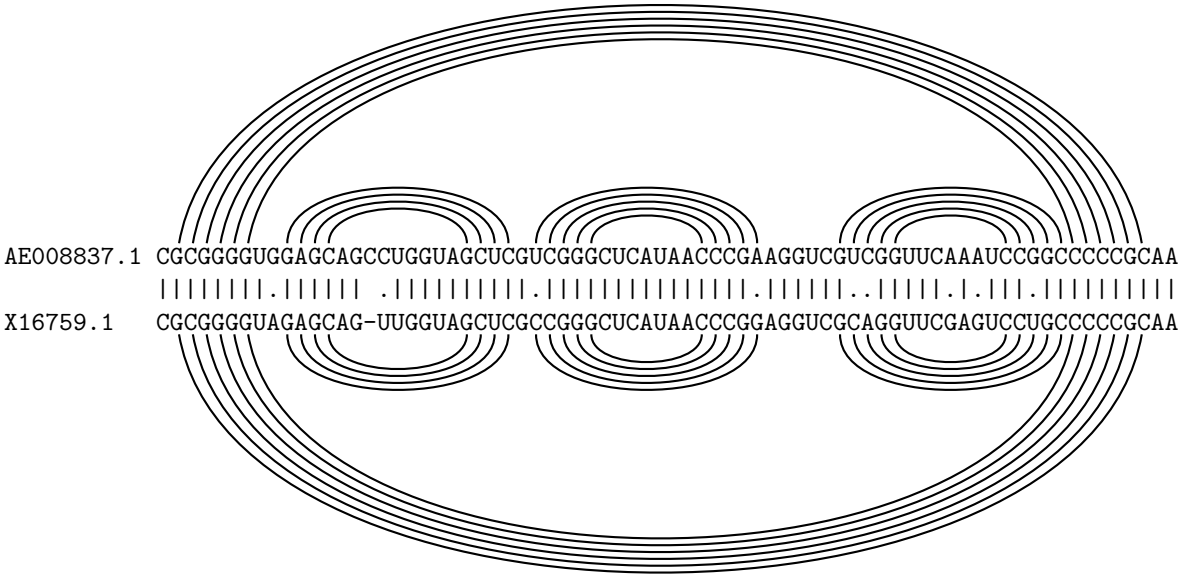
Construire un alignement de deux ou plusieurs séquences est toujours possible. Sans connaissance *a priori* sur la nature des séquences, l'exactitude d'un alignement est variable en fonction du degré de similarité des séquences à aligner : plus les séquences sont similaires, meilleur sera leur alignement, et inversement pour des séquences divergentes. Pour que les résultats apportés par une analyse comparative de séquences menée à partir d'un alignement fassent du sens, il est nécessaire que l'alignement employé soit fiable.

1.5.2 L'analyse de séquences codantes et de séquences structurées

L'analyse comparative de séquences nucléiques est de plus en plus employée à des fins prédictives, par exemple pour déterminer si des séquences sont des séquences codantes homologues, ou si elles partagent une structure commune. Sans connaissance *a priori* sur la fonction des séquences, les aligner semble un bon point de départ. Sur l'alignement, on devrait voir apparaître des mutations dont l'analyse permettra de déterminer si elles sont ou non corrélées à la conservation d'une fonction particulière. Pour illustrer notre propos, nous avons produit deux alignements de séquences homologues reportés sur la figure 1.17.



(a) Alignement correct de deux fragments de séquences codant pour une prostaglandine dont le pourcentage d'identité est 77,8%.



(b) Alignement correct de deux séquences d'ARN de transfert dont le pourcentage d'identité est de 86,5%.

FIG. 1.17 – Deux alignements semi-globaux optimaux corrects de séquences codantes homologues (a) et de séquences partageant une structure commune (b).

La sous-figure (a) de la figure 1.17 présente un alignement de deux fragments de séquences codantes homologues. Les séquences d'acides aminés codées par chacune des séquences sont également reportées. Sur cet alignement, on peut clairement voir apparaître les mutations silencieuses et synonymes entre les codons car les bases de chaque séquence sont ici correctement alignées avec leurs homologues dans l'autre séquence. Sans connaissance *a priori* de la nature de ces séquences nucléiques ni des séquences d'acides aminés qu'elles codent, cet alignement constitue donc un support fiable pour une analyse comparative dont le but est de prédire la séquence conservée d'acides aminés.

La sous-figure (b) de la figure 1.17 présente un alignement de séquences homologues d'ARN de transfert. Les structures secondaires individuelles de chaque séquence sont également reportées sous forme arc-annotée. Sur cet alignement se produit le phénomène analogue à celui

précédemment observé sur l'alignement des séquences codantes : les bases appariées de chaque séquence sont bien alignées avec leurs homologues dans l'autre séquence. Les mutations qui préservent les appariements, compensées ou non, sont ainsi clairement révélées. Sans connaissance *a priori* de la nature de ces séquences ni de leurs structures secondaires, cet alignement constitue donc un support de qualité pour une analyse comparative dont le but est de prédire la structure conservée.

1.5.3 Mise en œuvre bio-informatique

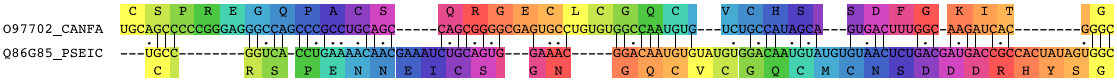
Sans en apporter la preuve ici, le raisonnement appliqué précédemment sur des alignements de deux séquences reste valable pour des alignements de plus deux séquences. Intuitivement, plus on dispose de séquences homologues à comparer, plus on détient d'information à même de servir l'analyse comparative, quel qu'en soit l'objectif. Traditionnellement, une analyse comparative portant sur un ensemble de plusieurs séquences s'appuie sur un alignement multiple de ces séquences. Selon le degré de conservation, l'alignement multiple n'est cependant pas toujours un support pertinent pour analyser un ensemble de séquences.

Degré de conservation et alignement

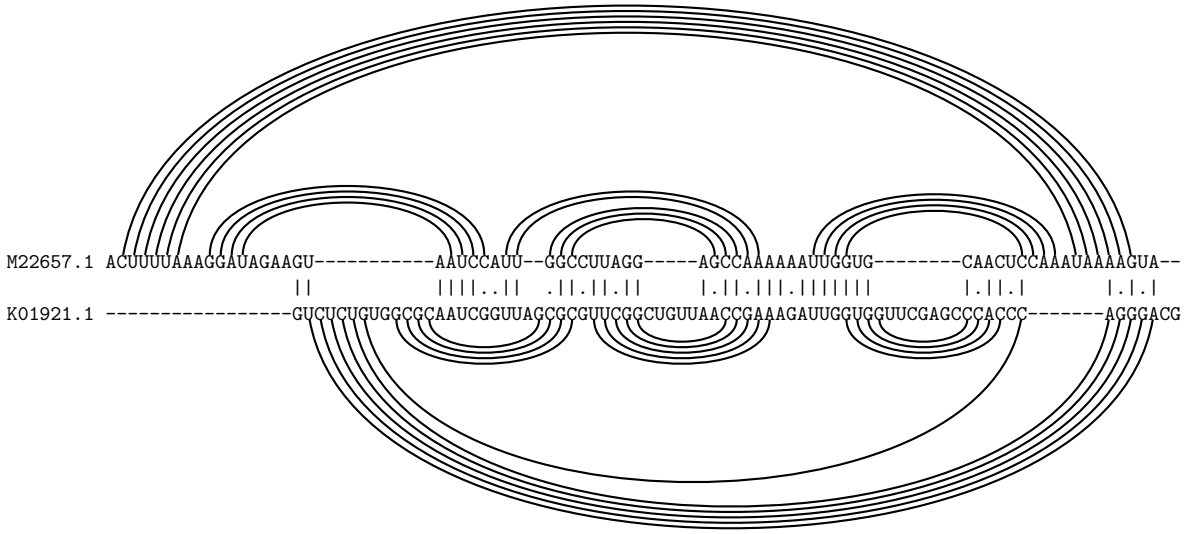
Les exemples présentés dans la section 1.5.2 impliquent des couples de séquences plutôt bien conservées. En effet, leurs pourcentages d'identité sont supérieurs à 75%, ce qui signifie que plus de trois quarts de leurs nucléotides sont identiques. Le degré de conservation de ces deux couples de séquences est suffisamment élevé pour pouvoir les aligner correctement sans connaissance supplémentaire de leurs fonctions communes. Plus le pourcentage d'identité entre deux séquences homologues est faible, plus il devient difficile de les aligner sans en connaître la nature.

La figure 1.18 présente deux alignements optimaux de couples de séquences homologues faiblement conservées. Sur l'alignement (a) de séquences codantes homologues, on peut remarquer qu'aucun nucléotide de la première séquence n'est correctement aligné avec son homologue dans la seconde séquence, à l'exception des trois derniers. Cet alignement n'est donc pas correct car il ne révèle aucune des mutations silencieuses et synonymes qui existent pourtant entre ces séquences quant on connaît les séquences d'acides aminés réellement codées. De plus, on peut également remarquer que les insertions et délétions introduites durant l'alignement sont décalantes et interviennent au beau milieu des codons véritablement traduits. Sur l'alignement (b) de séquences homologues d'ARN de transfert, aucune paire de bases appariées dans une séquence n'est alignée avec son homologue dans l'autre séquence. Cet alignement ne laisse donc pas apparaître les mutations qui préservent les appariements des deux structures secondaires identiques de ces séquences.

Ces simples exemples sur des couples de séquences montrent qu'un alignement n'est un support fiable pour une analyse comparative que si les séquences présentent un degré de conservation suffisamment élevé pour être alignées correctement sans information supplémentaire sur leur nature. Néanmoins, sur des ensembles de plus de deux séquences il n'est pas nécessaire d'adopter un point de vue binaire en ce qui concerne l'utilisation ou non d'un alignement multiple. Au sein d'un ensemble de séquences, tous les couples ne présentent en effet pas nécessairement le même degré de conservation.



(a) Alignement incorrect de deux fragments de séquences codant pour une intégrine dont le pourcentage d'identité est 37,8%.



(b) Alignement erroné de deux séquences d'ARN de transfert dont le pourcentage d'identité est de 34,7%.

FIG. 1.18 – Deux alignements semi-globaux optimaux incorrects de séquences codantes homologues (a) et de séquences partageant une structure commune (b).

Les méta-séquences

Afin d'analyser convenablement les ensembles de séquences hétérogènes en terme de conservation, nous proposons de créer des sous-ensembles de séquences suffisamment similaires pour pouvoir les aligner de manière fiable. Pour la construction des sous-ensembles de séquences similaires, nous définissons une représentation des données sous forme de graphe. Soit $\mathcal{S} = \{s_1, \ldots, s_n\}$ un ensemble de n séquences nucléiques et $\texttt{id}(s_i, s_j)$ le pourcentage d'identité entre deux séquences s_i et s_j. Nous créons ensuite une partition $\mathcal{P}$ de $\mathcal{S}$ comportant m parties $\{P_1, \ldots, P_m\}$. Chaque partie P_k est composée d'un sous-ensemble de séquences de $\mathcal{S}$ connectées par la relation de similarité $\texttt{id}(s_i, s_j) \geq \alpha$, où α est un seuil sur le pourcentage d'identité au delà duquel on considère que les séquences présentent un degré de similarité suffisant. Chaque partie de $\mathcal{P}$ est appelée une méta-séquence. Une méta-séquence fait ainsi référence soit à une seule séquence, soit à un ensemble de plusieurs séquences qui seront représentées par leur alignement. La figure 1.19 présente un exemple de quatre séquences regroupées en deux méta-séquences, symbolisées par les parties grisées. Sur cet exemple, le couple de séquences s_1 et s_2 d'une part, et le couple s_1 et s_3 d'autre part, présentent un pourcentage d'identité supérieur à α. Les séquences s_1, s_2 et s_3 sont donc regroupées pour former la partie P_1, et la séquence s_4 forme à elle seule la partie P_2.

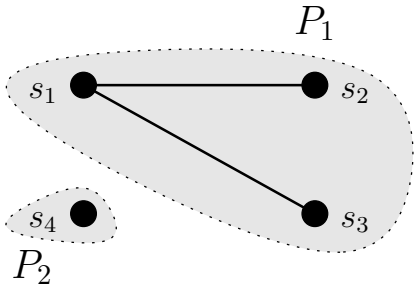


FIG. 1.19 – Exemple de création de deux méta-séquences à partir de quatre séquences. Chaque sommet noir correspond à une séquence, les arêtes relient les séquences dont le pourcentage d'identité est supérieur au seuil α. Les régions grisées correspondent aux méta-séquences correspondantes.

A l'issue de ce processus, l'ensemble des séquences originales est partitionné en plusieurs méta-séquences. La comparaison des méta-séquences entre elles dépend de l'analyse comparative à produire. Dans la suite de ce manuscrit, nous présentons deux utilisations des méta-séquences pour la prédiction par analyse comparative de séquences codantes homologues (chapitre 2) et de structures secondaires communes (chapitre 3).

Chapitre 2

Recherche de gènes et régions codantes

Dans ce chapitre, nous abordons le problème de l'identification de régions codantes. Historiquement, c'est l'un des premiers problèmes sur lequel s'est penchée la communauté bioinformatique. Ce problème constitue une partie importante de l'annotation structurale des génomes : où se trouvent les gènes codant pour des protéines ? Quelle est leur structure ? Pour répondre à ces questions, les informations utilisées peuvent être de nature différente : la présence de signaux qui balisent et concourent à la structure du gène et à son expression, le contenu de la séquence codante qui peut comporter des biais de composition, ou enfin la similarité avec d'autres molécules connues. Nous regroupons ces informations en deux groupes : les informations intrinsèques, c'est-à-dire les informations contenues dans la séquence nucléique considérée, et les informations extrinsèques, c'est-à-dire les informations obtenues par comparaison de la séquence nucléique d'intérêt avec des séquences déjà connues. Cela donne lieu à deux types d'approche de prédiction : les approches *ab initio* et les approches par homologie de séquences. Un bon nombre d'approches ne se contentent pas d'une seule séquence mais travaillent sur un alignement de deux ou plusieurs séquences. Ces méthodes exploitent l'information évolutive entre les séquences pour détecter un schéma spécifique de mutations qui pourrait trahir la présence d'une contrainte fonctionnelle codante (sections 1.4.3 et 1.5). Nous les regroupons sous le terme générique d'analyse comparative.

Le contenu de ce chapitre est le suivant. Nous commençons par dresser un état de l'art des principales méthodes de prédiction *ab initio* en section 2.1, par homologie en section 2.2 et par analyse comparative en section 2.3. Dans la section 2.4, nous présentons ensuite notre contribution au problème, avec la méthode Protea [FT07, FT09] qui s'inscrit dans le cadre de l'analyse comparative. La section 2.5 est consacrée à l'exposé des résultats expérimentaux de Protea.

2.1 Les méthodes *ab initio*

Les approches *ab initio* ont pour objectif de prédire l'ensemble des gènes présents dans une séquence nucléique sans autre connaissance extérieure. Pour cela, elles tirent parti des signaux présents dans la séquence et des biais de composition des séquences codantes.

2.1.1 Le cadre ouvert de lecture

Le premier signal qui peut être exploité provient simplement des bornes des gène, en particulier pour les organismes procaryotes. Nous avons vu dans le chapitre précédent (section 1.2) qu'un élément essentiel d'un gène codant est son cadre ouvert de lecture, débutant par un codon START suivi d'un enchaînement ininterrompu de codons et terminé par un codon STOP. Cette information est souvent suffisante pour identifier un bonne partie des gènes au niveau génomique quand le cadre ouvert de lecture est significativement long. L'absence de codon STOP peut cependant être statistiquement peu significative pour des séquences courtes ou lorsque la fréquence de ces codons est localement fortement réduite. A cause de l'épissage chez les eucaryotes et certaines bactéries, rechercher au niveau génomique un codon STOP après un codon START dans le même cadre de lecture n'a pas de sens à cause de la présence des introns [Fic95].

2.1.2 Les autres signaux liés à la structure du gène

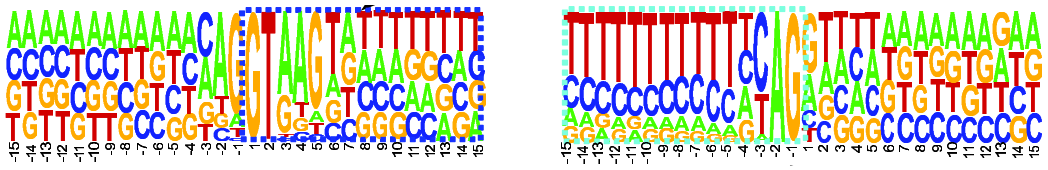
Des signaux plus fins que la détection d'un cadre ouvert de lecture peuvent être utilisés pour identifier les séquences codantes. Nous avons vu que toute une batterie de signaux balisent et concourent à la structuration d'un gène, que celui-ci soit d'origine eucaryote ou procaryote (section 1.2). Ces signaux interviennent dans les processus mis en œuvre lors de l'expression du gène. Les régions en amont et en aval de la région codante des gènes contiennent ainsi des sites de fixation de facteurs de transcription, le site d'initiation de la transcription, le signal de poly-adénylation, ... Chez les eucaryotes, d'autres signaux sont présents dans la région codante des gènes et servent à guider l'épissage des introns.

La grande majorité de ces signaux sont des motifs relativement courts, dont la longueur est inférieure à une vingtaine de nucléotides. Ce sont des motifs approchés, établis à partir d'un certain nombre d'observations. Ils doivent donc être décrits au moyen d'une représentation plus flexible qu'une simple séquence. Cela pose la question du modèle choisi pour les représenter.

L'une des premières représentations adoptée est la matrice poids-position, ou PWM pour *Position Weight Matrix*. La PWM associée à un motif contient pour chaque position du motif la probabilité d'apparition de chacun des quatre nucléotides. Sur la figure 2.1 sont représentées sous forme graphique deux PWM qui correspondent à des sites d'épissage.

La représentation par PWM est parfois insuffisante pour décrire certains signaux car elle ne capture aucune relation entre les positions d'un motif : les positions sont considérées comme indépendantes. Une représentation plus riche a donc été proposée : le modèle poids-tableau, également appelé WAM pour *Weight Array Models*. Les WAM s'appuient sur un théorie fréquemment utilisée dans l'analyse de séquences biologiques : les modèles de Markov [DEKM99].

Dans le cadre de la découverte de motifs dans une séquence nucléotidique, un modèle de Markov d'ordre k peut être succinctement défini comme un processus stochastique dans lequel la probabilité d'occurrence d'un nucléotide à une position donnée dépend uniquement des k positions précédentes. En pratique, un WAM qui décrit un motif de longueur k est modélisé par un modèle de Markov k-périodique d'ordre $k-1$, c'est-à-dire un ensemble de k modèles de Markov d'ordre $k-1$. Construire un tel modèle revient à déterminer toutes les probabilités conditionnelles du modèle. Par exemple, pour un modèle d'ordre 1 : quelle est la probabilité d'observer un `A` à une position donnée sachant que le nucléotide précédent est un `C` ? Pour



Source http://www.pnas.org/cgi/doi/10.1073/pnas.0703773104

FIG. 2.1 – Représentation graphique de deux PWM décrivant les sites d'épissage. La fréquence de chaque nucléotide à une position est proportionnelle à la taille dans la lettre qui le représente. Sur la représentation du site donneur, à gauche, on voit clairement apparaître une séquence consensus qui caractérise le début de l'intron `GT`. On remarque le même phénomène pour le site accepteur qui marque la fin des introns et qui est caractérisé par la séquence consensus `AG`.

obtenir ces probabilités, on procède par comptage sur un ensemble d'apprentissage constitué de séquences contenant le motif à caractériser. Cet ensemble doit contenir suffisamment de séquences pour que les observations réalisées fassent du sens. De plus, comme la forme des signaux peut être propre à chaque espèce, il est nécessaire de disposer des séquences de même provenance, ou dont on est sûr qu'elles sont biaisées de la même manière. Un modèle construit sur une espèce ne pourra donc pas être directement appliqué à une autre espèce. Toutefois, certains signaux comme les signaux liés à la transcription sont relativement bien conservés entre les espèces. Un modèle descriptif d'un signal peut donc sous certaines conditions être transféré d'une espèce à une autre. Plus l'ordre du modèle de Markov est important, plus la quantité de données d'apprentissage nécessaire est importante. Etant donné que quatre nucléotides différents sont possibles à une position donnée, il est nécessaire de déterminer 4^{k+1} probabilités pour un modèle d'ordre k. Une fois le modèle de Markov construit, son exploitation est assez simple : étant donnés une séquence et un modèle de Markov, on peut calculer la probabilité que la séquence ait été générée selon ce modèle avec, par exemple, l'algorithme de Viterbi [Vit67]. A l'heure actuelle, les WAM sont très largement utilisés pour décrire et détecter les sites d'épissage [TMM07, Sto90, Gel95].

2.1.3 Les biais de composition de la séquence codante

À côté des signaux liés à la structure d'un gène, on peut utiliser la composition de la séquence codante elle-même [SM82]. De façon générale, les régions codantes ont des asymétries et périodicités qui facilitent leur distinction des autres régions [GGG80]. Ces caractéristiques sont propres à chaque espèce. Les premières analyses de régions codantes ont révélé chez certaines bactéries un biais dans l'usage des codons [GG82]. Du fait de la redondance du code génétique, des codons différents codent pour le même acide aminé. Pour coder un acide aminé, plusieurs organismes affichent une préférence marquée pour un ou plusieurs codons. Pour un organisme, cette préférence est en partie corrélée à l'abondance de copies des gènes d'ARN de transfert correspondants dans son génome [Ike81b, Ike81a, Ike82]. Plusieurs autres mesures ont par la suite été testées pour caractériser les régions codantes : la fréquence d'apparition des nucléotides, des hexamers, la périodicité d'occurrence des nucléotides, ... L'étude menée par Fickett [FT92] a permis de mettre en évidence que la fréquence d'apparition des hexamers est la mesure qui discrimine le mieux les régions codantes.

La méthode la plus couramment employée pour modéliser les biais de composition en k-mers des régions codantes est un modèle de Markov 3-périodique d'ordre $k-1$. Ce type de modèle capturent simultanément plusieurs mesures "intuitives" : un modèle 3-périodique d'ordre 5 capture le biais d'usage des codons mais aussi les dépendances entre les paires de codons successifs.

L'utilisation de modèles de Markov ne se limite pas qu'à la caractérisation de régions codantes. Il est ainsi possible de construire des modèles pour les régions introniques, les régions non traduites ou encore les régions intergéniques. Pour ces régions, les modèles utilisés sont plus simples car l'information à capturer est moindre : les modèles utilisés sont non périodiques et d'ordre inférieur à 2. De plus, il est également fréquent de considérer toutes les séquences qui ne codent pas pour des protéines simultanément et donc de ne construire qu'un seul modèle de Markov pour les caractériser.

2.1.4 Les mises en œuvre logicielles

Les informations disponibles dans la séquence sont donc multiples, et complémentaires. Pour les intégrer, chaque méthode *ab initio* adopte une modélisation plus ou moins fine des éléments constitutifs d'un gène qui dépend directement de la nature et de la quantité des signaux pris en compte. Les modélisations les plus simples ne discernent que deux types de séquences, codantes ou non, tandis que les plus élaborées identifient également les extrémités transcrites mais non traduites des gènes, les régions introniques, exoniques, le site d'initiation de la transcription, le site de poly-adénylation, ... Plus la modélisation est précise, plus la segmentation de la séquence nucléique sera détaillée [RAG97]. Deux techniques majeures ont été déployées pour segmenter la séquence nucléique suivant la modélisation [Gui98] : des algorithmes *ad hoc* par programmation dynamique et les modèles de Markov cachés.

Les méthodes qui emploient la programmation dynamique pour obtenir une segmentation de la séquence nucléique se déroulent en deux temps [Sea92, DS94, SS95] : la détection dans la séquence des éléments constitutifs considérés de manière indépendante, puis la recherche d'un assemblage optimal de ces éléments par programmation dynamique qui respecte des contraintes de dépendances entre les éléments. En fait, un score est attribué à chaque élément, et un assemblage optimal correspond à une sélection d'éléments compatibles entre eux dont la somme des scores est maximale. Plusieurs méthodes *ab initio* se servent de cette technique, notamment Glimmer [DHK+99, SDKW98] pour les procaryotes, et GeneID [PBG00], Fgenes [SS00], Grail [XMU94, XU97], GlimmerM [DBPS07, SPD+99] et EuGène [SMR01] pour les eucaryotes. Les différences entre ces programmes se situent au niveau des éléments constitutifs considérés et la manière de les détecter. Parmi ces programmes, EuGène, Glimmer et GlimmerM sont les seuls à intégrer une modélisation du biais de composition dans les séquences codantes par modèle de Markov périodique. EuGène est le programme dont la modélisation est la plus détaillée : il distingue les exons, les introns, les régions intergéniques et les extrémités non traduites des gènes, et dispose d'un modèle de Markov pour reconnaître chacune de ces régions.

Bon nombre de méthodes s'appuient sur les modèles de Markov cachés, ou HMM pour *Hidden Markov Model*. Dans ce cadre, le modèle de Markov caché permet de segmenter la séquence en combinant différents modèles de Markov, propre à chaque élément constitutif du gène. La motivation principale à l'usage de cette technique est l'homogénéité apportée par la confusion entre la modélisation de la structure du gène, c'est-à-dire l'alternance de régions, et la caractérisation de ces régions. Par conséquent, pour toutes les méthodes à base

de HMM la détection des régions codantes et non codantes est confiée à des modèles Markoviens dont les caractéristiques varient selon la nature de la région modélisée. Historiquement, EcoParse [KMH94] est le premier programme à avoir intégré un modèle de Markov caché. D'autres ont suivi comme Genie [KHRE96], Fgenesh [SS00], Augustus [SW03, Sta03], Genie [KHRE96], GeneMark.hmm [LB98, LTHCB05] et Genscan [BK97]. Genscan et GeneMark.hmm sont les programmes les plus utilisés actuellement pour la richesse de leur modélisation de la structure des gènes et parce qu'ils ont été paramétrés sur plus d'une centaine d'organismes.

2.2 Les approches par homologie de séquence

Les méthodes par homologie de séquences utilisent comme première et principale source d'information la similarité avec des séquences connues et déjà annotées. L'hypothèse de travail est la suivante : deux séquences significativement similaires ont généralement une fonction identique ou proche (section 1.4.3). En effet, durant l'évolution, les séquences fonctionnelles, *a fortiori* les séquences codantes, sont soumises à une contrainte fonctionnelle. Sous cette contrainte, les séquences codantes tendent à être plus conservées entre les espèces que les séquences non fonctionnelles. D'après cette assertion, les informations disponibles à propos d'une séquence connue peuvent être transférées à toute séquence significativement similaire. Au moins trois types de séquences sont susceptibles d'apporter de l'information pour détecter des régions codantes dans une séquence nucléique : des séquences de protéines, des séquences d'ARN ou d'EST et des séquences génomiques annotées.

2.2.1 Similarité avec des séquences peptidiques

La manière la plus simple et la plus utilisée pour déterminer si une séquence nucléique est codante est de chercher à identifier la protéine qu'elle code, si celle-ci est connue, ou une protéine homologue dans le cas contraire.

Les banques de données les plus utilisées pour effectuer ce type de recherche sont SwissProt [BAW+05] et Pir [BGH+98], car elles contiennent exclusivement des séquences de protéines vérifiées expérimentalement. En complément, d'autres banques proposent des séquences d'acides aminés issues de la traduction automatique de séquences codantes. TrEMBL [HBB+02, BAB+04] contient ainsi uniquement les traductions automatiques des séquences annotées codantes du projet Ensembl où toutes les séquences déjà présentes dans SwissProt sont exclues. Afin de simplifier les recherches, UniProt [Con08] réunit les banques SwissProt et TrEMBL. Au même titre qu'UniProt, RefSeq [WBB+08] propose des séquences vérifiées expérimentalement et des séquences issues de l'annotation automatique basée sur les données et l'expertise des membres du NCBI.

Pour effectuer la recherche dans toutes ces banques, on a recours en première approche à des algorithmes d'alignement deux à deux, tels que Fasta et Blast. Il existe plusieurs déclinaisons de ces méthodes qui permettent de fouiller une banque de séquences protéiques à partir d'une séquence nucléique, telles que FastX, FastY et BlastX. L'idée est de comparer indépendamment les six traductions potentielles d'une séquence nucléique contre toutes les séquences d'acides aminés contenues d'une banque. FastX et FastY diffèrent dans leur manière de traiter les éventuels décalages du cadre de lecture : FastX ne prend en compte que les décalages positifs de cadre de lecture (+1 ou +2), tandis que FastY ne considère

que les décalages négatifs (-1 ou -2). Contrairement à ses homologues, BlastX ne gère pas les décalages potentiels du cadre de lecture. Enfin, il existe une version enrichie de BlastX, nommée BlastC [SG94], qui prend en compte dans son système de score les éventuels biais d'usage des codons connus et avérés. On estime qu'environ 50% des gènes peuvent être identifiés à partir des séquences présentes dans SwissProt et Pir [MSSR02] à l'aide de ces outils.

Même lorsque l'on dispose d'une protéine similaire, il est difficile de déterminer la structure complète d'un gène, particulièrement les bornes des extrémités $5'$ et $3'$ non traduites. Qui plus est, l'identification précise de la séquence codante peut également s'avérer incomplète : tous les domaines protéiques ne sont pas nécessairement partagés, en cas d'épissage alternatif notamment, et certains exons peuvent ne pas être attrapés à cause de leur petite taille.

La recherche de similarité peut ensuite donner lieu à des traitement plus sophistiqués. Historiquement, Procrustes [GMP96] est la première méthode de prédiction de gènes à exploiter la similarité de séquences au niveau peptidique. Etant donnée une protéine similaire à la séquence nucléique d'intérêt, trouvée manuellement par FastX ou BlastX par exemple, la méthode déployée par Procrustes consiste à aligner la séquence nucléique et la séquence protéique. Dans un premier temps, Procrustes identifie tous les exons potentiels en recherchant les bornes exons/introns selon leurs séquences consensus strictes. Les exons potentiels sont ensuite traduits puis alignés sur la séquence peptidique fournie. Enfin, Procrustes assemble par programmation dynamique un enchaînement d'exons qui maximise le score de similarité avec la protéine tout en respectant la structure minimale du gène, c'est-à-dire une séquence codante qui commence par un codon START, ne contient pas de codon STOP dans le cadre de lecture et se termine par un codon STOP. GeneWise [BCD04] ou encore GenomeScan [YLB01] effectuent la même tâche que Procrustes. Ces derniers sont globalement plus tolérants que Procrustes : ils n'interdisent pas strictement les décalages de cadres de lecture ainsi que les codons STOP dans le cadre de lecture mais ces deux éléments sont fortement pénalisés. De plus, les sites d'épissage dans GeneWise et GenomeScan sont détectés à l'aide d'un HMM, fortement inspiré de celui de Genscan. Dans GenomeThreader, les sites d'épissage sont décris par un représentation *ad hoc* équivalente aux HMM précédents. La différence entre ces trois programmes réside essentiellement dans l'évaluation de la similarité entre les séquences : GenomeScan utilise les résultats de BlastX, GeneWise évalue leur similarité à l'aide d'un modèle pair-Markov caché (pair-HMM), et GenomeThreader utilise un algorithme d'alignement local par programmation dynamique. ORFGene2 [RMK96] et PredictGenes [GHKB00] sont deux méthodes équivalentes à GeneWise dans la modélisation adoptée. Toutefois elles intègrent la recherche de protéines similaires en interrogeant la banque SwissProt. Toutes ces méthodes sont destinées à la prédiction de gènes eucaryotes, en majorité entraînées et testées chez l'Homme, la souris ou des plantes. Elle ont été conçues pour travailler sur des séquences provenant d'organismes proches. Elles fournissent d'ailleurs d'excellents résultats sur des séquences très conservées, mais leurs performances se révèlent moyennes sur des séquences relativement peu conservées [TPP99].

2.2.2 Similarité avec des séquences transcrites

Le second type de séquences auquel on peut faire appel pour identifier des régions codantes sont les séquences d'ARN matures, ou des fragments d'ARN matures. Pour des raisons expérimentales de séquençage, la majorité des séquences d'ARN présentes dans les banques de données comme RefSeq [WBB+08] ou dbEST [BLT93] se trouvent sous forme d'ADN

complémentaires. Ces ADN complémentaires, notés ADNc, sont obtenus par transcription inverse d'ARN matures. Il existent plusieurs protocoles pour obtenir les séquences d'ADNc rétro-transcrits. Le séquençage "classique" d'un ADNc permet d'en obtenir la séquence complète et ce de manière fiable. Avant que soit mis au point ce protocole, le séquençage des ARN messagers se faisait par un protocole à haut débit moins fiable. Ce protocole consiste à séquencer quelques centaines de nucléotides en une seule fois à chaque extrémité d'un ADNc. Ces fragments nommés des EST, acronyme pour *Expressed Sequence Tags*, ne représentent donc qu'une information partielle par rapport à la taille de certains ADNc qui peuvent atteindre plusieurs milliers de nucléotides. La figure 2.2 illustre de manière schématique des séquences d'EST obtenues pour un ARN messager mature.

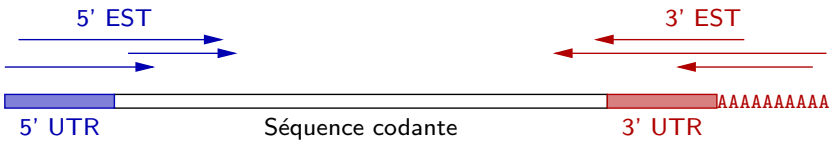


FIG. 2.2 – Les EST sont issus du séquençage partiel des extrémités d'un ARN mature, ici un ARN messager.

Les ADNc et les EST représentent les informations les plus pertinentes dont on peut disposer pour établir la structure précise des gènes surtout s'ils sont issus du même organisme que la séquence nucléique à annoter [FSY+99]. En effet, comme les ADNc proviennent d'ARN transcrits, ils contiennent en plus de la séquence codante les extrémités 5′ et 3′ non traduites. Les données d'ADNc et d'EST disponibles dans les banques dépendent des conditions expérimentales dans lesquelles les ARN ont été extraits : dans quel type de tissus ? A quel stade de développement ? ... Les données disponibles pour un organisme ne couvrent donc qu'une partie de son transcriptome, elles ne sont pas nécessairement représentatives de tous ses gènes. De plus, bien que les ADNc et les EST correspondent à des séquences transcrites, ces séquences ne sont pas obligatoirement traduites en protéines. Ces caractéristiques des ADNc et des EST en font des indices précieux pour confirmer des prédictions effectuées par ailleurs, notamment pour déterminer la structure des gènes. A ce titre, les ADNc sont, par nature, une source d'information plus complète que les EST qui ne représentent qu'une source d'information partielle.

Les banques de données d'EST sont fortement redondantes du fait de la technique employée pour les obtenir. Pour traiter ce problème de redondance, EBEST [JJ98] et PAGAN [KS01] procèdent à un regroupement des EST trouvés dans les banques. Les EST chevauchants sont regroupés puis un EST représentatif de chaque groupe est sélectionné. Les EST représentatifs sont enfin réalignés avec la séquence nucléique. Les séquences d'EST comportent un certain nombre d'erreurs introduites durant leur séquençage. La procédure d'alignement est donc paramétrée pour tolérer les substitutions, les insertions et les délétions d'un nucléotide.

Quelques méthodes se servent des séquences d'ADN complémentaires présentes dans les banques de données. Ces programmes, tels que AAT [HAZK97] ou SIM4 [FHZ+98] interrogent les banques de données d'ADNc à la recherche d'ADNc similaires puis réalignent les ADNc trouvés avec la séquence d'intérêt afin de localiser plus précisément les exons. Lorsque l'on dispose d'ADNc complémentaires provenant du même organisme que la séquence à annoter, l'identification des exons peut se révéler particulièrement fiable et précise [FSY+99].

2.2.3 Séquences génomiques

La similarité avec des séquences génomiques peut également permettre l'identification de régions codantes, même si ces génomes ne sont pas annotés. L'idée est que sous la pression de sélection (section 1.4) les séquences codantes présentent un niveau de conservation plus élevé que les régions non fonctionnelles. Plusieurs protocoles de recherche peuvent être envisagés : une comparaison intra-génomique à la recherche de séquences paralogues, c'est-à-dire de séquences homologues au sein du même génome, ou une comparaison inter-génomique pour trouver des séquences orthologues, c'est-à-dire des séquences homologues chez d'autres organismes.

Les comparaisons de séquences peuvent être réalisées au niveau nucléotidique ou au niveau peptidique, en traduisant "à la volée" selon les six cadres de lecture possibles. Quelque soit le niveau de comparaison utilisé, l'exploitation des résultats est relativement plus laborieuse que la comparaison avec des banques de protéines ou d'ARN messagers. En effet, les résultats sont ici bruités par la présence d'autres types de séquences conservées dans les génomes que des séquences codantes : des séquences non codantes fonctionnelles telles que des gènes à ARN non-codants ou des éléments répétés, des séquences régulatrices, ... De plus, la détection d'une ou plusieurs séquences significativement similaires à une séquence nucléique d'intérêt dépend essentiellement des génomes utilisés pour les comparaisons. Dans les faits, les résultats que l'on peut espérer obtenir de comparaisons inter-génomiques varient selon les distances évolutives qui séparent les organismes dont les séquences sont comparées. Entre deux espèces distantes, il est plus facile de discriminer les régions codantes car celles-ci seront significativement plus conservées que le reste des séquences génomiques. Inversement sur deux espèces proches dont les séquences génomiques complètes sont globalement ressemblantes, il est plus difficile de distinguer les régions plus conservées. Enfin, plus la distance évolutive entre les espèces comparées est élevée, plus la recherche de séquences similaires dépend de la sensibilité de la méthode d'alignement utilisée. Le choix de la méthode d'alignement et son paramétrage sont donc des critères importants qui influencent la qualité des résultats obtenus lorsqu'on compare des espèces séparées par une distance évolutive élevée.

2.3 Les approches par analyse comparative

Les approches de prédiction de gènes codants par analyse comparative travaillent sur des alignements de deux ou plusieurs séquences (section 1.5.2). L'originalité de ces méthodes est de caractériser des biais liés à l'évolution des séquences codantes observables entre un couple ou une famille de séquences. Ces biais peuvent être de plusieurs nature : la synténie, c'est-à-dire la conservation de l'ordre des gènes entre génomes, un biais de conservation de certaines régions ou encore la caractérisation d'un biais dans les mutations entre des séquences codantes homologues.

Syncod [RDM99] est la première méthode qui exploite réellement les biais de mutations entre des séquences codantes homologues. La méthode calcule un ratio entre les mutations silencieuses et les mutations faux-sens (section 1.4.3) observables entre deux cadres ouverts de lecture alignés avec Blast. Les séquences correspondantes sont identifiées comme des séquences codantes homologues si ce ratio est significativement plus élevé que ce qu'on pourrait observer par hasard sur des séquences ayant le même pourcentage d'identité. L'estimation du comportement attendu par hasard est ici confiée à une procédure de type Monte-Carlo. Qrna [RE01] s'appuie également sur un biais en mutations silencieuses et faux-sens entre les

séquences codantes. Cependant, QRNA emploie trois modèles de Markov cachés différents pour évaluer la probabilité de trois hypothèses : le modèle COD permet d'évaluer la probabilité que les séquences alignées soient des séquences codantes homologues, le modèle RNA évalue la probabilité que les séquences alignées soient des séquences non-codantes structurées homologues, et enfin le modèle OTH qui évalue la probabilité que les mutations entre les séquences soient fortuites. L'avantage certain de QRNA par rapport à SYNCOD est de distinguer dans son modèle COD les mutations faux-sens qui produisent des acides aminés aux propriétés physico-chimiques proches. Cette caractéristique lui permet de considérer des séquences séparées par une distance évolutive plus importante.

Alors que SYNCOD et QRNA travaillent sur des séquences sorties de leur contexte génomique, ROSETTA [BPM+00a, BPM+00b] propose un point de vue différent en travaillant sur un alignement complet de deux génomes. ROSETTA est cependant exclusivement paramétré et destiné à la comparaison des génomes de l'Homme et de la souris. ROSETTA utilise deux critères essentiels pour effectuer ses prédictions : la synténie d'une part, et la similarité au niveau nucléique et peptidique d'autre part. Les régions codantes des deux organismes sont en effet exceptionnellement bien conservées, approximativement 85% d'identité au niveau nucléique, en comparaison de leurs séquences introniques, environ 35% d'identité [MZB96, MB98, LMS+95, KH94]. D'autres méthodes telles que TWINSCAN [KFDB01], AGENDA [TRG+03], UTOPIA [BRS03], PRO-GEN [NGM01], CEM [BH00] et SGP-1 [WGJMOG01] fonctionnent de manière analogue à ROSETTA. Plus récemment, PROJECTOR [MD04] et GENEMAPPER [CP06] constituent un compromis entre ROSETTA et QRNA. Ces méthodes proposent une adaptation de la technique déployée dans QRNA en intégrant la synténie comme le fait ROSETTA pour la comparaison de deux génomes complets alignés.

Enfin, un dernier ensemble de méthodes travaille non plus sur deux séquences, mais sur un ensemble de séquences alignées. EXONIPHY [SH04] et EVOGENE [PH03] notamment étendent et affinent l'approche de QRNA. Pour cela elles requièrent en plus de l'alignement multiple un arbre phylogénétique correspondant aux séquences alignées. La connaissance des distances évolutives permet d'estimer avec plus précision les taux de mutations attendus entre séquences codantes, notamment le taux de mutations silencieuses. Ces deux méthodes utilisent des modèles phylogénétiques de Markov cachés. Le principe est identique à celui d'un modèle de Markov caché classique à la différence que les probabilités conditionnelles du modèle sont obtenues par des fonctions paramétrées par l'arbre phylogénétique. Dans le cadre de la comparaison de génomes complets alignés, N-SCAN [GB06] propose une adaptation de TWINSCAN pour traiter des alignements multiples. L'approche de N-SCAN, plus simple que celle de EXONIPHY et EVOGENE, ne requiert pas d'arbre phylogénétique mais est limitée à la prédiction de gènes pour l'Homme et la drosophile.

2.4 Protea

Dans cette section, nous présentons la méthode PROTEA, que nous avons développée pour l'identification de séquences codantes. Par rapport aux approches existantes que nous venons de présenter, PROTEA présente au moins deux caractéristiques qui font son originalité. En premier lieu, PROTEA peut traiter un nombre quelconque de séquences sans nécessiter d'alignement multiple préalable entre ces séquences. Disposer d'un alignement multiple correct est en effet une tâche délicate quand la distance évolutive entre les séquences est importante [Mar08], et peut se révéler une source d'erreurs comme nous l'avons mis en évidence

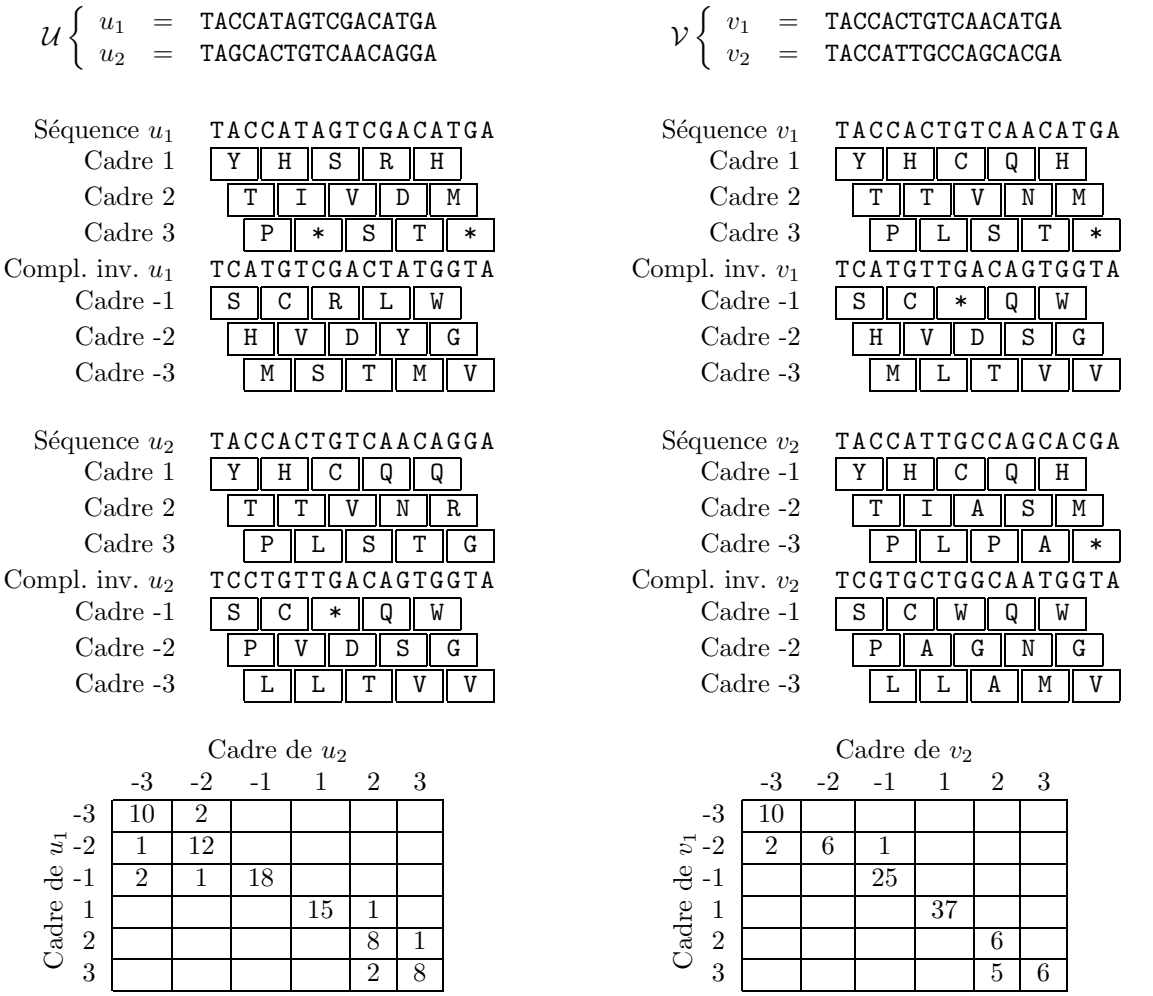
dans la section 1.5. En second lieu, PROTEA repose sur les principes de l'analyse comparative avec une idée générale qui à notre connaissance n'a jamais été utilisée : il s'agit d'exploiter le schéma évolutif observé entre séquences codantes à travers la conservation du cadre de lecture. Les séquences non codantes ne sont pas censées présenter ce schéma évolutif particulier.

Nous commençons par illustrer et valider cet argument pour un couple de séquences. Nous expliquons ensuite comment étendre ce principe à une famille comprenant un nombre quelconque de séquences, sans passer par un alignement multiple, mais en utilisant une structure de graphe.

2.4.1 Le modèle sur deux séquences

Nous avons vu que lors de l'évolution, une séquence codante fonctionnelle pouvait être altérée par un certain nombre de mutations ponctuelles dont les effets sur la séquence d'acides aminés codée varient selon les positions affectées des codons et les bases substituées (section 1.4.3). Cependant, sous la pression de sélection, la séquence d'acides aminés codée tend à être préservée. Entre des séquences codantes homologues, les substitutions suivent donc un schéma d'évolution spécifique à cause notamment de la redondance du code génétique : les mutations silencieuses et les mutations qui ont pour effet de produire un acide aminé aux propriétés physico-chimiques proches de l'acide aminé original sont privilégiées. Par nature, ce schéma d'évolution n'affecte pas les séquences fonctionnelles non codantes ou non fonctionnelles. Cette constatation laisse supposer qu'il est possible de distinguer des séquences codantes homologues d'autres ensembles de séquences en comparant de manière systématique toutes les séquences d'acides aminés potentielles obtenues par traduction selon les six cadres de lecture.

Pour une séquence nucléique, il existe six cadres de lecture possibles : $\{1, 2, 3\}$ sur le brin donné, et $\{-1, -2, -3\}$ sur le brin opposé. Etant données deux séquences nucléiques, il faut donc comparer 36 couples de cadres de lecture, et donc 36 couples de séquences d'acides aminés potentielles. La similarité entre chaque paire de séquences peut être estimée en les alignant. L'hypothèse à vérifier est alors que pour deux séquences codantes homologues, le score de similarité d'un couple de cadres de lecture se détache nettement des autres, et que ce couple de cadres de lecture est celui qui permet d'obtenir la paire de séquences d'acides aminés effectivement traduites. De plus, sur tout autre type de séquences, tous les scores de similarité doivent être sensiblement les mêmes. Un exemple est donné en figure 2.3 : la sous-figure (a) montre les résultats obtenus pour deux séquences similaires $\mathcal{U} = \{u_1, u_2\}$ qui ne codent pas pour des séquences peptidiques homologues, tandis que la sous-figure (b) montre les résultats obtenues pour deux séquences codantes homologues $\mathcal{V} = \{v_1, v_2\}$. Ces deux jeux de données présentent strictement les mêmes caractéristiques en terme de longueur et de pourcentage d'identité. Dans ces deux exemples, on observe que six couples de cadres de lecture obtiennent des scores positifs élevés. On parlera alors de couples *compatibles* : $(-3, -3)$, $(-2, -2)$, $(-1, -1)$, $(1, 1)$, $(2, 2)$, $(3, 3)$. Ces couples s'obtiennent en incrémentant ou décrémentant simultanément les cadres de lecture de chaque séquence. Les scores de similarité positifs de ces couples sont la conséquence immédiate de la similarité au niveau nucléique des séquences u_1 et u_2 d'une part, et des séquences v_1 et v_2 d'autre part. Concernant les séquences u_1 et u_2, on constate qu'aucun couple de traductions potentielles ne semble nettement plus conservé que les autres. Au contraire, un couple de traductions se dégage clairement des autres pour v_1 et v_2. Le couple $(1, 1)$ atteint un score de similarité (37) significativement plus élevé que les autres couples. Ce couple correspond en effet aux cadres de lecture de v_1 et v_2



Cadre de u_1 \ Cadre de u_2	-3	-2	-1	1	2	3
-3	10	2				
-2	1	12				
-1	2	1	18			
1				15	1	
2					8	1
3					2	8

(a) Comparaison de toutes les paires de traductions de deux séquences similaires.

Cadre de v_1 \ Cadre de v_2	-3	-2	-1	1	2	3
-3	10					
-2	2	6	1			
-1			25			
1				37		
2					6	
3					5	6

(b) Comparaison de toutes les paires de traductions de deux séquences codantes homologues.

FIG. 2.3 – On considère deux paires de séquence nucléiques, $\mathcal{U} = \{u_1, u_2\}$ et $\mathcal{V} = \{v_1, v_2\}$. Ces deux paires de séquences ont le même pourcentage d'identité, 76,5%. u_1 et u_2 sont des séquences similaires quelconques, alors que v_1 et v_2 présentent un motif évolutif représentatif de séquences codantes homologues. Pour chaque séquence, les six traductions en séquences d'acides aminés sont données. Pour chacune des paires de séquences $\mathcal{U}$ et $\mathcal{V}$, les 36 comparaisons entre les séquences d'acides aminés possibles sont produites. Le score de chaque comparaison est calculé grâce à la matrice BLOSUM62. Les scores négatifs et nuls sont volontairement omis du tableau car ils dénotent un très faible degré de conservation.

qui permettent de produire les bonnes séquences d'acides aminés.

Pour valider notre hypothèse de conservation d'un cadre de lecture, nous avons mené une expérience sur la base de données PANDIT [WdBQ+06]. Dans sa version 17.0, cette base répertorie 7 738 familles de séquences codantes homologues accompagnées des séquences peptidiques correspondantes. Pour chaque famille, nous avons regroupé les séquences par paires de manière aléatoire. Pour chaque paire de séquence, on réalise ensuite la comparaison des 36 couples de traductions potentielles. On compte enfin le nombre de couples de séquences pour lesquels le couple de cadres de lecture correct obtient le meilleur score parmi les 36 couples comparés. La comparaison des séquences peptidiques est réalisée par alignement semi-global exact, une adaptation de l'algorithme de Needleman&Wunsch [NW70] où les insertions et délétions aux extrémités des séquences ont un coup nul. Pour chaque comparaison, la matrice BLOSUM appropriée est automatiquement choisie en fonction de la distance évolutive des séquences nucléiques [HH92]. La figure 2.4 montre les résultats obtenus en fonction de la longueur moyenne et du pourcentage d'identité entre les séquences. Comme on pouvait l'espérer, dans la majorité des cas le bon couple de cadres de lecture obtient le meilleur score parmi les 36 couples de cadres de lecture possibles. Les quelques exceptions constatées apparaissent pour des couples de séquences dont la longueur moyenne est inférieure à 300 nucléotides ou le pourcentage d'identité est supérieur à 80%. Cependant, même dans ces cas de figure, la fréquence où le couple correct obtient le meilleur score est significativement plus élevée que ce que l'on pourrait observer par hasard.

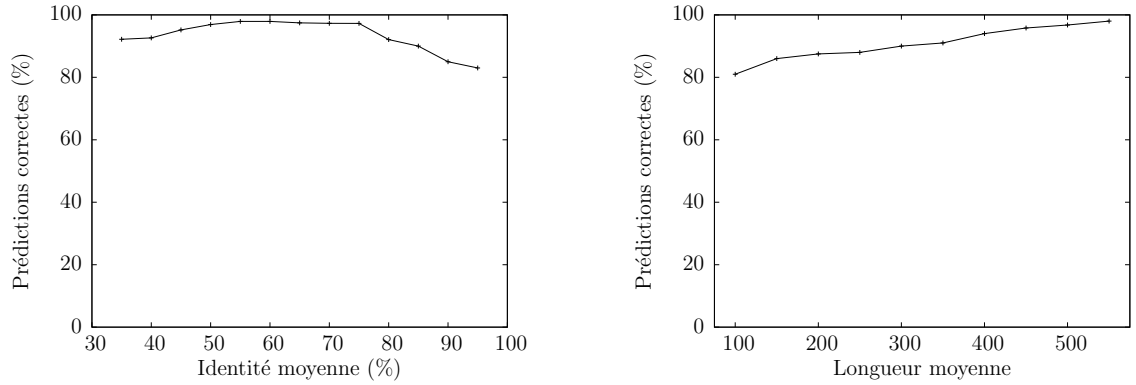


FIG. 2.4 – Proportion de couples de cadres de lecture correctement prédits parmi tous les couples de séquences de chaque famille de PANDIT. Les résultats sont exprimés en fonction du pourcentage d'identité des séquences (gauche) et de leur longueur moyenne (droite).

2.4.2 L'extension à une famille de séquences, le graphe des cadres de lecture

Nous allons maintenant nous intéresser à la définition d'un algorithme pour identifier les ensembles de séquences codantes homologues qui tire parti des observations précédentes réalisées sur deux séquences.

La démarche proposée pour deux séquences ne peut pas être étendue directement à un ensemble de séquences de taille quelconque en travaillant sur un alignement multiple. En effet, comme il existe six cadres de lecture possibles pour une séquence, il existe 6^n n-uplets de cadres de lecture possibles pour un ensemble de n séquences. Pour chacun des n-uplets,

il faudrait alors produire un alignement multiple des n traductions potentielles afin de les comparer. Cette solution n'est évidemment pas praticable car elle requiert de générer une quantité d'alignements multiples exponentielle par rapport au nombre de séquences à traiter. Pour surmonter cette difficulté, il faut trouver une autre manière de généraliser l'approche pour deux séquences à un ensemble de séquences de taille quelconque. Si l'on suppose qu'il existe un cadre de lecture globalement conservé entre toutes les séquences homologues, alors on doit pouvoir détecter cette conservation à partir des comparaisons des traductions potentielles uniquement entre paires de séquences.

Soit $\mathcal{S} = \{s_1, \ldots, s_n\}$ un ensemble de n séquences. Nous utilisons la notion de méta-séquence introduite dans la section 1.5.3. Selon ce processus, on crée une partition $\mathcal{P}$ des séquences de $\mathcal{S}$, où chaque partie correspond à une méta-séquence. On construit ensuite le *graphe des cadres de lecture* à partir des m parties de $\mathcal{P}$. Le but de cette structure est de combiner les comparaisons deux à deux des traductions potentielles.

Définition 1 (Graphe des Cadres de Lecture (GCL))**.** *Le graphe des cadres de lecture $G = (\mathcal{P}, E, \phi)$ est un 36-graphe valué non orienté tel que*

- *$\mathcal{P}$ est l'ensemble des m sommets de G correspondants aux m méta-séquences ;*
- *$E \subset \mathcal{P} \times \mathcal{P} \times C \times C$ est l'ensemble des arêtes du graphe où $C = \{-3, -2, -1, 1, 2, 3\}$ est l'ensemble des cadres de lecture possibles ;*
- *$\phi(P_i, P_j, c, c')$ associe à chaque arête de E une valuation issue de la comparaison entre la méta-séquence P_i traduite selon le cadre c et la méta-séquence P_j traduite selon c'.*

La définition de ϕ suppose que la traduction selon un cadre de lecture soit correctement définie pour une séquence unique et pour une méta-séquence. La définition la plus naturelle de la traduction d'une méta-séquence selon un cadre de lecture donné est de prendre l'ensemble de séquences déduites de la traduction de leur alignement. La traduction de l'alignement se fait ainsi de la même manière que pour une séquence unique, mais au lieu d'obtenir une seule traduction on obtient une traduction par séquence alignée. Chaque triplet composé de trois bases est traduit normalement. Les triplets composés de trois gaps produisent un gap. Les triplets contenant un ou deux gaps sont traduits par un pseudo acide aminé `X`. Ce caractère apparaît dans toutes les matrices de substitution de type BLOSUM et les substitutions qui l'impliquent sont fortement pénalisées. Grâce à cette définition de la traduction d'un alignement, les décalages de cadres de lecture positifs sont supportés entre les séquences regroupées dans une même méta-séquence. Ainsi, à chaque cadre de lecture d'une méta-séquence correspond un ensemble de séquences obtenues par traduction simultanée des séquences bases alignées. La valuation par ϕ d'une arête du GCL est ainsi donnée par la formule suivante

$$\phi(P_i, P_j, c, c') = \sum_{u \in P_i, v \in P_j} \texttt{sim}(u, v, c, c')$$

où $\texttt{sim}(u, v, c, c')$ est le score d'alignement des séquences u et v traduites respectivement selon les cadres de lecture c et c'. A noter que lorsque P_i correspond à une méta-séquence, la séquence u correspond à la séquence correspondante extraite de la traduction de la méta-séquence P_i selon le cadre c. Et respectivement pour P_j. Dans l'approche mise en œuvre pour deux séquences, les scores d'alignements négatifs ou nuls sont ignorés car ils ne reflètent pas un degré de conservation suffisant pour notre investigation. Dans le GCL, cette propriété est généralisée en omettant de placer une arête entre deux sommets lorsque sa valuation est négative ou nulle.

2.4.3 La classification à partir du graphe des cadres de lecture

L'objectif d'un GCL est de pouvoir distinguer les familles de séquences codantes homologues des autres types de séquences. Pour de telles séquences, on s'attend à ce que les comparaisons deux à deux des cadres de lecture soient globalement cohérentes entre elles : pour chaque séquence, un seul cadre de lecture doit être impliqué dans les meilleures comparaisons deux à deux. A l'inverse, sur tout autre type de séquences, on ne s'attend pas à trouver de cohérence dans les comparaisons deux à deux.

On définit le *score de consistance* qui s'applique à un GCL et qui reflète la consistance des prédictions deux à deux qu'il contient. Ce système de scores est cependant peu adapté aux petits GCL. Pour les GCL construits à partir d'un faible nombre de séquences, on utilise alors un autre système : le *score d'alignement*.

Score de consistance

Etant donnés deux sommets, les arêtes qui les relient sont triées par valuation décroissante. On attribue alors à chacune un score qui dépend de son `rang` : l'arête dont la valuation est la plus élevée se voit attribuée un score de 6, la seconde 5, jusqu'à 1. Seules les six meilleures arêtes sont retenues, car comme il est mentionné en section 2.4.1, seuls six couples de cadres de lecture sont censés émerger de la conservation au niveau nucléique.

On définit une *affectation globale* de cadres de lecture comme une fonction de $\mathcal{P}$ dans $C = \{-3, -2, -1, 1, 2, 3\}$ qui attribue un cadre de lecture à chaque sommet d'un GCL. Pour chaque affectation globale $A = (c_1, \ldots, c_m)$, on définit le *score de consistance* noté $\texttt{consistance}(A)$, comme la somme des rangs des arêtes induites par A.

$$\texttt{consistance}(A) = \sum_{1 \leq i < j \leq m} \texttt{rang}(P_i, P_j, c_i, c_j)$$

La valeur de ce score est comprise entre M et $6M$, où M est le nombre de sommets connectés dans le GCL. Dans la plupart des cas, $M = \frac{m(m-1)}{2}$ car chaque couple de sommets est au moins relié par une arête. La valeur optimale $6M$ est atteinte lorsqu'une affectation globale couvre toutes les meilleures comparaisons deux à deux.

La qualité d'un GCL est donnée par le score de consistance le plus élevé obtenu par une affectation globale. En pratique, il est rarement nécessaire d'énumérer toutes les affectations pour trouver celle dont le score de consistance est maximal. Les scores des arêtes étant bornés, on applique les techniques standards de rebroussement et d'élagage pour éviter des calculs inutiles.

Significativité d'un score de consistance

Pour évaluer la significativité d'un score de consistance, on calcule sa P-valeur, c'est-à-dire la probabilité d'observer un score de consistance égal ou plus grand étant donnée la topologie du GCL. On suppose que le rang de chaque arête est une variable discrète uniformément distribuée dans l'intervalle $[1; 6]$. On suppose également que les scores d'alignement obtenus par deux paires différentes de séquences sont des variables aléatoires indépendantes. Selon ces hypothèses, la distribution d'une somme de rangs est calculable au moyen d'un produit de convolution discret d'un nombre fini de variables uniformément distribuées. La formule analytique pour ce calcul a été établie par Uspensky [Usp37]. Cette formule nous permet d'obtenir la formule de la P-valeur d'un score de consistance.

$$\Pr[\texttt{consistency\_score}(A) \geq s] = \sum_{i=s}^{6M} \frac{1}{6^M} \sum_{j=0}^{\lfloor (i-M)/6 \rfloor} (-1)^j \left(\begin{array}{c} M \\ j \end{array} \right) \left(\begin{array}{c} i-6j-1 \\ M-1 \end{array} \right)$$

Preuve de la formule de Uspensky. On souhaite calculer la probabilité d'obtenir une somme p de N variables aléatoires indépendantes identiquement distribuées selon une loi uniforme discrète à valeurs dans $[1;6]$. Le nombre d'arrangements permettant d'obtenir p est donné par le coefficient c de x^p dans

$$f(x) = (x + x^2 + \cdots + x^6)^N,$$

où chaque arrangement possible correspond à un terme. $f(x)$ peut également s'écrire comme une série multinômiale

$$\begin{aligned} f(x) &= x^N \left(\sum_{i=0}^{5} x^i \right)^N \\ &= x^N \left(\frac{1-x^6}{1-x} \right)^N \\ &= x^N (1-x^6)^N (1-x)^{-N} \end{aligned}$$

Selon le théorème binômial,

$$\begin{aligned} x^N (1-x^6)^N &= \sum_{k=0}^{N} (-1)^k \left(\begin{array}{c} N \\ k \end{array} \right) x^{N+6k} \\ &= x^N \sum_{k=0}^{N} (-1)^k \left(\begin{array}{c} N \\ k \end{array} \right) x^{6k} \\ (1-x)^{-N} &= \sum_{l=0}^{\infty} \left(\begin{array}{c} N+l-1 \\ N-1 \end{array} \right) x^l \end{aligned}$$

D'où l'expression,

$$x^N \sum_{k=0}^{N} (-1)^k \left(\begin{array}{c} N \\ k \end{array} \right) x^{6k} \sum_{l=0}^{\infty} \left(\begin{array}{c} N+l-1 \\ l \end{array} \right) x^l,$$

donc le coefficient c de x^p inclus tous les termes où

$$p = N + 6k + l.$$

c est par conséquent donné par

$$c = \sum_{k=0}^{N} (-1)^k \left(\begin{array}{c} N \\ k \end{array} \right) \left(\begin{array}{c} p-6k-1 \\ p-6k-N \end{array} \right).$$

Cependant, $p - 6k - N > 0$ seulement si $k < (p-N)/6$, donc les autres termes ne contribuent pas au calcul. De plus,

$$\begin{pmatrix} p-6k-1 \\ p-6k-N \end{pmatrix} = \begin{pmatrix} p-6k-1 \\ N-1 \end{pmatrix},$$

donc

$$c = \sum_{k=0}^{\lfloor (p-N)/6 \rfloor} (-1)^k \begin{pmatrix} N \\ k \end{pmatrix} \begin{pmatrix} p-6k-1 \\ N-1 \end{pmatrix}.$$

La probabilité $P(p, N)$ que p soit la somme de N variables aléatoires indépendantes suivant la même loi discrète uniforme dans $[1; 6]$ est donc donnée par

$$P(p, N) = \frac{1}{6^N} \sum_{k=0}^{\lfloor (p-N)/6 \rfloor} (-1)^k \begin{pmatrix} N \\ k \end{pmatrix} \begin{pmatrix} p-6k-1 \\ N-1 \end{pmatrix}.$$

□

En fixant un seuil sur la P-valeur de la meilleure affectation globale, on est en mesure de déterminer la nature des séquences analysées. Si cette P-valeur est inférieure au seuil fixé, on suppose alors être en présence de séquences codantes homologues. Dans le cas contraire, on suppose que les séquences ne correspondent pas à des séquences codantes homologues. Nous avons déterminé de manière empirique la valeur de ce seuil durant l'élaboration de la méthode.

Petits GCL : z-score du score d'alignement

Le calcul d'un score de consistance n'est pas approprié pour des GCL ayant moins de trois sommets. Sur de tels GCL, les P-valeurs des scores de consistance ont en effet toutes des valeurs trop élevées. Pour évaluer la qualité des GCL comportant peu d'arêtes, on utilise alors l'information des scores d'alignement. Comme il a déjà été remarqué précédemment, on s'attend à ce que le meilleur alignement fasse intervenir les bons cadres de lecture sur des séquences codantes homologues. Qui plus est, plus le score de cet alignement se détache des autres, plus le biais en faveur d'une conservation de la séquence d'acide aminés est important. On définit donc le *score d'alignement* d'une affectation globale $A = (c_1, \ldots, d_m)$, noté `alignment_score(A)`, comme la somme des valuations des arêtes induites par A.

$$\texttt{alignment\_score}(A) = \sum_{1 \leq i < j \leq m} \phi(P_i, P_j, f_i, f_j)$$

On sélectionne pour un GCL l'affectation globale de score d'alignement maximal. Contrairement au score de consistance, un score d'alignement n'est pas informatif en soi car il reflète partiellement la similarité entre les séquences nucléiques initiales. Ce score ne peut être interprété que s'il est mis en relation avec les scores d'alignement atteints par les autres affectations globales. Plus précisément, son interprétation dépend des scores des autres affectations globales qui lui sont compatibles. Pour une affectation globale, il existe cinq affectations compatibles obtenues par décalage simultané de tous les cadres de lecture. Par exemple, pour une affectation globale $(1, 2, 3, 1)$, il existe deux affectations compatibles sur le même brin, $(2, 3, 1, 2)$ et $(3, 1, 2, 3)$, et trois sur l'autre brin qui dépendent des longueurs des séquences modulo 3. La différence entre les scores d'alignements des affectations compatibles provient

ainsi uniquement d'un motif évolutif particulier et non d'une similarité au niveau nucléique. On exclut toutefois une autre affectation des affectations compatibles, l'affectation sur le brin opposé où la troisième position des codons coïncide avec celle de l'affectation à évaluer. Le score de cette affectation est en effet clairement biaisé à cause d'une propriété du code génétique. En effet, les mutations silencieuses apparaissent le plus souvent sur la troisième base des codons (section 1.4.2). Lorsque l'on observe un grand nombre de mutations silencieuses entre deux cadre de lecture, on peut donc naturellement obtenir deux cadres de lecture sur les brins opposés qui présentent ce même biais, comme illustré sur la figure 2.5 où les trois mutations silencieuses entre les cadres +1 produisent sur les brins opposés deux mutations silencieuses fortuites. Sur l'exemple de la figure 2.3, le couple de cadres de lecture correct pour les séquences de $\mathcal{V}$ est $(1, 1)$. On constate un certain nombre de mutations silencieuses entre ces deux cadres, c'est pourquoi le cadre $(-1, -1)$ obtient également un bon score. Pour interpréter le score d'alignement d'une affectation, on ne garde donc finalement que quatre des cinq affectations compatibles. La significativité de la déviation du score d'alignement de la meilleure affectation globale d'un GCL est mesuré par un z-score, même si le nombre d'observations est faible.

Dans le cas des petits GCL, la détection d'ensembles de séquences codantes homologues se fait au moyen d'un seuil sur le z-score du score de la meilleure affectation globale. A l'instar de la classification réalisée sur les GCL de plus grande taille, la valeur de ce seuil a été déterminée de manière empirique au cours de la validation de la méthode.

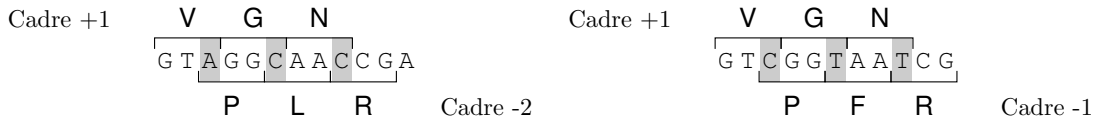


FIG. 2.5 – Du fait de la redondance du code génétique, les mutations dites silencieuses (positions grisées) n'ont aucun effet sur les acides aminés codés. Ces mutations apparaissent plus fréquemment à la troisième position des codons.

2.4.4 Mise en œuvre logicielle

Une implémentation de PROTEA a été réalisée en C. PROTEA est un logiciel quasiment autonome qui s'appuie sur deux librairies, GMP (`http://gmplib.org`) et MPFR [FHL+07], pour le calcul exact de la P-valeur d'un score d'affectation. En effet, bien que l'on dispose d'une formule analytique pour ce calcul, celle-ci fait intervenir des calculs de combinaisons qui dépassent rapidement les capacités des types primitifs disponibles en C. Grâce aux librairies GMP et MPFR, il nous est possible de spécifier précisément la capacité nécessaire et suffisante pour les variables servant au calcul des P-valeurs.

Les alignements deux à deux sont directement réalisés par PROTEA en utilisant les matrices BLOSUM, tandis que la construction des alignements multiples est déléguée à un programme externe. Actuellement, PROTEA offre la possibilité d'utiliser indifféremment trois programmes d'alignement multiple : CLUSTALW [THG94], DIALIGN2-2 [Mor99] et T-COFFEE [NHH00].

L'implémentation en C de PROTEA est complétée par un ensemble de scripts CGI écrits en Python permettant une utilisation du logiciel via un navigateur Web. Cette interface, disponible à l'adresse `http://bioinfo.lifl.fr/protea` permet d'éviter aux utilisateurs l'installation locale du logiciel et offre en plus une présentation plus lisible et efficace des résultats.

2.5 Résultats expérimentaux de Protea

Nous avons conduit une série d'expériences dans le but de valider notre méthode et d'évaluer ses performances par rapport aux méthodes existantes (section 2.5.1). Nous nous sommes par la suite intéressés à une application plus concrète de PROTEA : la prédiction de nouvelles séquences codantes dans le génome humain (section 2.5.2). Au cours de cette expérience, nous avons également pu confronter les prédictions de PROTEA à celles réalisées par des méthodes *ab initio* et des méthodes destinées à l'annotation de séquences génomiques complètes.

2.5.1 L'évaluation des performances de Protea

A l'heure actuelle, il n'existe aucun de jeu de données de référence pour évaluer la prédiction de gènes ou de régions codantes [PAA+03]. Bien que plusieurs expériences aient été menées dans ce sens [RMO01, BG96, KC07], leur objectif est d'évaluer la précision de prédictions des bornes des gènes et des régions codantes dans des séquences génomiques individuelles et non par approche comparative sur plusieurs séquences. Les jeux de données utilisés par les auteurs de méthodes de prédiction par analyse comparative sont limités exclusivement à des paires de séquences conservées à plus de 75%. Bien que ce degrés de conservation est adapté pour pouvoir utiliser des méthodes plus conventionnelles travaillant uniquement sur des alignements, ces séquences ne peuvent constituer à elles seules un jeu de données adapté et représentatif pour évaluer PROTEA. Pour mener à bien notre expérience, nous avons donc constituer des jeux de données plus variés à partir de séquences disponibles dans les banques publiques. Sur ces jeux de données, nous aurions souhaité pouvoir comparer les performances de PROTEA à celles des trois méthodes apparentées : QRNA, ROSETTA et SYNCOD. ROSETTA et SYNCOD sont relativement âgées et ne sont hélas plus maintenues depuis déjà quelques années. De plus, elles ne sont plus disponibles ni en ligne, ni auprès de leurs auteurs. Pour cette raison, les performances de PROTEA n'ont pu être comparées qu'à celles de QRNA.

Les jeux de données

Trois jeux de données ont été construits pour mesurer les performances de PROTEA : des familles de séquences codantes, des ARN non-codants et des séquences aléatoires. Pour chaque jeu de données, des sous-ensembles ont été construits aléatoirement de façon à ce que chaque famille ne soit représentée qu'une seule fois et que les sous-ensembles soient les plus équivalents possible en terme de pourcentage d'identité et de longueur.

Le jeu de données CODANT : des familles de séquences codantes. Ce jeu de données est produit à partir de la banque de données PANDIT (section 2.4.1) complétée de séquences extraites de POPSET [WBB+08]. La conservation moyenne des familles des séquences de PANDIT est relativement faible, c'est pourquoi nous avons ajouté dans ce jeu de données des séquences provenant de POPSET. POPSET contient des ensembles de séquences nucléiques collectées pour analyser les relations évolutives d'une population. Les populations peuvent être de deux natures : intra-espèce ou inter-espèces. Nous avons ainsi sélectionné 58 familles de POPSET qui présentent un degré de conservation élevé. Au total, le jeu de données CODANT est composé de 7 796 familles de séquences codantes homologues.

Le jeu de données NONCODANT : des familles d'ARN non-codants. Pour construire ce jeu de données, nous avons extrait des familles de séquences de RFAM, complétées de séquences provenant de la banque européenne de données ribosomiques [WPVdP04]. Seule une partie des séquences de RFAM a été utilisée car certaines séquences présentes dans cette base recouvrent des séquences codantes. A partir de RFAM version 8.0, qui contient 574 familles d'ARN, nous avons filtré les familles qui contiennent ou recouvrent une séquence codante. Le filtre appliqué consiste à éliminer toute famille contenant au moins une séquence pour laquelle il existe une séquence peptidique hautement similaire dans SWISSPROT. La recherche de séquence similaire a été conduite avec BLASTX paramétré avec un seuil sur la E-valeur à 10^{-4} et une couverture minimale de 50% de la séquence requête. 110 familles ont ainsi été supprimées. La longueur moyenne des séquences des 464 familles restantes étant relativement faible, nous avons donc ajouté 32 familles d'ARN ribosomiques (grosse sous-unité) provenant de la banque européenne de données ribosomiques. Au final, le jeu de données NONCODANT est composé des 496 familles.

Le jeu de données ALEATOIRE : des familles de séquences aléatoires. Ce jeu de données est composé de séquences aléatoires générées à partir du jeu de données CODANT. L'alignement multiple de chaque famille de CODANT fourni dans PANDIT est mélangé selon un processus conservatif [WH04]. Cette procédure assure que les deux jeux de données CODANT et ALEATOIRE ont les mêmes propriétés en terme de longueur, de composition nucléotidique et de conservation globale mais surtout de conservation locale. La figure 2.6 montre un exemple d'alignement mélangé où le degré de conservation locale de l'alignement originale est préservé.

```
VGJ_BPG4      AAAAAATCAATTCGCCGCTCTGGT----------------------------------------GGCAAATCTAAGGGTGCCCGTCTCTGGTATGTAGGCGGAACACAATAC
Q9G087_BPS13  TCTAAAGGTAAAAAACGTTTTGGCGCTCGCTCCGGTCGTCCACAGCCGTTGCGAGGCACTAAAGGCAAGCGTAAAGGCGCTCGTCTTTGGTATGTAGGCGGTCAACAATTT
VGJ_BPAL3     ATGAAGAAAGCACGTCGTTCTCCT----------------------------------------AGTCGTCGTAAAGGTGCTCGCCTCTGGTATGTAGGCGGTTCTCAGTTT
                 **          ** * *                                            *      *** ** ** ** ** **************    ** *

seq0          GACATGTTCAACAAATATGCGAAT----------------------------------------ATCAGATCTGACTACCGGTTAGGTAGTCAGGCGGTCGCACCGAGTCAC
seq1          ACGATGAGTACTTTTTACGAGAACCCGGGGTTCTGAATCTGGAGTGCCCCGAACCGCACCACTATCAGGAGTGATTATCGATTAGGCAGTCAGGCGGTCGCTAAGAGTCTT
seq2          GTAATACACGTTAAGTACGCGTGT----------------------------------------CTTGCTAGTGATTACCGATTGGGTAGTCAGGCGGTCGCTTCCAGCCTT
                 **          ** * *                                            *      *** ** ** ** ** **************    ** *
```

FIG. 2.6 – En haut, l'alignement multiple de la famille PF04726 du jeu de données CODANT ; en bas, un alignement multiple obtenu par la procédure de mélange conservative. Les caractères '*' marquent les positions parfaitement conservées.

Les résultats de Protea

Les jeux de données décrits précédemment sont soumis à PROTEA qui les classifie selon deux classes : "codant" pour les ensembles de séquences prédits comme étant des familles de séquences codantes homologues, "autre" dans le cas contraire. Contrairement à PROTEA, QRNA classifie les séquences selon trois classes : "codant", "non-codant" ou "autre". La classification "non-codant" correspond à l'identification de séquences présentant une structure secondaire conservée (section 3.2.4). Etant donné que nous nous intéressons ici à la prédiction de séquences codantes homologues, les prédictions "non-codant" sont considérées comme des prédictions "autre". Ce stratagème permet d'obtenir une classification des jeux de données en deux classes équivalentes pour PROTEA et QRNA : "codant" et "autre".

Mesure de performances. Pour évaluer la qualité des prédictions effectuées par un classifieur binaire, deux mesures sont couramment utilisées : la sensibilité et la spécificité. La sensibilité de la méthode correspond ici à la proportion de séquences codantes homologues classifiées "codant", tandis que la spécificité correspond à la proportion de séquences d'un autre type classifiées "autre". La sensibilité S_n sur un jeu de données est donnée par

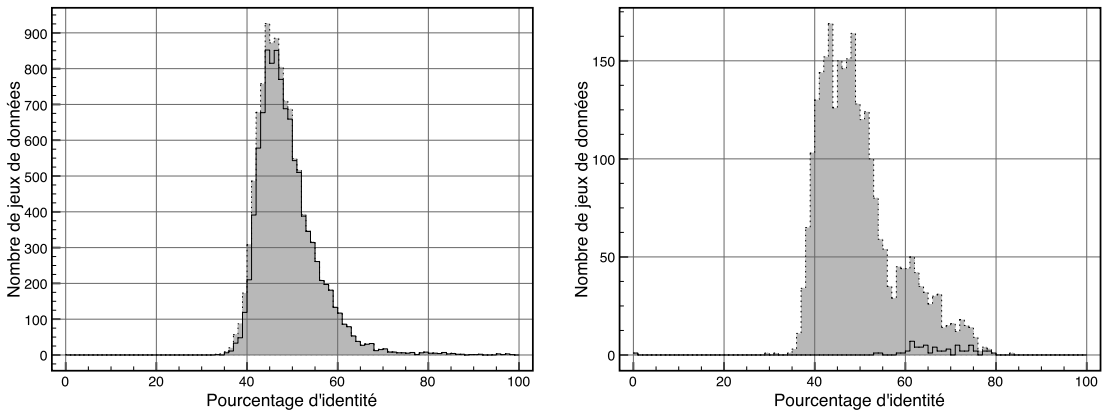
$$S_n = \frac{TP}{TP + FN}$$

où TP désigne la quantité de vrais positifs, c'est-à-dire les prédictions "codant" correctes, et FN la quantité de faux négatifs, c'est-à-dire les prédictions "autre" incorrectes. De manière analogue, la spécificité S_p sur un jeu de données est donnée par

$$S_p = \frac{TN}{TN + FP}$$

où TN désigne la quantité de vrais négatifs, c'est-à-dire les prédictions "autre" correctes, et FP, c'est-à-dire les prédictions "codant" incorrectes.

Résultats généraux. Les résultats de Protea sur les jeux de données CODANT, NONCODANT et ALEATOIRE pour des ensembles de 3, 5 et 11 séquences sont répertoriés dans la table 2.1 selon le pourcentage d'identité moyen et la longueur moyenne des séquences. Ces résultats sont également présentés graphiquement sur la figure 2.7.



(a) Répartition des prédictions "codant" sur le jeu de données CODANT.

(b) Répartition des prédictions "codant" sur le jeu de données NONCODANT.

Fig. 2.7 – Répartition des prédictions "codant" de Protea sur les ensembles de 11 séquences des jeux de données CODANT (a) et NONCODANT (b). Les histogrammes tracés en pointillés représentent les ensembles de données initiaux, tandis que les histogrammes en trait plein représentent les prédictions "codant" de Protea.

Dans la plupart des cas, la sensibilité et la spécificité sont supérieures à 80%. Comme on pouvait l'espérer, les performances des Protea augmentent avec le nombre de séquences ainsi que leur longueur. En dessous de 50% d'identité moyenne, Protea est particulièrement performant et affiche une spécificité supérieure à 90%. En revanche, les performances de Protea se dégradent au delà de 90% d'identité. Sur de telles séquences très conservées, la

quantité de mutations est insuffisante pour détecter un schéma de substitution en lien avec la conservation d'une séquence d'acides aminés particulière. Le comportement de PROTEA sur des courtes séquences est soumis à un biais causé par la présence d'acides aminés rares, tel que le tryptophane par exemple, conservés entre des traductions potentielles. Les scores de substitution de ces acides aminés dans les matrices BLOSUM sont relativement élevés et la conservation fortuite d'un seul acide aminé de ce type sur des séquences de moins de cinquante nucléotides entraîne une élévation mécanique des scores lors de leur alignement.

Comparaison avec Qrna. PROTEA a été conçu pour traiter des ensembles de plusieurs séquences non alignées. Néanmoins, il peut quand même être utilisé sur des paires de séquences, bien qu'il ne soit pas spécialement conçu pour ce cas de figure.

Comme QRNA nécessite en entrée des séquences alignées, nous avons testés plusieurs méthodes d'alignement : CLUSTALW, T-COFFEE, DIALIGN2-2 et BLAST. Les performances de QRNA étant meilleures sur les alignements produits par CLUSTALW, seuls les résultats de QRNA obtenus avec cette méthode sont présentés.

Les résultats de QRNA et de PROTEA sur les couples de séquences des jeux de données CODANT et NONCODANT sont reportés dans la table 2.2. Globalement, QRNA est plus spécifique que PROTEA, tandis que PROTEA est plus sensible sur les mêmes données. Si l'on considère le compromis entre sensibilité et spécificité, PROTEA se montre plus performant que QRNA qui dispose de plus, rappelons le, d'un modèle pour détecter les séquences non-codantes homologues qui l'avantage sur le jeu de données NONCODANT. De plus, PROTEA est clairement plus performant que QRNA en dessous de 50% d'identité : 38% de sensibilité et 100% de spécificité pour QRNA, contre 81,3% et 95,5% respectivement pour PROTEA. Cette observation permet de mettre en évidence les limites intrinsèques des méthodes basées sur des alignements en présence de séquences divergentes. En terme de temps de calculs, QRNA s'avère beaucoup plus gourmand que PROTEA notamment à cause de l'évaluation du modèle non-codant très coûteuse. Pour réaliser cette expérience, il a fallu moins d'une heure à PROTEA contre plus de 40 heures à QRNA.

2.5.2 Une application au génome humain

L'annotation des éléments conservés des génomes nouvellement séquencés ou partiellement annotés reste à l'heure actuelle une tâche difficile. En complément des approches de prédiction classiques par homologie qui permettent de retrouver des séquences codantes connues, et des approches *ab initio* qui permettent d'identifier de nouvelles séquences codantes putatives, les approches comparatives peuvent jouer un rôle important en apportant des prédictions de nouvelles séquences codantes plus fiables car supportées par plus d'une séquence. Nous avons donc conduit une étude visant à découvrir de nouvelles séquences codantes sur le génome humain grâce à PROTEA. Cette étude menée sur des séquences conservées entre le génome humain et plus d'une dizaine d'organismes nous a permis d'identifier de nouvelles séquences codantes putatives.

Les séquences conservées entre plusieurs espèces

L'UCSC GENOMEBROWSER propose des séquences conservées entre plusieurs espèces [KBD+03, KHF+04] extractibles à partir de la piste nommée multiz17way. Ces

(a) Sensibilité sur le jeu de données CODANT.

Nb séq.	Longueur moyenne	Pourcentage d'identité moyen						
		<50	50-60	60-70	70-80	80-90	90-95	>95
3	<100	57,1	68,9	71,9	77,8	62,9	41,7	38,2
	100-200	72,8	91,7	87,3	85,2	70,0	61,3	44,4
	200-300	78,9	93,4	92,8	92,1	71,4	66,7	58,8
	>300	86,7	97,0	96,5	93,3	81,6	64,7	61,5
5	<100	70,5	76,6	81,8	83,3	63,6	57,1	56,1
	100-200	86,9	96,1	91,1	89,3	71,4	55,6	64,3
	200-300	88,1	98,3	97,3	95,0	78,7	60,0	66,7
	>300	94,0	98,5	99,4	94,4	80,8	66,7	71,4
11	<100	82,2	92,0	92,6	89,0	74,1	65,1	55,6
	100-200	93,2	96,6	93,8	92,8	75,0	65,6	56,8
	200-300	95,1	98,5	96,3	95,4	79,2	66,4	61,7
	>300	96,4	99,7	100	96,8	92,6	78,6	73,8

(b) Spécificité sur le jeu de données NONCODANT.

Nb séq.	Longueur moyenne	Pourcentage d'identité moyen						
		<50	50-60	60-70	70-80	80-90	90-95	>95
3	<100	91,5	90,0	88,7	88,9	83,6	79,0	76,3
	100-200	96,4	92,0	85,7	85,7	84,8	82,4	81,3
	200-300	91,7	80,0	85,7	86,0	83,3	80,4	85,7
	>300	100	93,0	81,8	90,0	87,7	76,3	78,6
5	<100	94,3	94,0	93,3	90,5	79,8	79,5	77,8
	100-200	95,8	87,6	86,0	82,8	82,8	78,3	80,0
	200-300	97,8	93,3	86,7	86,7	84,8	81,5	76,3
	>300	100	91,7	90,6	90,6	81,8	83,8	78,4

(c) Spécificité sur le jeu de données ALEATOIRE.

Nb séq.	Longueur moyenne	Pourcentage d'identité moyen						
		<50	50-60	60-70	70-80	80-90	90-95	>95
3	<100	96,8	93,2	81,1	80,7	80,0	78,1	60,0
	100-200	97,4	97,2	83,3	81,1	83,3	66,7	61,9
	200-300	97,9	95,7	86,6	87,0	85,2	70,2	62,5
	>300	98,1	95,0	93,3	90,5	88,9	83,3	64,3
5	<100	97,4	95,1	87,5	74,8	77,6	68,4	61,4
	100-200	98,3	95,9	94,3	93,2	86,9	78,7	62,5
	200-300	98,3	97,5	96,4	95,8	93,4	81,8	68,1
	>300	100	98,5	97,2	96,6	94,4	83,3	73,3
11	<100	99,9	98,1	99,6	97,9	89,0	87,5	61,7
	100-200	100	100	99,8	98,6	88,9	80,0	61,1
	200-300	100	100	100	100	100	81,8	64,3
	>300	100	100	100	100	100	96,0	79,4

TAB. 2.1 – Les résultats de PROTEA sur les jeux de données CODANT, NONCODANT et ALEATOIRE de 3, 5 et 11 séquences. Les résultats avec 11 séquences pour le jeu de données NONCODANT ne sont pas mentionnés car très peu de familles de ce jeu de données contiennent autant de séquences. Le tableau (a) contient la proportion de données correctement classifiées "codant". Les tableaux (b) et (c) les proportions de données correctement classifiées "autre" par PROTEA.

(a) Les résultats de QRNA.

Id. moy. %	Sensibilité (%) CODANT	Spécificité (%) NONCODANT
<50	38,0	100,0
50-60	63,1	98,4
60-70	73,4	97,9
70-80	69,6	93,7
80-90	43,1	91,8
90-95	38,2	90,2
>95	30,7	88,9

(b) Les résultats de PROTEA.

Id. moy. %	Sensibilité (%) CODANT	Spécificité (%) NONCODANT
<50	81,3	95,5
50-60	90,6	76,8
60-70	88,9	74,2
70-80	82,5	85,3
80-90	66,7	83,2
90-95	61,5	84,2
>95	54,8	64,3

TAB. 2.2 – Les résultats de QRNA et PROTEA sur les couples de séquences.

ensembles de séquences sont construits à partir d'alignements multiples générés par MULTIZ [BKR+04] puis filtrés par PHASTCONS [SH05]. Dix huit génomes d'eucaryotes supérieurs dont l'Homme, la souris, le chien et le poulet ont ainsi été comparés.

L'objectif de notre expérience est de découvrir de nouvelles séquences codantes chez l'Homme, c'est pourquoi nous avons retiré de ce jeu de données toute ensemble contenant une séquence déjà identifiée comme telle de manière expérimentale. A cet effet, nous avons utilisé les ressources fournies par l'UCSC TABLEBROWSER pour filtrer les séquences chevauchantes ou incluses dans les pistes KnownGene [HKC+06] ou MGC (Mammalian Gene Collection). Etant donné les performances de PROTEA sur les séquences courtes ou trop conservées, les éléments de moins de cinquante nucléotides ou dont le pourcentage d'identité est supérieur à 90% ont été écartés. Au total, 97 956 ensembles d'au moins douze séquences ont été soumis à PROTEA. Ce jeu de données peut être séparé en deux groupes : les *séquences annotées* dont la fonction putative a déjà été prédite par d'autres méthodes bio-informatiques, et les *séquences non annotées*. Parmi les 97 956 ensembles analysés, on compte 37 318 ensembles contenant des séquences annotées et 60 638 sans annotation. Les résultats obtenus sont schématisés sur la figure 2.8.

Les séquences avec annotations putatives

Une partie des séquences traitées comportent des annotations réalisées par des méthodes de prédiction *ab initio* ou par analyse comparative. L'UCSC Table Browser nous a permis de récupérer les annotations réalisées par AUGUSTUS, EXONIPHY, EXONWALK, GENEID, GENSCAN et N-SCAN. Nous avons également réalisé des annotations en utilisant une approche classique par homologie de séquences au niveau peptidique grâce à BLASTX sur la base SWISSPROT en filtrant les alignements dont la E-valeur est inférieure à 10^{-4}.

Globalement, 23 220 ensembles de séquences annotées sont confirmés par PROTEA, soit 62% des séquences annotées par d'autres méthodes. La table 2.3 explicite la répartition des prédictions de PROTEA en fonction des autres méthodes. Cette table contient également le cœfficient de corrélation entre les prédictions de chaque méthode et celles de PROTEA. Les valeurs de ce cœfficient comprises entre 0,1 et 0,26 sont relativement faibles. Cette expérience montre que l'approche comparative de PROTEA est complémentaire des approches existantes, notamment des approches *ab initio* et par analyse comparative.

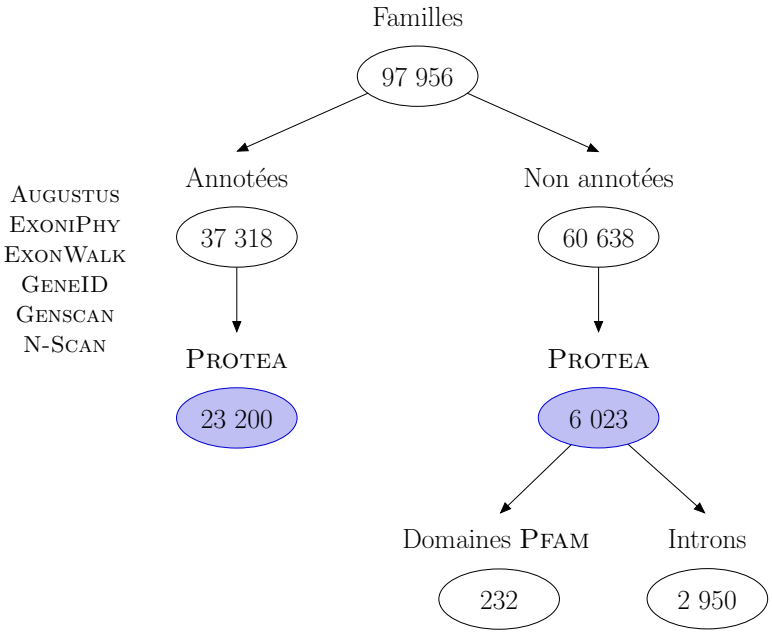


Fig. 2.8 – Découpage des résultats de Protea sur les groupes de séquences similaires de l'UCSC.

Annotation	Nombre d'éléments	Prédictions chevauchantes	Cœfficient de corrélation
Augustus	8 383	6 702	0,20
ExonWalk	3 994	3 284	0,14
ExoniPhy	19 882	14 497	0,24
GeneID	20 410	13 919	0,14
Genscan	23 810	15 691	0,10
N-Scan	10 920	7 814	0,12
SwissProt	23 174	16 676	0,26

Tab. 2.3 – Les résultats de Protea sur les éléments conservés comportant des annotations putatives.

Les séquences sans annotation

PROTEA prédit 6 023 des 60 638 ensembles de séquences comme des ensembles de séquences codantes homologues, soit 9,93%. Afin d'estimer le taux de faux positifs, nous avons construit un jeu de données de contrôle selon la même procédure que pour le jeu de données ALEATOIRE construit pour la validation de PROTEA. Parmi ce jeu de données, seuls 0,8% des ensembles de séquences sont prédits comme "codant". Les prédictions positives de PROTEA sur les ensembles de séquences non annotées ne sont donc pas un artefact de PROTEA. Si l'on s'intéresse plus en détails aux 6 023 prédictions positives de PROTEA, on remarque que 232 d'entre elles contiennent des domaines protéiques connus répertoriés dans PFAM [FMSB+06], contre 272 pour les 54 615 prédictions négatives de PROTEA. De plus, si l'on s'intéresse aux positions dans le génome humain des prédictions positives, on en compte 2 050 dans des régions introniques de gènes vérifiés expérimentalement et 900 à proximité immédiate d'exons prédits. Cette constatation laisse supposer qu'une partie des prédictions de PROTEA sont impliquées dans l'épissage alternatif, ou correspondent à des exons non annotés de gènes prédits.

2.5.3 Conclusions

Les expériences présentées dans la section 2.5.1 permettent d'apprécier en pratique les forces et les faiblesses de PROTEA. PROTEA est une méthode efficace et performante, capable de traiter des séquences non alignées très faiblement conservées. Les performances de PROTEA sont, qui plus est, cohérentes avec le comportement attendu pour une méthode à base d'analyse comparative : ses performances croissent avec le nombre et la longueur des séquences comparées. Au cours de ces expériences, nous avons noté que PROTEA n'était pas à l'aise sur des séquences courtes ou très bien conservées. Ces faiblesses proviennent du principe même de l'analyse comparative de séquences. Comparer des séquences quasiment identiques n'apporte pas plus d'information qu'une seule séquence. Les séquences trop courtes ne contiennent pas suffisamment d'information pour réaliser des observations significatives et pertinentes. L'application de PROTEA à l'annotation de séquences codantes sur le génome humain (section 2.5.2) a permis d'une part de mettre en évidence la complémentarité de PROTEA par rapport aux méthodes de prédictions existantes, et d'autre part de détecter de nouveaux fragments de séquences codantes putatives avec un degré de confiance élevé.

Chapitre 3

Prédiction de structures communes d'ARN non-codants homologues

Dans le chapitre précédent, nous avons abordé le problème de l'identification des régions codantes. Nous nous intéressons maintenant aux ARN non-codants. Comme pour les ARN codants, plusieurs types d'informations peuvent être pris en compte tels que l'homologie et les biais de composition. Toutefois, la majorité des ARN non-codants présentent une particularité supplémentaire qui est la formation d'une structure spatiale stable (section 1.3.1), que l'on peut capturer partiellement à travers la structure secondaire. C'est un signal important qui s'avère fort utile pour la prédiction de gènes à ARN. De ce fait, nous commençons par présenter dans la première section les approches principales pour la prédiction de structures secondaires. En section 3.2, nous expliquons ensuite comment ces méthodes s'appliquent à la prédiction de gènes à ARN. Enfin, dans la section 3.3, nous présentons notre contribution au problème, avec une évolution du logiciel CARNAC et des premiers résultats pour la prédiction de gènes.

3.1 La prédiction de structures secondaires, état de l'art

La prédiction de la structure secondaire d'un ARN est un problème de bio-informatique relativement ancien. Les premières approches virent le jour au début des années 80. En effet, les techniques expérimentales pour obtenir des informations structurales de macromolécules biologiques, en particulier d'acides nucléiques, sont délicates. La cristallographie aux rayons X, qui est la technique de référence pour cela, nécessite par exemple d'emprisonner la molécule d'intérêt dans un cristal afin de figer sa structure et de pouvoir l'observer. De plus, dans la cellule, les molécules ne sont pas isolées dans leur milieu mais en interaction avec d'autres molécules. Il est alors d'autant plus difficile de résoudre la structure d'un complexe entier. Peu de structures d'ARN ont pu être caractérisées par cette méthode qui fait appel à des techniques expérimentales lourdes et relativement coûteuses en temps et en argent.

Les méthodes bio-informatiques pour traiter ce problème se révèlent donc être une alternative peu coûteuse à mettre en œuvre. Plusieurs approches ont été proposées. Initialement, on peut distinguer schématiquement deux écoles : approches thermodynamiques émanant de la physique statistique basée sur la stabilité d'une molécule, et approches comparatives qui exploitent le schéma évolutif d'un ensemble de séquences supposées homologues pour en déterminer une structure commune. A ce tableau s'ajoutent les méthodes hybrides qui s'ap-

puient généralement sur un modèle thermodynamique soutenu par des signes d'évolution des séquences selon le schéma évolutif des ARN non-codants. Nous détaillons dans la suite de cette section les méthodes thermodynamiques, comparatives et hybrides.

3.1.1 La prédiction par approche thermodynamique

Hypothèses de travail

Le premier principe de la thermodynamique affirme qu'au cours d'une transformation quelconque d'un système fermé, la variation de son énergie est égale à la quantité d'énergie échangée avec le milieu extérieur, sous forme de chaleur et sous forme de travail. Autrement dit, l'énergie totale d'un système isolé reste constante. Les événements qui s'y produisent ne se traduisent que par des transformations de certaines formes d'énergie en d'autres formes d'énergie. L'énergie ne peut donc pas être produite *ex nihilo* ; elle est en quantité invariable dans la nature. L'énergie libre est une fonction d'état d'un système dont la variation permet d'obtenir le travail utile susceptible d'être fourni par un système thermodynamique fermé, à température constante. Dans un système composé uniquement d'une molécule d'ARN, la stabilité structurale de cette molécule est mesurée par la perte d'énergie libre accompagnant la transition d'un état non replié ou dénaturé à un état natif, à température constante. L'état le plus stable pour une molécule d'ARN dans ce contexte est la structure dont l'énergie libre est minimale. L'approche thermodynamique pour la prédiction de structures d'ARN consiste par conséquent, étant donnée une séquence d'ARN, à trouver un repliement de cette molécule dont l'énergie libre est minimale.

Ce contexte de travail s'accompagne de plusieurs choix et contraintes. La première limite est que la quasi totalité des approches existantes se restreignent à la prédiction de structures secondaires, sans pseudonœuds, pour des raisons de complexité algorithmique et de paramétrage du modèle thermodynamique. On suppose en effet que les interactions tertiaires sont plus faibles que les interactions secondaires et que la somme des énergies libres des éléments de structure secondaire constitue une approximation raisonnable de l'énergie libre totale. Par nature, les pseudonœuds sont constitués d'appariements entre des bases relativement éloignées, qui complètent en général une structure secondaire déjà stable. La formation de pseudonœuds peut altérer la structure secondaire par quelques remaniements, mais ne la modifie qu'à titre exceptionnel en une structure radicalement différente.

Une seconde limite est que les interactions potentielles avec d'autres molécules telles que des protéines ou d'autres acides nucléiques ne sont pas prises en compte, bien que celles-ci puissent concourir à la stabilité de la structure d'un ARN.

Enfin, l'existence sur le papier de plusieurs solutions avec des niveaux d'énergie proches de l'optimal oblige à considérer un ensemble de solution potentielles, et non une unique solution. Il faut donc raisonner en termes de solution sous-optimales, et pas simplement optimales.

L'algorithme de Nussinov et Jacobson

La première approche pour la prédiction de structures d'ARN suivant un modèle thermodynamique a été introduite en 1978 par Nussinov [NPGK78]. Entant donnée une séquence d'ARN, il s'agit de trouver une structure secondaire où le nombre d'appariements est maximal. Nussinov et Jacobson [NJ80] ont ensuite proposé une adaptation de cette méthode pour intégrer un modèle énergétique simple où l'énergie libre d'une structure secondaire est obtenue en sommant la contribution énergétique (négative) des appariements individuels. Dans

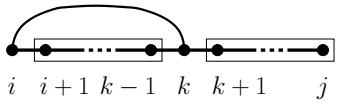
les deux cas, le problème se résout par programmation dynamique. Le calcul de la structure d'énergie libre minimale se décompose ainsi en deux étapes : le remplissage de la table de programmation dynamique afin de calculer l'énergie libre minimale atteignable, puis la reconstruction d'une structure optimale par remontée dans la matrice.

Soit une séquence d'ARN $s = s[1..n]$, c'est-à-dire un mot de longueur n sur l'alphabet $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{U}\}$. On définit une matrice carrée E de $n \times n$ cellules. Chaque cellule $E(i,j)$ de la matrice E correspond à l'énergie libre de la structure d'énergie libre minimale de la sous-séquence $s[i..j]$, avec $i \leq j$, du mot s. Le remplissage de la matrice E s'effectue en suivant la relation suivante, illustrée de manière schématique sur la figure 3.1

$$E(i,j) = \min \begin{cases} E(i+1,j) \\ \min\limits_{i<k<j} \{E(i+1,k-1) + E(k+1,j) + \alpha(i,k)\} \end{cases}$$

où $\alpha(i,k)$ correspond à la contribution énergétique de l'appariement formé entre le nucléotide i et le nucléotide j. En pratique, $\alpha(i,k) < 0$ si les nucléotides aux positions i et k forment un appariement canonique et si $k - i > 3$, $\alpha(i,k) = +\infty$ dans les autres cas.

(a) La base i est libre, l'énergie libre de la séquence $s[i..j]$ est alors celle de la séquence $s[i+1..j]$.



(b) La base i est appariée avec une certaine base k, l'énergie libre de la séquence $s[i..j]$ est alors la somme des énergies associées à l'appariement de i avec k et aux séquences $s[i+1..k-1]$ et $s[k+1..j]$.

FIG. 3.1 – Les récurrences de Nussinov et Jacobson présentées de manière schématique.

Par construction, l'énergie libre de la structure d'énergie libre minimale de s se trouve dans la cellule $E(1,n)$. Pour reconstruire une structure optimale associée à cette valeur d'énergie libre, on remonte la matrice en partant de la cellule $E(1,n)$ afin de retracer le chemin suivi dans la matrice pour obtenir cette valeur. La complexité spatiale de l'algorithme est en $\mathcal{O}(n^2)$ à cause du stockage de la matrice carrée E. Chaque cellule de la matrice nécessite un calcul en temps linéairement proportionnel à longueur de la sous-séquence correspondante. Cet algorithme a donc une complexité temporelle en $\mathcal{O}(n^3)$.

Bien que cette modélisation soit fortement limitée, l'algorithme défini par Nussinov et Jacobson est la base de la plupart des algorithmes de prédiction de structures d'ARN qui visent à déterminer une structure d'énergie libre minimale.

L'algorithme de Zuker

La modélisation adoptée par l'algorithme de Nussinov et Jacobson ne prend pas en compte de nombreux éléments qui contribuent à stabiliser une structure, comme les empilements d'appariements, ou à la déstabiliser, comme les régions non appariées.

L'algorithme proposé par Zuker [ZS81, JTZ89, JTZ90] est une extension de l'algorithme de Nussinov et Jacobson pour le modèle d'énergie plus réaliste de Freier-Turner [TSF88, MSZT99, MDC+04] : empilements d'appariements pour former des tiges, boucles terminant

une tige, branchements multiples,... La figure 3.2 montre une partie des éléments pris en compte dans ce modèle d'énergie.

Quatre matrices $F, C, MetM^1$ sont nécessaires au découpage d'une structure dans ce modèle, comme illustré en figure 3.3. Pour une séquence $s = s[1..n]$,

- $F(i,j)$ correspond à l'énergie libre de la structure d'énergie libre minimale de la sous-séquence $s[i..j]$;
- $C(i,j)$ correspond à l'énergie libre de la structure d'énergie libre minimale de la sous-séquence $s[i..j]$ où l'appariement entre les nucléotides aux positions i et j est forcé ;
- $M(i,j)$ correspond à l'énergie libre de la structure d'énergie libre minimale de la sous-séquence $s[i..j]$ sachant que cette sous-séquence fait partie d'un embranchement multiple comportant au moins une composante, c'est-à-dire une structure quelconque fermée un appariement ;
- $M^1(i,j)$ correspond à l'énergie libre de la structure d'énergie libre minimale de la sous-séquence $s[i..j]$ sachant que cette sous-séquence fait partie d'un embranchement multiple comportant exactement une composante.

Les récurrences établies par Zuker sont les suivantes.

$$F(i,j) = \min \begin{cases} F(i+1,j) \\ \min\limits_{i<k<j} \{C(i,k) + F(k+1,j)\} \end{cases}$$

$$C(i,j) = \min \begin{cases} \mathcal{H}(i,j) \\ \min\limits_{i<k<l<j} \{C(k,l) + \mathcal{I}(i,j,k,l)\} \\ \min\limits_{i<u<j} \{M(i+1,u) + M^1(u+1,j-1) + a\} \end{cases}$$

$$M(i,j) = \min \begin{cases} \min\limits_{i<u<j} \{(u-i+1).c + C(u+1,j) + b\} \\ \min\limits_{i<u<j} \{M(i,u) + C(u+1,j) + b\} \\ M(i,j-1) + c \end{cases}$$

$$M^1(i,j) = \min \begin{cases} M^1(i,j-1) + c \\ C(i,j) + b \end{cases}$$

où $\mathcal{H}(i,j)$ est l'énergie d'une boucle terminale fermée par l'appariement entre les nucléotides en position i et j, $\mathcal{I}(i,j,k,l)$ est l'énergie d'une boucle interne formée des deux sous-séquences $s[i..j]$ et $s[k..l]$ et où les variables a, b et c sont des constantes qui proviennent du modèle d'énergie linéaire des embranchements multiples, à savoir que l'énergie d'un embranchement multiple de degré degree et de longueur size est $E = a + b.\mathsf{degree} + c.\mathsf{size}$. Le degré d'un embranchement multiple est le nombre de sous-séquences non appariées qui séparent ses composantes. La longueur d'un embranchement multiple est la somme des longueurs de ses régions non appariées, comme illustré sur la figure 3.2.

L'algorithme de Zuker a une complexité temporelle en $\mathcal{O}(n^4)$ et spatiale en $\mathcal{O}(n^2)$. Historiquement, on compte deux implémentations strictes de l'algorithme de Zuker : MFOLD [Zuk89] et RNAFOLD [HFS+94]. Ces deux logiciels sont les plus utilisés pour la prédiction de structures secondaires. Toutefois, une complexité temporelle en $\mathcal{O}(n^3)$ de l'algorithme a pu être atteinte grâce à un traitement différent des boucles internes basé sur une fonction de coût concave ou convexe [LZP99]. Des travaux récents de Roytberg *et al* ont permis d'améliorer

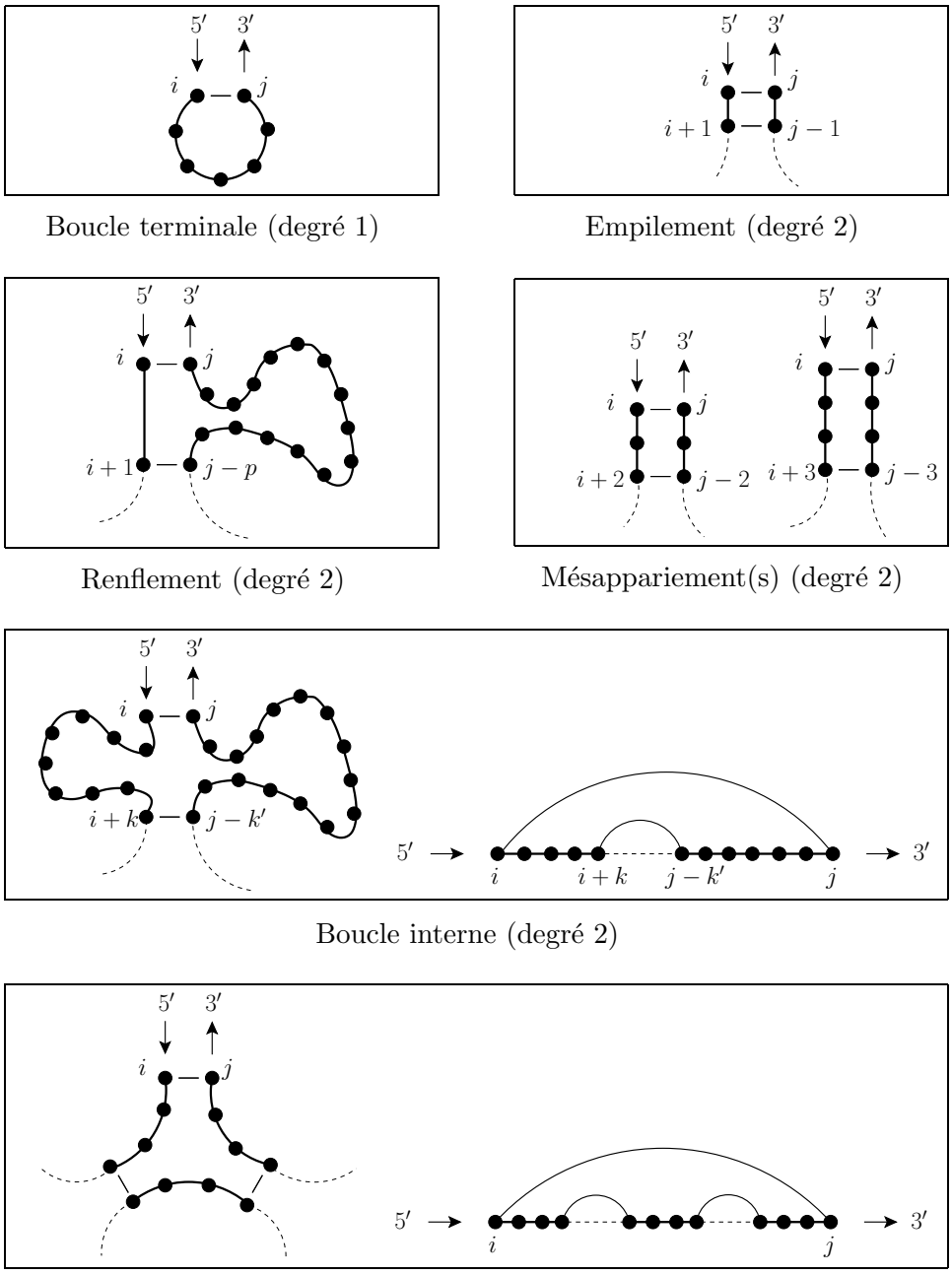
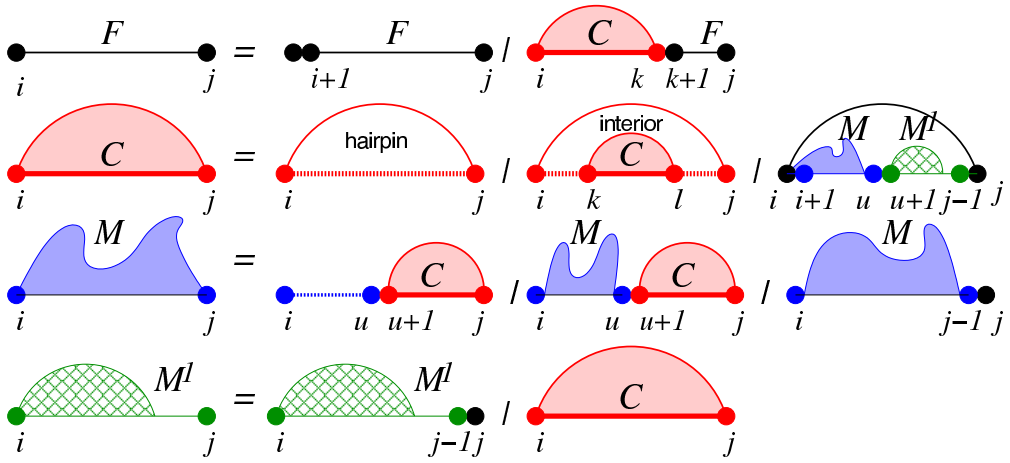


FIG. 3.2 – Classification des composantes fermées par un appariement. Les empilements, renflements et mésappariements sont des cas particuliers de boucles internes.



Source http://www.zbit.uni-tuebingen.de/pas/EMBO-RNACourse/handouts/HandoutBook.pdf

FIG. 3.3 – Schématisation des règles de décomposition appliquées dans l'algorithme de Zuker. Les arcs pleins correspondent à des appariements entre des bases reliées. Les traits discontinus indiquent des régions qui ne contiennent aucun appariement.

encore la complexité temporelle de l'algorithme en prenant en plus en compte les informations du modèle thermodynamique pour l'évaluation des boucles internes [OSKR06]. Leur algorithme a une complexité temporelle en $\mathcal{O}(n^2 \log^2 n)$. On compte plusieurs variantes heuristiques de cet algorithme, notamment RDFOLDER [YLLL04] basé sur les simulations de type Monte-Carlo, et SARNA-PREDICT [TW06, TW07] où la technique retenue est le recuit simulé.

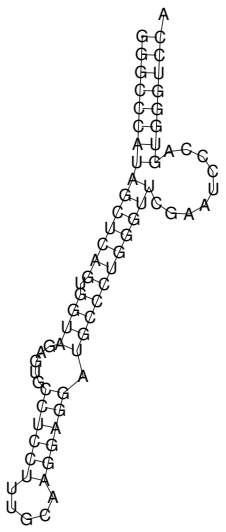
Plusieurs études montrent que la structure d'énergie libre minimale ne correspond pas toujours à la conformation adoptée par la molécule dans la cellule [ZS81, TSF88, JTZ89, JTZ90, MSZT99]. Par exemple, la structure optimale prédite par l'algorithme de Zuker de certaines séquences d'ARN de transfert n'est pas la structure secondaire correcte en feuille de trèfle. Sur la figure 3.4 sont représentées la structure optimale et la structure correcte d'un ARN de transfert qui se replie mal. L'énergie libre de la structure correcte est ici légèrement plus élevée. Cette structure fait partie des structures sous-optimales.

RNASUBOPT [WFHS99], basé sur RNAFOLD, permet l'énumération de toutes les structures sous-optimales distantes de l'optimale d'une certaine quantité d'énergie. Cependant, parmi les structures sous-optimales beaucoup ne diffèrent que de quelques appariements et il faut donc manuellement rechercher des structures dont l'aspect global est radicalement différent. MFOLD propose par défaut en plus de la structure d'énergie libre minimale, une sélection de structures sous-optimales ayant un aspect général différent. Cette idée a été cristallisée de manière plus formelle dans le logiciel RNASHAPES [SVR+06]. Ce logiciel propose plusieurs niveaux d'abstraction des tiges, avec ou sans renflement, avec ou sans mésappariements, ... Il utilise une représentation sous forme parenthésée des tiges détectées pour représenter les structures et sélectionne ainsi des structures différentes, selon le niveau d'abstraction choisi, parmi les résultats produits par RNASUBOPT. Ainsi, pour l'exemple de l'ARN de transfert présenté en figure 3.4, les trois résultats produits par RNASHAPES avec un delta d'énergie de -5 kcal/mol sont les suivants

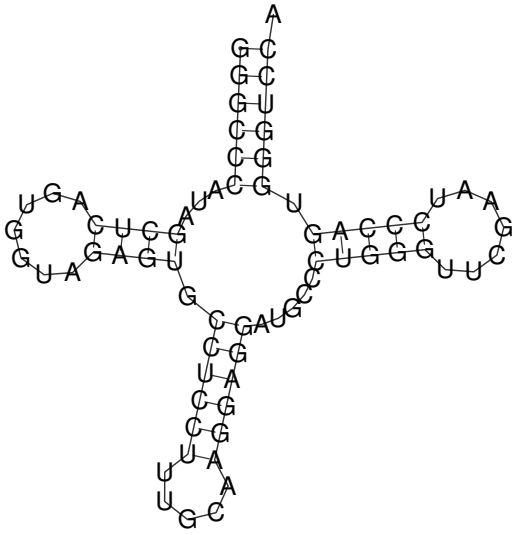
```
Shape       GGGCCCAUAGCUCAGUGGUAGAGUGCCUCCUUUGCAAGGAGGAUGCCCUGGGUUCGAAUCCCAGUGGGUCCA
[]          ((((((((((((((.((((.....(((((((...)))))))..)))))))))))).........)))))))).  -35.9 kcal/mol
[][]]       ((((((((.....((.((((.....(((((((...)))))))..))))))).((((.....))))..)))))))).  -32.2 kcal/mol
[][][]]     ((((((...((((.......)))).(((((((...)))))))).....(((((.......)))))..))))))).  -31.7 kcal/mol
```

GGGCCCAUAGCUCAGUGGUAGAGUGCCUCCUUUGCAAGGAGGAUGCCCUGGGUUCGAAUCCCAGUGGGUCCA

(a) Structure primaire d'un ARN de transfert associé à l'alanine provenant du génome de *Natronobacterium pharaonis* (AB003409.1)



(b) Structure secondaire d'énergie libre minimale (−35.9 kcal/mol) prédite par RNAfold



(c) Structure secondaire sous-optimale prédite par RNAfold (−31.7 kcal/mol) qui correspond à la structure secondaire réelle de l'ARN de transfert

Fig. 3.4 – La structure secondaire de gauche correspond à la structure d'énergie libre minimale prédite pour un ARN de transfert par l'algorithme de Zuker et le modèle d'énergie de Freier-Turner. La structure de droite est la structure secondaire réelle des ARN de transfert.

L'algorithme de Zuker a été étendu par Eddy et Rivas [RE99] afin de permettre la prédiction de structures tertiaires d'ARN incluant des pseudonœuds. Ces derniers ont ainsi pu montrer que la prédiction de structures tertiaires est un problème dont la complexité temporelle en $\mathcal{O}(n^6)$ est quasiment impraticable sur des séquences de plus d'une centaine de bases. Toutefois, en restreignant l'investigation à certaines classes de pseudonœuds, des solutions algorithmiques exactes et efficaces ont été proposées telles que Pknots [RE99] et Pknots-RG [RSG07]. Le modèle énergétique sous-jacent à la formation des pseudonœuds reste néanmoins trop flou pour permettre d'établir des prédictions aussi pertinentes que pour les structures secondaires.

Bon nombre d'approches se sont inspirées de l'algorithme de Zuker en utilisant d'autres modèles énergétiques pour la prédiction de structures secondaires. CONTRAfold [DWB06] est une méthode qui adopte un modèle d'énergie obtenu par apprentissage sur un ensemble de séquences annotées par leur structure connue et vérifiée. Une méthode plus récente, MC-fold [PM08], adopte un schéma différent du modèle d'énergie de Turner en utilisant une base de motifs identifiés *in silico* sur des structures tertiaires d'ARN, les NCM, acronyme pour *Nucleotide Cyclic Motifs*. Plusieurs structures jusqu'ici incorrectement prédites grâce

au modèle de Turner ont ainsi pu être prédites par ce logiciel. Cependant, la recherche de ces motifs particuliers et leur assemblage pour former une structure secondaire a un coût algorithmique non négligeable qui limite son utilisation à des séquences relativement courtes.

La fonction de partition

L'approche thermodynamique peut être abordée sous un autre angle avec la *fonction de partition*. Le but n'est alors plus de minimiser l'énergie libre mais de maximiser la probabilité d'une structure donnée d'ARN connaissant l'ensemble des structures que la séquence peut adopter et la probabilité de formation d'un appariement dans ce contexte. Selon les principes de la thermodynamique, la probabilité d'une structure Ψ dans un système équilibré est proportionnelle à son facteur de Boltzmann

$$\exp\left(\frac{-E(\Psi)}{RT}\right)$$

où $E(\Psi)$ est l'énergie libre de la structure Ψ, R est la constante du gaz parfait (en Joules/(Kelvin mol)), T est la température absolue (en Kelvin). L'ensemble des structures est déterminée par la fonction de partition notée Z. Cette fonction est une grandeur fondamentale qui englobe les propriétés statistiques d'un système à l'équilibre thermodynamique. Le système considéré ici étant l'ensemble des structures secondaires possibles, la fonction de partition est définie par

$$Z = \sum_{\Psi} \exp\left(\frac{-E\left(\Psi\right)}{RT}\right)$$

Grâce à cette fonction, on peut déterminer la probabilité d'une structure Ψ dans l'ensemble des structures possibles considérées :

$$p(\Psi) = \frac{\exp\left(\frac{-E(\Psi)}{RT}\right)}{Z}$$

Cette approche pour calculer la probabilité d'une structure nécessite de calculer la fonction de partition complète. Directement, ce calcul est impraticable car il demande de calculer l'énergie libre de toutes les structures secondaires possibles dont le nombre croît de manière exponentielle en fonction de la longueur de la séquence [WFHS99]. Grâce aux travaux de McCaskill [McC90], la fonction de partition peut être calculée de manière partielle et récursivement par programmation dynamique avec une complexité spatiale en $\mathcal{O}(n^2)$ et temporelle en $\mathcal{O}(n^3)$. Pour expliquer l'idée mise en œuvre par McCaskill, on se place dans le cas du modèle énergétique simple utilisé dans l'algorithme de Nussinov et Jacobson où l'énergie d'un structure est obtenue en sommant les contribution des appariements individuels. Soit $Z(i,j)$ la fonction de partition pour toutes les structures de la séquence $s[i..j]$

$$Z(i,j) = Z(i+1,j) + \sum_{k} Z(i+1,k-1)Z(k+1,j)\exp\left(\frac{-\alpha(i,j)}{RT}\right)$$

Cette formule peut être obtenue en transformant l'équation de récurrence utilisée dans l'algorithme de Nussinov et Jacobson présentée à la page 58 en remplaçant les opérations de minimisation par des sommes, les sommes par des multiplications, et les énergies par les

facteurs de Boltzmann correspondants. L'intérêt de cette décomposition est également de pouvoir en dériver le calcul de la probabilité de la formation d'un appariement entre deux nucléotides i et j.

$$p(i,j) = \sum_{(i,j)\in\Psi} p(\Psi)$$

Toujours grâce aux travaux de McCaskill, cette probabilité peut être calculée récursivement grâce à la relation suivante

$$p(i,j) = \hat{Z}(i,j)Z(i+1,j-1)\frac{\exp\left(\frac{-\alpha(i,j)}{RT}\right)}{Z}$$

où $\hat{Z}(i,j)$ est la fonction de partition de l'ensemble des structures qui ne font pas intervenir la séquence $s[i..j]$.

Il existe plusieurs implémentations de la fonction de partition pour la prédiction de structure secondaire : RNAfold, Sfold [DL99, DCL04] et une implémentation pour machines massivement parallèles [FHS00]. Bien que l'algorithme de prédiction "classique" et la fonction de partition apportent des résultats équivalents en terme de prédiction de structure secondaire [HGK97], la fonction de partition ouvre d'autres possibilités applicatives telle que l'échantillonnage de structures et de motifs structuraux [DL01, DL03].

3.1.2 La prédiction par analyse comparative

L'analyse comparative aborde le problème de la prédiction de structure lorsque l'on dispose de plusieurs séquences homologues. Dans ce contexte, le gain d'information apporté par l'utilisation de plusieurs séquences est double. Des ARN non-codants homologues partagent une fonction induite par une structure commune mieux conservée que leur structure primaire durant l'évolution. Lorsque l'on dispose de plusieurs séquences dont on suppose qu'elles partagent une fonction liée à leur structure, il est donc naturel de rechercher leur structure commune, et les programmes de prédiction de structures communes sont donc plus fiables et plus robustes que les programmes de prédiction de structures à partir d'une seule séquence. De plus, les prédictions de structures communes peuvent être confortées par la présence de mutations compensatoires induites par la conservation d'une structure au cours de l'évolution (section 1.4.4). La figure 3.5 montre un exemple de mutations compensatoires sur un alignement de sept séquences d'ARN de transfert.

On distingue deux approches dans les méthodes de prédiction par analyse comparative : celles qui recherchent une structure commune sur des séquences préalablement alignées, et celles qui travaillent sur des séquence non alignées. Ces deux approches sont complémentaires et le choix d'une approche plutôt que l'autre dépend principalement du degré de conservation des séquences à replier (section 1.5). Un approche plus récente et toute aussi originale, RNAcast [RG05], combine les prédictions individuelles de RNAshapes pour détecter parmi les structures sous-optimales individuelles une structure globalement conservée selon son aspect général.

Aligner et replier simultanément, l'algorithme de Sankoff

Une première manière d'exploiter l'information contenue dans un ensemble de séquences est de rechercher à replier simultanément toutes les séquences. L'algorithme de Sankoff est

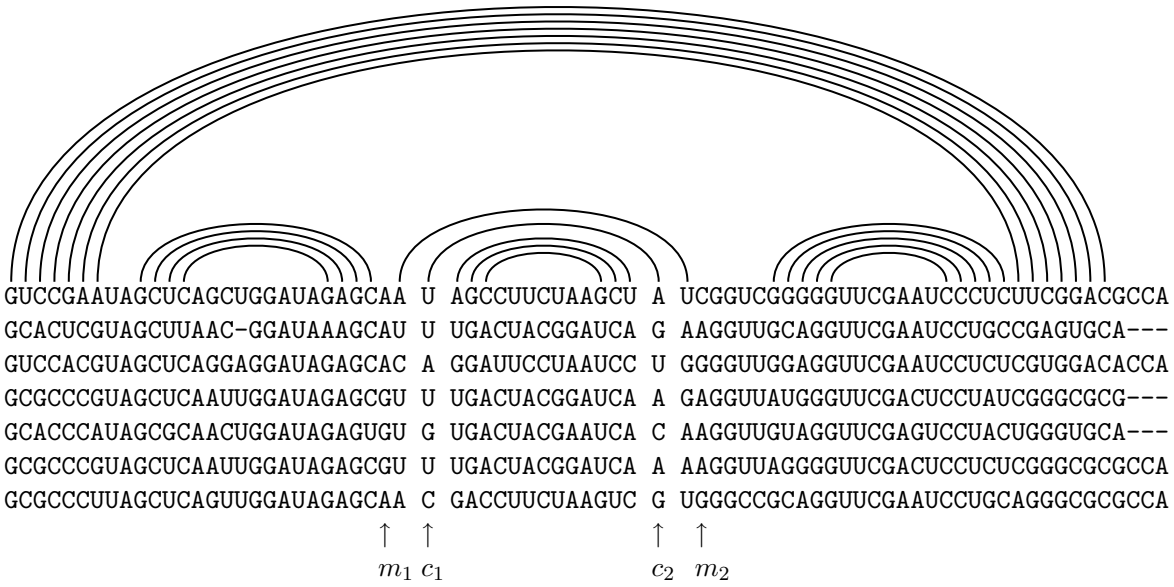


FIG. 3.5 – Alignement de sept séquences d'ARN de transfert, avec représentation de la structure commune. Les deux colonnes isolées c_1 et c_2 participent à un même appariement et font apparaître des mutations compensatoires.

l'algorithme de référence pour ce problème [ZS84, San85]. Il procède au repliement simultané de deux séquences par programmation dynamique et produit donc par la même un alignement de ces séquences. Par souci de clarté et de lisibilité, nous allons présenter la version de l'algorithme de Sankoff basée sur le modèle énergétique utilisé par Nussinov et Jacobson plus simple à appréhender que la version basée sur le modèle énergétique de Turner où la multiplicité des décompositions rend les relations rapidement illisibles. Toutes les remarques effectuées par la suite restent néanmoins valables, en particulier les notions de complexité.

Soient deux séquences d'ARN $s_1 = s_1[1..n]$ et $s_2 = s_2[1..m]$ et une matrice S de dimension $n \times n \times m \times m$. Chaque cellule $S(i,j,k,l)$ de la matrice S correspond à l'énergie libre minimale du repliement commun des sous-séquences $s_1[i..j]$ et $s_2[k..l]$. Le remplissage de la matrice s'effectue suivant la relation

$$
S(i,j,k,l) = \min \begin{cases}
S(i+1,j,k,l) \\
S(i,j,k+1,l) \\
\min\limits_{i<p<j} \{S(i+1,p-1,k,l) + S(p+1,j,0,0) + \alpha(s_1[i], s_1[p])\} \\
\min\limits_{i<p<j} \{S(i+1,p-1,0,0) + S(p+1,j,k,l) + \alpha(s_1[i], s_1[p])\} \\
\min\limits_{k<q<l} \{S(i,j,k+1,q-1) + S(0,0,q+1,l) + \alpha(s_2[k], s_2[q])\} \\
\min\limits_{k<q<l} \{S(0,0,k+1,q-1) + S(i,j,q+1,l) + \alpha(s_2[k], s_2[q])\} \\
\min\limits_{\substack{i<p<j \\ k<q<l}} \{S(i+1,p-1,k+1,q-1) + S(p+1,j,q+1,l) + \alpha'(s_1[i], s_1[p], s_2[q], s_2[k])\}
\end{cases}
$$

où $\alpha(s_1[i], s_1[j])$ correspond à la contribution énergétique apportée par l'appariement entre $s_1[i]$ et $s_1[j]$ telle qu'elle est utilisée dans l'algorithme de Nussinov et Jacobson, et

$\alpha'(s_1[i], s_1[j], s_2[k], s_2[l])$ correspond à la contribution énergétique apportée par l'appariement conjoint entre $s_1[i]$ et $s_1[j]$ d'une part, et $s_2[k]$ et $s_2[l]$ d'autre part. La définition la plus naturelle pour α' consiste à prendre la somme des contributions énergétiques des appariements individuels

$$\alpha'(s_1[i], s_1[j], s_2[k], s_2[l]) = \alpha(s_1[i], s_1[j]) + \alpha(s_2[k], s_2[l])$$

On peut introduire un facteur bonifiant les appariements conjoints afin de favoriser les corepliements en présence de mutations, particulièrement en présence de mutations compensatoires. Un malus peut également être considéré en cas d'introduction d'insertion ou de délétion, c'est-à-dire dans les deux premières règles.

L'algorithme de Sankoff a une complexité temporelle en $\mathcal{O}(n^3m^3)$ et une complexité spatiale en $\mathcal{O}(n^2m^2)$. Cet algorithme peut être étendu à plus de deux séquences. Pour N séquences, sa complexité temporelle est alors en $\mathcal{O}(l^{3N})$ et sa complexité spatiale en $\mathcal{O}(l^{2N})$, où l est la longueur de la plus longue des séquences traitées.

La complexité élevée de l'algorithme de Sankoff, même sur deux séquences, le rend impraticable sur des séquences qui dépassent la centaine de nucléotides. Il existe cependant de nombreuses déclinaisons de cet algorithme qui tentent de traiter ce problème. Seule une partie d'entre elles sont présentées, les plus originales. FoldAlign [HLG05, HTG07] implante une version où les embranchements multiples sont interdits et seules les tiges terminées par une boucle terminale sont considérées. La restriction appliquée dans Dynalign [MT02, HSM07] est une borne maximale sur la distance qui sépare des nucléotides alignés, ce qui permet de restreindre l'exploration de la matrice S à son hyperdiagonale. Dans Consan [DE06] et Stemloc [Hol05], des régions fortement conservées sont identifiées entre les séquences pour produire un alignement grossier des séquences, ce qui permet de contraindre la formation d'appariements respectueux de cet alignement.

Toujours inspirées de l'algorithme de Sankoff, d'autres heuristiques s'attachent à la prédiction de structures communes à plus de deux séquences. FoldAlignM [THG07] réalise un alignement global de manière progressive à partir des alignements deux à deux générés par FoldAlign pour produire une structure globalement conservée. Cette manière de passer de deux à plus de séquences est fortement inspirée de PMcomp/PMmulti [HBS04], nouvellement LocaRNA [WRH+07], également repris dans StrAl [DWMS06]. Dans ces logiciels, les repliements deux à deux sont réalisés par comparaison des matrices de probabilité d'appariement produite à l'aide de la fonction de partition. Murlet [KTKA07] réalise de manière itérative une alignement multiple des séquences à partir des repliements deux à deux calculés par l'algorithme de Sankoff restreint à la manière de Consan, c'est-à-dire en établissant un alignement préliminaire entre les séquences. Cette idée d'alignement préliminaire est reprise dans MxscaRNA [TTKA06, TKKA08] où cette fois seules les parties ouvrantes et fermantes des tiges sont alignées. L'alignement par morceaux ainsi obtenu est ensuite utilisé pour inférer une structure globale.

Aligner puis replier

Nous avons vu que l'algorithme de Sankoff et ses déclinaisons permettent d'inférer une structure commune pour un ensemble de séquences non alignées. On peut également aborder le problème en commençant par aligner les séquences sur la base de la structure primaire, puis en cherchant une structure commune compatible avec l'alignement. Cela présente deux

avantages : la complexité algorithmique est moindre, et on peut améliorer la prédiction en utilisant la présence de mutations compensatoires, observables directement entre les couples de positions de l'alignement.

La corrélation des colonnes, l'information mutuelle La mesure de corrélation des colonnes la plus utilisée est tirée de la théorie de l'information de Shannon : elle mesure l'information mutuelle entre deux colonnes [CK91]. Etant donné un alignement multiple, $f_i(x)$ désigne la fréquence d'apparition du nucléotide x dans la colonne i de l'alignement et $f_{ij}(x, y)$ la fréquence d'apparition du couple de nucléotides (x, y) dans les colonnes i et j. L'information mutuelle des deux colonnes i et j est définie par

$$M_{ij} = \sum_{x,y} f_{ij}(x, y) \log_2 \left(\frac{f_{ij}(x, y)}{f_i(x) f_j(y)} \right)$$

La valeur de M_{ij} varie entre 0 et 2 bits et mesure le degré de corrélation des deux colonnes. Elle est maximale lorsque les deux colonnes sont parfaitement corrélées, c'est-à-dire qu'un appariement est totalement absent ou conservé, et que leur contenu individuel est pourtant totalement aléatoire, c'est-à-dire que toutes les nucléotides apparaissent de manière équiprobable. M_{ij} est nulle en l'absence de mutation dans les deux colonnes ou lorsque les colonnes varient de façon indépendante, c'est-à-dire $f_{ij}(x, y) = f_i(x) f_j(y)$. Des corrections peuvent être apportées à cette mesure pour prendre en compte la composition globale des séquences ou encore un arbre phylogénétique pour prendre en compte le taux de mutations attendues par colonne [KH99, GHH$^+$94]. Sur l'exemple de la figure 3.5, l'information mutuelle des deux colonnes appariées c_1 et c_2 qui présentent des mutations compensatoires est de 1,37 bit, alors qu'entre les colonnes non appariées m_1 et m_2 cette valeur est de 0,52 bit.

L'information mutuelle constitue la noyau de bon nombre de méthodes de prédiction de structure à partir d'un alignement. La plupart de ces méthodes utilisent cette information comme score soit à la place de l'information énergétique pour les méthodes à base de l'algorithme de Zuker, soit comme bonus ou malus complémentaire à une approche énergétique.

RNAalifold [HFS02] est actuellement la méthode plus utilisée pour la prédiction de structure secondaire à partir d'un alignement, car cette méthode est celle qui exploite le modèle énergétique de Turner. Son fonctionnement repose sur l'algorithme de Zuker généralisé à un alignement. La contribution énergétique d'un appariement entre deux colonnes est simplement calculée en moyennant les contributions individuelles. Un bonus est appliqué en fonction de la corrélation entre les colonnes appariées. Ilm [RSZ04a, RSZ04b] est une méthode strictement analogue à RNAalifold où l'algorithme de repliement adapté à l'alignement multiple est celui de Nussinov et Jacobson. Plus récemment, RNAlishapes [Vos06] est en fait une variante de RNAalifold où l'algorithme de repliement est une version modifiée de RNAshapes : le repliement est effectué entre des représentations abstraites des structures optimales et sous-optimales prédites individuellement. L'algorithme travaille ainsi au niveau des tiges en tentant de faire correspondre les parties ouvrantes d'une même tige conservée d'une part, et les parties fermantes correspondantes d'autre part. Cove [ED94], Pfold [KH99, KH03] et CMFinder [YWR06] sont trois méthodes à base de grammaire stochastiques hors contexte entraînées sur des séquences exemples et où l'information mutuelle mesurée entre les colonnes est utilisée comme pondération des informations apprises. Pfold présente toutefois une originalité supplémentaire : la mesure de l'information mutuelle peut

être affinée pour tenir compte de la distance évolutive qui sépare les espèces dont sont issues les séquences.

P-DCfold [TGR02, TER03, **?**, Eng06, ET07] propose une alternative originale aux méthodes citées précédemment. Le logiciel commence par chercher des séquences palindromiques conservées et alignées qui exhibent une ou plusieurs mutations compensatoires. Puis en appliquant une heuristique gloutonne, il construit successivement des ensembles de palindromes tous compatibles entre eux, c'est-à-dire sans croisement ni chevauchement. Cette approche de type "diviser pour régner" permet d'une part de réduire de manière drastique la complexité du repliement et d'autre part d'autoriser la formation de pseudonœuds.

3.1.3 BRAliBase I, le benchmark de référence

En 2004, Gardner propose BRAliBase I [GG04], un benchmark pour évaluer les méthodes de prédiction de structure. Dans un premier temps, nous présentons les données et les critères sur lesquels sont évaluées les méthodes. Par la suite nous présentons les résultats des méthodes testées dans BRAliBase I parmi lesquelles figure caRNAc, notre méthode de prédiction de structure dont nous parlons plus en détails dans la section 3.3.

Description des données

BRAliBase I contient quatre familles d'ARN non-codants : des ARN ribosomiques, petite sous-unité et grosse sous-unité, des ARN de transfert et des ARN de RNase P. Chaque famille est divisée en deux groupes dont les caractéristiques sont rappelées dans la table 3.1 : un groupe de séquences moyennement conservées (medium) dont l'identité moyenne est comprise entre 60% et 80%, et un autre groupe de séquences bien conservées (high) dont l'identité moyenne est supérieure à 80%. Pour chaque groupe, la structure correcte d'une séquence est donnée pour permettre l'évaluation des prédictions réalisées. Ces structures sont déduites d'alignements multiples construits manuellement accompagnés des structures individuelles vérifiées provenant de la littérature.

Jeu de	Longueur	Identité moyenne (%)		Nb. séq.	
données	moyenne	medium	high	medium	high
LSU	2904	72,0	88,1	11	11
SSU	1542	80,0	90,7	11	11
RNaseP	377	67,1	81,5	11	9
tRNA	73	60,0	84,4	11	11

Tab. 3.1 – Description générale des jeux de données de BRAliBase I.

Evaluation de la qualité des structures prédites

Les structures prédites sont évaluées par rapport à la structure de référence de chaque groupe grâce à trois mesures : la sensibilité, la spécificité, le cœfficient de corrélation de Matthews (MCC). La spécificité est calculée de la manière suivante

$$S_p = \frac{TP}{TP + (FP - \xi)}$$

où TP est le nombre de vrais positifs, c'est-à-dire le nombre d'appariements correctement prédits, et FP est le nombre de faux positifs, c'est-à-dire le nombre d'appariements prédits qui n'existent pas dans la structure de référence. Les faux positifs sont ici séparés en trois groupes : les appariements inconsistants, les appariements contrariant et les appariements compatibles. Un appariement prédit entre deux bases i et j est inconsistant si, et seulement si, i ou j est appariée avec une autre base dans la structure de référence. Un appariement prédit entre deux bases i et j est contrariant si, et seulement si, il existe un appariement entre deux bases k et l dans la structure de référence tel que $k < i < l < j$, c'est-à-dire que l'ajout de l'appariement entre les bases i et j dans la structure de référence produit un pseudonœud. Enfin, un appariement prédit entre deux bases i et j est compatible si, et seulement si, il n'est pas inconsistant ou contrariant. Le paramètre ξ présent dans le calcul de la spécificité désigne le nombre d'appariements compatibles.

Le cœfficient de corrélation de Matthews peut être vu comme une mesure synthétisant la sensibilité et la spécificité, dont la définition adaptée dans BRAliBase I est

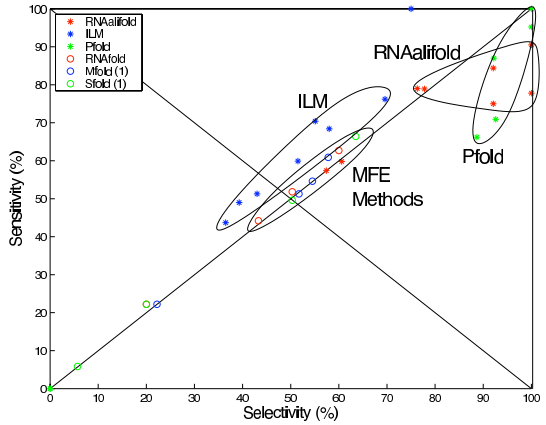
$$MCC = \frac{TP \times FN - (FP - \xi) \times FN}{\sqrt{(TP + (FP - \xi))(TP + FN)(TN + (FP - \xi))(TN + FN)}}$$

Résultats des méthodes de prédiction testées

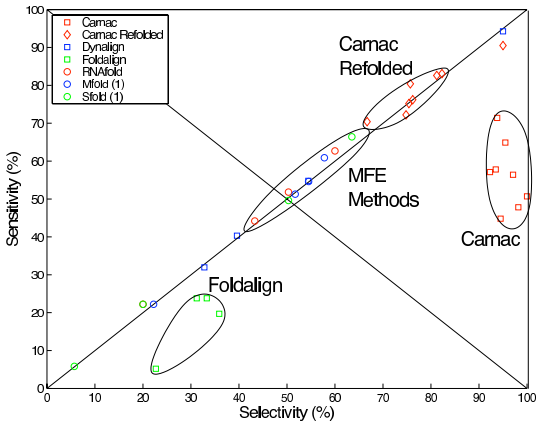
La figure 3.6 présente les résultats de BRAliBase I provenant de l'article de Gardner [GG04]. Globalement, les méthodes qui travaillent sur des ensembles de séquences, quelque soit leur approche, se distinguent nettement des approches purement thermodynamiques. Parmi les méthodes les plus performantes, trois méthodes se dégagent significativement : RNAalifold, Pfold et caRNAc. Les auteurs de BRAliBase I ont remarqué que caRNAc était une méthode particulièrement spécifique qui avait tendance à prédire moins d'appariements que les autres méthodes, mais qu'en contrepartie les appariements prédits étaient plus fiables. Ils ont donc proposé un protocole pour compléter les structures prédites par caRNAc. Pour cela ils effectuent un repliement purement thermodynamique avec RNAfold d'une séquence de chaque jeu de données où ils contraignent RNAfold d'intégrer tous les appariements prédits par caRNAc. Ce protocole permet d'améliorer nettement la sensibilité de caRNAc, sans altérer sa spécificité. Dans cette configuration, les résultats de caRNAc sont comparables à ceux de RNAalifold et Pfold qui, rappelons le, travaillent sur séquences alignées. De plus, les résultats de caRNAc sont beaucoup plus stables d'un jeu de données à un autre avec une sensibilité et une spécificité minimum supérieure à 70%, contre moins de 60% pour RNAalifold.

3.2 La prédiction de gènes à ARN

Dans le chapitre précédent, nous avons vu qu'il existait diverses approches pour la prédiction de gènes à protéines ou plus généralement de régions codantes : les approches *ab initio* (section 2.1), les approches par homologie (section 2.2), et les approches comparatives (section 2.3). Des approches analogues sont disponibles pour la prédiction de gènes à ARN. Chronologiquement, le schéma adopté montre également une certaine symétrie. Les premières investigations menées ont porté sur la recherche de biais de composition dans les gènes à ARN dont le perfectionnement a donné lieu à des approches de prédiction *ab initio*. En parallèle, plusieurs méthodes de recherche par homologie avec des séquences connues contenues dans les



(a) Résultats des approches qui suivent le paradigme "Aligner puis replier" comparés aux approches thermodynamiques.



(b) Résultats des approches qui suivent le paradigme "Aligner et replier simultanément" comparés aux approches thermodynamiques.

FIG. 3.6 – Résultats de BRALIBASE I présentés par type d'approche en fonction de la spécificité, en abscisse, et de la sensibilité, en ordonnée.

banques de données se sont développées. Par la suite, les approches comparatives ont fait leur apparition, intégrant à la fois des informations de similarité et des informations intrinsèques aux séquences.

La prédiction de gènes à ARN est un problème plus complexe que la prédiction de gènes codants pour plusieurs raisons. On ne dispose pas des signaux forts présents dans les gènes codants : l'existence d'un cadre ouvert de lecture et les biais dans l'usage des codons. De plus, contrairement aux gènes codants, la production de certains ARN non-codants ne suit pas le schéma classique "un gène pour une molécule" : certains ARN non-codants sont localisés dans les introns d'autres gènes ou encore dans les régions non traduites des ARN messagers. A cause de ce type d'ARN, il devient plus compliqué d'exploiter les signaux qui balisent traditionnellement les gènes. Enfin, les propriétés à l'origine de la fonction des ARN varient d'une famille d'ARN à une autre : certaines familles sont caractérisées par la seule conservation d'un motif de séquence, d'autres par une structure commune. Pour toutes ces raisons, des méthodes de prédiction d'ARN non-codants *ad hoc* ont été développées, c'est-à-dire des méthodes qui ciblent une seule famille d'ARN non-codants. L'idée est alors de rechercher des séquences ou de vérifier si des séquences respectent un ensemble de contraintes qui décrivent les propriétés conservées au sein d'une famille particulière.

L'organisation de cet état de l'art est calquée sur celui des méthodes de prédiction de séquences codantes du chapitre 2. Nous envisageons dans un premier temps les approches *ab initio*, c'est-à-dire l'exploitation de signaux exclusivement présents dans les séquences d'ARN non-codants. Dans la section 3.2.1, nous nous intéressons donc à l'analyse de différents biais de composition des séquences d'ARN non-codants liés à la formation d'une structure. Cette analyse nous conduit naturellement vers l'analyse de la stabilité des structures d'ARN non-codants présentée dans la section 3.2.2. Ensuite, nous nous intéressons aux méthodes de prédiction d'ARN non-codants par homologie, c'est-à-dire la recherche de séquences homologues dans les banques de données. Dans la section 3.2.3, deux types de similarité sont ainsi envisagées : au niveau nucléique et au niveau structural. Enfin, nous clôturons cet état de l'art par les approches comparatives de prédiction d'ARN non-codants (section 3.2.4).

3.2.1 Les biais de composition en séquence

Comme il existe un biais de composition dans la séquence codante d'un gène à protéine, on peut supposer qu'il existe également un biais de composition dans la séquence d'un ARN non-codant structuré. Certains appariements sont plus stables que d'autres, ce qui peut introduire un biais de composition en mono-nucléotides. L'adjacence des appariements est également importante. Ces empilements contribuent en effet beaucoup à la stabilité des structures, ce qui peut introduire un biais de composition en di-nucléotides (sections 1.3.1 et 3.1.1).

Dans [Sch02], Schattner s'est intéressé à l'existence de biais de composition dans les séquences d'ARN non-codants dont la fonction dépend essentiellement de leurs structures. Cette étude a été menée sur des ARN de transfert, des ARN ribosomiques, des ARN nucléaires, des ARN nucléolaires et des SRP de trois organismes : la bactérie *Methanococcus jannaschii*, le ver *Caenorhabditis elegans* et le parasite *Plasmodium falciparum*. Les mesures effectuées sont la fréquence d'apparition des bases `G` et `C` et la fréquence d'apparition du di-nucléotide `CG` normalisée par les fréquences d'apparition des nucléotides `G` et `C` notée $\rho(CG)$. Les résultats obtenus sont rapportés dans la table 3.2.

(a) Résultats sur les ARN non-codants.

Organisme	Nb séq.	(G+C)%	ρ(CG)
Methanococcus jannaschii	44	63.1 (7.3)	0.75 (0.24)
Caenorhabditis elegans	59	32.1 (7.2)	0.94 (0.56)
Plasmodium falciparum	59	53.5 (8.2)	0.96 (0.23)

(b) Résultats sur les génomes.

Source	(G+C)%	ρ(CG)
Methanococcus jannaschii	31.4 (6.9)	0.34 (0.47)
Caenorhabditis elegans (chr. II)	20.0 (8.4)	0.75 (1.30)
Plasmodium falciparum (chr. I)	35.9 (8.8)	1.03 (0.68)

Tab. 3.2 – Résultats des mesures effectuées dans [Sch02]. (G+C)% correspond à la moyenne du pourcentage en `G` et en `C` observé. ρ(CG) correspond à la moyenne de la fréquence normalisée du di-nucléotide `CG`. Les valeurs entre parenthèses sont les écarts-types associés.

Dans chacun des trois organismes, le pourcentage en `G` et en `C` des séquences d'ARN non-codants est en moyenne plus élevé que celui de leur génome. Pour *Methanococcus jannaschii* et *Caenorhabditis elegans*, le di-nucléotide `CG` apparaît plus fréquemment dans les ARN non-codants que dans le reste de leur génome. Cependant, ce di-nucléotide est globalement sous-représenté dans les génomes de ces organismes par rapport aux nucléotides `C` et `G`, c'est-à-dire que les di-nucléotides `CG` et `GC` n'apparaissent pas de manière équiprobable, compte tenu des fréquences d'apparition des nucléotides `C` et `G`. Dans *Plasmodium falciparum*, le phénomène inverse se produit puisque le di-nucléotide `CG` est légèrement moins fréquent dans les ARN non-codants que dans son génome. Globalement, ces observations font apparaître une grande variabilité du pourcentage en `G` et `C` ainsi que de la fréquence d'apparition du di-nucléotide `CG` entre les organismes. Les valeurs des écarts-types montrent également que cette variabilité existe au sein même d'un organisme. Par la suite, les expériences ont été focalisées sur la prédiction de gènes à ARN chez *Methanococcus jannaschii* en mesurant localement les fréquences en mono- et di-nucléotides. Après divers ajustements, les meilleurs résultats obtenus permettent de retrouver les 44 ARN non-codants contenus dans le génome

de *Methanococcus jannaschii*, mais également 28 régions supplémentaires, ce qui représente tout de même près de 40% de prédictions fausses. Bien que peu d'organismes aient été pris en compte dans ces expériences, les résultats de cette étude démontrent qu'il existe des biais de composition significatifs en di-nucléotides dans les séquences d'ARN non-codants pour certains organismes mais que le signal apporté est à lui seul insuffisant pour détecter de manière fiable des ARN non-codants.

En parallèle des travaux de Schattner, d'autres se sont intéressés aux organismes hyperthermophiles dont fait partie *Methanococcus jannaschii*. Le génome de ces organismes, riche en `A` et en `T`, constitue un terrain plus propice à détecter des biais de composition dans les séquences d'ARN non-codants structurés [KME02]. Les auteurs utilisent ici un modèle de Markov caché entraîné sur tous les ARN non-codants connus d'organismes hypertermophiles afin de prédire des régions susceptibles de contenir de nouveaux ARN non-codants. Néanmoins, cette étude utilise un logiciel supplémentaire pour confirmer ces prédictions, QRNA présenté dans la section 3.2.4, dont les fondements dépassent la simple recherche par biais de composition.

En marge de la recherche de biais de composition en mono- et di-nucléotides, une autre méthode, RNAGENIE [CDH01], s'attache à la recherche de motifs structuraux qui pourraient trahir la présence d'ARN non-codants. Des ARN fonctionnels partagent en effet des éléments de structure dont une grande partie sont représentées par des motifs dans les séquences correspondantes (section 1.3.1). L'idée maîtresse est que ces motifs apparaissent plus fréquemment dans les séquences d'ARN non-codants que dans d'autres types de séquences. La classification à partir des fréquences d'occurrences de toute une série de motifs, complétées par d'autres mesures comme la composition en mono- et di-nucléotides, est confiée à un réseau de neurones entraîné à partir de séquences de deux souches d'*Escherichia coli*. Les performances de RNAGENIE sont évaluées sur les génomes de huit organismes. Entre 80% et 90% des ARN non-codants sont correctement détectés avec une proportion de prédictions positives erronées en moyenne inférieure à 15%. Ces résultats restent néanmoins assez variables selon les organismes allant de 64% de spécificité pour 68% de sensibilité à plus de 90% de spécificité pour 90% de sensibilité. Si l'on regarde de plus près ce que le réseau de neurones a appris, on constate que les entrées les plus informatives sont les fréquences des nucléotides `G` et `U`, ainsi que des fréquences d'apparition des di-nucléotides `CU`, `GU` et `GG`. Les motifs structuraux ne participent que faiblement au processus de décision. Les résultats de RNAGENIE sur *Methanococcus jannaschii* sont meilleurs que les résultats obtenus par la méthode de Schattner. L'amélioration provient essentiellement d'un processus d'apprentissage plus fin et de l'utilisation des fréquences de tous les mono- et di-nucléotides. Ces travaux montrent que les motifs structuraux apportent moins d'information que les biais de composition en mono- et di-nucléotides pour la détection d'ARN non-codants. Le peu de variété des organismes utilisés ne permet toutefois pas de tirer des conclusions générales. De plus, les résultats de RNAGENIE ne sont pas reproductibles car la méthode n'est pas disponible librement et les auteurs ne sont pas disposés à le fournir à titre académique.

3.2.2 La stabilité thermodynamique

La recherche de biais de composition décrite précédemment n'offre pas un signal suffisant pour une détection systématique des ARN non-codants structurés. L'une des raisons est que les biais de composition recherchés porte sur la détection de régions susceptibles de contenir des appariements particulièrement stables mais ne capte pas nécessairement le potentiel total

d'une région à former une structure complète fonctionnelle. Les programmes de prédiction de structures qui adoptent une approche thermodynamique (section 3.1.1) sont conçus pour fournir la structure d'énergie libre minimale qui peut se former à partir d'une séquence. Quelque soit la séquence choisie, ces programmes prédisent toujours une structure. Seule, l'existence d'une structure prédite n'est donc pas informative. C'est pourquoi il est nécessaire de s'intéresser plus précisément à la qualité, en terme de stabilité thermodynamique, des structures prédites.

Pour qu'une structure se forme, de nombreux appariements se font puis se défont jusqu'à ce qu'un état stable soit atteint. On peut donc s'attendre à ce que les structures des ARN non-codants soient remarquablement stables et caractérisées par une énergie libre particulièrement faible. Evaluer la significativité de la stabilité d'une structure nécessite de disposer d'une distribution de l'énergie libre avec laquelle effectuer la comparaison. Il n'existe cependant aucune théorie pour la construire, elle est donc établie de manière empirique grâce à de nombreuses séquences aléatoires équivalentes. Le protocole suivi pour évaluer la significativité de la stabilité d'une structure est le suivant. A partir d'une séquence s,

1. calculer E, l'énergie libre minimale de la structure optimale prédite pour s ;
2. construire la distribution de l'énergie libre, c'est-à-dire inférer les structures d'un grand nombre de séquences aléatoires obtenues par mélange de s ou en utilisant un processus Markovien, de telle sorte que la composition en mono-nucléotides et/ou en di-nucléotides de s soit conservée ;
3. évaluer la significativité de E à partir de la distribution obtenue grâce au z-score ou à la P-valeur de E ; ces mesures sont équivalentes en terme d'information apportée.

Le z-score de E mesure l'écart de E par rapport à la distribution. On l'obtient en calculant le rapport

$$\frac{E - \mu}{\sigma}$$

où μ et σ sont respectivement la moyenne et l'écart-type de la distribution. Comme l'énergie libre est à valeur négative, plus le z-score de E est faible, plus la structure optimale de s est stable. La seconde mesure que l'on utilise est la P-valeur de E, qui correspond à la probabilité d'obtenir une valeur d'énergie libre inférieure ou égale à E dans la distribution. Plus la P-valeur de E est proche de 0, plus le nombre de structures prédites ayant une énergie libre inférieure est faible. Par conséquent, plus la P-valeur de E est faible, plus la stabilité de la structure optimale de s est significative.

Toute la difficulté dans ce processus d'évaluation réside dans la constitution de la distribution, et donc dans le choix de la composition des séquences. La conservation de la composition mono-nucléotidique de s permet de tenir compte d'un éventuel biais de composition en `G` et en `C` de s dû à l'existence d'une structure. La conservation de la composition di-nucléotidique de s a une propriété supplémentaire : prendre en considération la formation éventuelle d'empilements d'appariements.

Composition mono-nucléotidique équivalente

Dans un première étude de la stabilité des structures d'ARN non-codants [RE00], Rivas *et al* ont cherché à estimer si l'énergie libre d'une structure optimale potentielle était significative par rapport à des séquences de composition équivalente. Ils ont donc mesuré les variations du pourcentage en `G` et en `C` et les variations de l'énergie libre d'une structure optimale locale

sur un fragment du génome de *Caenorhabditis elegans* contenant deux ARN de transfert, et sur le même fragment où les structures des ARN de transfert ont été détruites sans dénaturer la composition locale en mono-nucléotides.

Les observations réalisées montrent que les variations de l'énergie libre d'une structure optimale sont liées à un biais de composition en `G` et en `C` et n'apportent donc pas plus d'information que les variations de composition en mono-nucléotides. Cette affirmation est vérifiée en plongeant un ARN de transfert de *Caenorhabditis elegans* dans une séquence aléatoire de même composition mono-nucléotidique : l'ARN de transfert est alors indétectable en observant les variations locales de l'énergie libre.

Pour pouvoir généraliser leurs observations sur les ARN de transfert, ils ont calculé les z-scores de l'énergie libre de 243 ARN non-codants. Ces ARN sont issus de diverses familles : des SRP, des petits ARN nucléolaires, des RNaseP et des télomérases. La distribution des z-scores obtenus est donnée en figure 3.7. Sur ce graphique, la négation des z-scores est représentée, c'est-à-dire que les z-scores les plus élevés correspondent aux énergies libres les plus faibles.

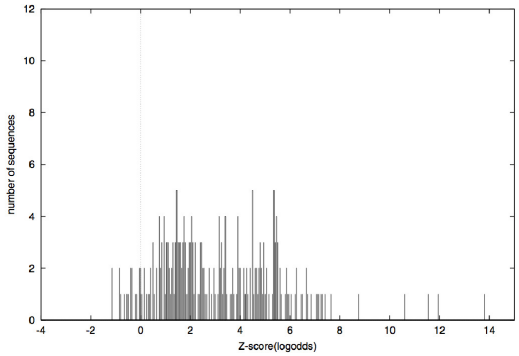


FIG. 3.7 – Distribution de la négation des z-scores de l'énergie libre des structures de 243 ARN non-codants par rapport à des structures optimales de séquences aléatoires de même composition en mono-nucléotides.

L'observation réalisée sur les ARN de transfert n'est pas valable pour toutes les familles d'ARN non-codants : en moyenne, les structures d'ARN non-codants sont plus stables que les structures optimales de séquences aléatoires équivalentes. Cependant, la stabilité moyenne des ARN non-codants n'est pas assez significative pour constituer un signal suffisant pour les détecter lorsque la distribution de référence est construite avec des séquences de même longueur et de même composition en mono-nucléotides. Cette étude a tout de même donné lieu au développement d'un logiciel, NCRNASCAN, qui réalise des prédictions d'ARN non-codants structurés selon ce procédé.

Composition di-nucléotidique équivalente

Les expériences précédentes ont été reprises dans [BWRVdP04] en considérant des séquences de même composition en di-nucléotides pour 500 ARN de transfert, 581 ARN ribosomiques et 506 micro ARN. Les résultats révèlent que les micro ARN possèdent systématiquement des structures plus stables que des structures de séquences aléatoires de même composition en di-nucléotides. Les structures des ARN ribosomiques et des ARN de transfert ne sont pas systématiquement plus stables, mais en moyenne plus stables.

Ces investigations ont été étendues à 300 familles d'ARN non-codants [CFKK05]. Cette étude ouvre des perspectives intéressantes quant à la conservation de la composition di-nucléotidique pour mesurer la significativité de la stabilité des structures d'ARN non-codants. Le tableau 3.3 reprend une partie de leurs résultats.

Famille	Nb. séq.	z-score moyen	Ecart-type des z-scores	z-score maxi	z-score mini	P-valeur moyenne
ARNt	530	−1,591	0,890	0,732	−4,035	0,123
Hammerhead III	114	−3,188	0,871	−1,203	−5,345	0,008
SECIS	5	−4,736	1,123	−3,482	−6,694	0,000
SRP	94	−3,564	2,140	−0,099	−9,255	0,046
U1	53	−1,750	0,931	0,157	−4,041	0,102
U2	62	−4,225	1,216	−1,831	−7,068	0,002

TAB. 3.3 – Extraits des résultats de Clote *et al* [CFKK05]. Les z-scores et P-valeurs sont ceux de l'énergie libre des structures.

Les structures des ARN non-codants sont en moyenne plus stables que ce qui est attendu par hasard. Pour certaines familles, comme les ARN de transfert et les petits ARN nucléolaires U1, les résultats sont plus modérés : la stabilité moyenne des structures n'est pas aussi significative que pour les autres familles d'ARN. A l'instar de l'étude de Rivas *et al*, cette étude a donné lieu au développement d'un logiciel, RANDFOLD, qui évalue la stabilité thermodynamique d'une structure par rapport à une distribution d'énergie libre construite à partir de séquences aléatoires de même longueur et composition en di-nucléotides.

3.2.3 L'homologie de séquence et de structure

A l'image de techniques déployées pour les séquences codantes présentées dans la section 2.2, la recherche d'ARN non-codants peut se faire par homologie à deux niveaux. D'une part l'homologie au niveau nucléotidique, d'autre part l'homologie en lien étroit avec la fonction des séquences nucléotidiques, c'est-à-dire l'homologie au niveau peptidique dans le cas des séquences codantes et l'homologie au niveau de la structure pour les séquences d'ARN non-codants. Dans un premier temps, nous présentons rapidement comment mener une recherche d'ARN non-codants par homologie de séquences. Ensuite, nous nous intéressons plus longuement à la recherche d'ARN non-codants par homologie de structures. Enfin, pour conclure nous présentons une évaluation des méthodes pour la recherche d'ARN non-codants par homologie menée par Gardner.

Recherche par pure similarité de séquences

La recherche d'ARN non-codants par similarité de séquences fait intervenir les outils d'alignement de séquences déjà évoqués dans les sections 1.5 et 2.2. La similarité avec d'autres ARN non-codants fait cependant appel à d'autres bases de données. Il existe à cet effet des bases généralistes : RFAM [GJBM+03, GJMM+05], NONCODE [CBG+05] et FRNADB [KYT+07, MYH+08], et des bases spécifiques pour quelques familles d'ARN non-codants ou pour certains organismes comme par exemple MIRBASE [GJGvD+06, GJ06] pour les microARN, RNADB [PSE+05, PSD+07] pour les ARN de mammifères, RNASE P DATABASE [BHGP94] pour les RNase P ou encore RIBOSOMAL DATABASE PROJECT [CWC+09]

pour les différentes sous-familles d'ARN ribosomiques. De manière générale, une similarité significative entre des régions non codantes peut également constituer une information pertinente sur la présence de séquences soumises à une contrainte fonctionnelle.

Recherche par similarité de séquences et de structures

Les méthodes de recherche d'ARN non-codants par similarité de structures font intervenir deux processus : la construction manuelle ou automatique d'un *profil* représentatif d'un ensemble d'ARN non-codants homologues, et la recherche effective de séquences similaires au regard d'un profil. Le profil intègre des informations sur la formation d'appariements mais également des informations de séquence dans les régions appariées et non appariées, c'est pourquoi on parle d'homologie de séquences et de structures. On peut distinguer deux types d'approches pour ce problème : les méthodes probabilistes et les méthodes déterministes.

Méthodes à base de modèles probabilistes. Les méthodes à base de modèles probabilistes sont analogues aux méthodes d'alignement classiques. Globalement, leur objectif est d'établir un alignement optimal au regard d'un système de score entre une structure et une séquence. La mise en œuvre de ce type d'alignement repose sur plusieurs opérations d'édition relatives aux structures d'ARN : la substitution d'un nucléotide non apparié, la substitution d'une paire de nucléotides appariés, l'insertion et la délétion de nucléotides et d'appariements. Concrètement la construction de ce type d'alignement est plus complexe que l'alignement "classique" de séquences à cause des interactions créées entre les nucléotides. L'espace des alignements possibles est, dans la majorité des cas, modélisé par une grammaire stochastique hors contexte, c'est-à-dire un ensemble d'états et de règles pour transiter entre les états étiquetés par une opération d'édition. Chaque alignement possible correspond alors à une dérivation de ces règles. Un alignement optimal dans ce système correspond donc à une dérivation de score maximal. L'obtention d'un alignement optimal au regard du système de score fixé est déléguée à un algorithme de programmation dynamique, comme l'algorithme CYK, dont la complexité temporelle est en $\mathcal{O}(mn^3)$ et la complexité spatiale en $\mathcal{O}(mn^2)$ où n est la longueur de la séquence à aligner, et m le nombre de règles de la grammaire [DEKM99].

L'une des premières méthodes d'alignement structure/séquence à base de modèles probabilistes est RSEARCH [KE03]. Cette méthode repose sur les matrices RIBOSUM conçues pour évaluer la substitution d'appariements. A l'image des matrices BLOSUM pour les séquences d'acides aminés, ces matrices sont construites à partir d'alignements multiples d'excellente qualité d'ARN non-codants. RSEARCH est relativement gourmand en ressources et s'avère par conséquent difficilement praticable pour traiter des génomes entiers d'eucaryotes supérieurs. Inspiré de RSEARCH, RSMATCH [LWHT05] propose un découpage en modules élémentaires de la structure d'origine : multiboucle, tige, région non appariée. Les modules sont évalués indépendamment à l'aide des matrices RIBOSUM, puis sont assemblés pour former une structure complète. Cette heuristique s'avère légèrement plus rapide que RSEARCH, mais à spécificité égale elle se montre en moyenne moins sensible. L'algorithme déployé dans FASTR [BZ04] applique un filtre pour ne conserver que les régions susceptibles de donner de bons alignements. Le filtrage repose sur l'expression de contraintes sur la formation de tiges aux propriétés ressemblantes aux tiges présentes dans la structure à aligner. Ces propriétés sont la distance qui sépare les parties ouvrante et fermante d'une tige et la longueur de la tige. Aucune information de séquence n'est prise en compte à ce niveau. Une fois les régions d'intérêts filtrées, l'alignement est réalisé par programmation dynamique à l'aide des

matrices RIBOSUM. HomoStRscan [LMZ04] est une méthode en tout point équivalente à Rsearch adaptée au traitement du génome humain : les matrices de substitutions sont construites à partir de séquences d'ARN non-codants provenant du génome humain.

Rsearch et FastR ont la propriété d'être génériques au regard des structures recherchées grâce à l'utilisation des matrices RIBOSUM. Toutefois, cette approche générale ne permet pas de prendre en compte certaines caractéristiques propres à un ensemble de séquences structurées homologues, telle que la conservation locale de motifs de séquence ou de structure (section 1.3.1). Si l'on souhaite cibler les recherches à un ensemble de séquences homologues particulier, les modèles de covariance sont alors plus appropriés [ED94, DEKM99]. Sans entrer dans les détails, un modèle de covariance est une grammaire stochastique hors contexte profilée par plusieurs matrices de substitutions de nucléotides et d'appariements. A partir d'un alignement multiple de séquences homologues annoté par la structure commune partagée par les séquences, on construit une matrice de substitutions d'appariements pour chaque couple de colonnes appariées et une matrice de substitutions de nucléotides pour chaque colonne non appariée. Ainsi, un modèle de covariance capture à la fois des informations locales sur la structure, mais également sur la séquence dans les régions appariées et non appariées.

Plusieurs méthodes s'appuient sur les modèles de covariance, notamment Infernal [GJBM$^+$03] utilisée pour maintenir la banque Rfam. Un modèle de covariance est en effet disponible pour chaque famille de Rfam, construit à partir d'un alignement multiple vérifié manuellement et annoté par la structure conservée. Infernal est une implémentation stricte et rigoureuse des modèles de covariance, c'est-à-dire qu'aucun traitement ne précède l'alignement. L'alignement à l'aide d'un modèle de covariance est une tâche plus gourmande en temps de calculs et en espace mémoire qu'un alignement basé sur une grammaire "générique", telle que celle utilisée dans Rsearch. L'exploitation des modèles de covariance est particulièrement gourmande en ressources, d'autant plus que la taille d'un modèle de covariance, c'est-à-dire le nombre de règles dans la grammaire associée, est proportionnelle à la longueur de l'alignement multiple ayant servi à sa construction et au nombre d'appariements impliqués dans la structure commune.

Plusieurs manières de contourner ce problème ont été envisagées. Dans RNAcad [Bro99], un modèle de Markov caché sert à détecter des régions non appariées précises de la séquence à aligner. Ce marquage permet d'imposer des contraintes lors de l'évaluation du modèle de covariance, sous forme de points de passage forcés dans la dérivation des règles de la grammaire associée au modèle. Cette solution s'avère extrêmement efficace, bien qu'elle nécessite beaucoup de données supplémentaires pour entraîner le modèle de Markov. Dans le même esprit, RaveNnA [WR04, WR06] construit un modèle de Markov particulier, un *profile HMM*, à partir d'un modèle de covariance. Un profile HMM est un modèle de Markov caché adapté pour modéliser un alignement multiple et non une simple séquence comme les modèles de Markov cachés classiques. Comme les modèles de Markov classiques, ce type de modèle ne permet pas de décrire la formation d'appariements entre nucléotides. Dans RaveNnA, le profile HMM est construit à l'aide d'une heuristique basée sur le principe du maximum de vraisemblance pour approximer au mieux le modèle de covariance original. Le filtrage à l'aide du profile HMM permet de cibler les régions dont l'évaluation à l'aide du modèle de covariance est susceptible de donner un score significatif. En moyenne, 90% des alignements trouvés avec un modèle de covariance peuvent ainsi être retrouvés avec le profile HMM correspondant dans RaveNnA, et ce environ 600 fois plus rapidement. Jusqu'ici nous n'avons qu'une seule facette des modèles de covariance : construire un alignement structure/séquence. CMFinder [YWR06] propose

d'utiliser ce type de modèles pour tenter d'améliorer un alignement classique réalisé uniquement sur la séquence primaire. A partir d'un ensemble d'alignements classiques, CMFinder sélectionne des bons candidats potentiels sur la base de critères énergétiques, puis affinent les alignements retenus grâce au modèle de covariance selon le principe de maximum de vraisemblance. Dans sa première version, Infernal était relativement gourmand en ressources car il ne procédait à aucune optimisation spécifique ou filtrage particulier en amont. Récemment, Infernal a évolué [NE07] et intègre maintenant un filtre qui permet de déterminer de manière exacte les dérivations de la grammaire qui n'aboutiront pas à un alignement significatif étant donné un modèle de covariance. Cette optimisation permet de diminuer de manière drastique les ressources physiques et temporelles nécessaires à Infernal, et par extension, aux approches à base de modèles de covariance.

Comme nous venons de le voir, le modèle de covariance est un outil puissant pour modéliser des séquences homologues partageant une structure commune. L'exploitation de cette modélisation s'avère cependant généralement coûteuse sans un filtrage approprié de l'espace des dérivations à explorer. Erpin [GL01, LFL$^+$04, LLFG05] propose une version simplifiée des modèles de covariance : une matrice de score pour la substitution d'appariements est calculée pour chaque tige, et non pour chaque appariement, et une matrice pour la substitution de nucléotides non appariées pour chaque sous-séquence non appariée. A la manière de FastR, un ensemble de contraintes est associé à chaque tige, notamment l'intervalle des distances autorisées entre la partie ouvrante et la partie fermante. Les tiges sont détectées de manière indépendante, puis assemblées pour former un alignement par programmation dynamique : il n'y a donc pas de notion de structure globale dans l'algorithme, ce qui réduit drastiquement sa complexité. Le facteur déterminant de la complexité de Erpin est la distance maximale autorisée qui sépare les parties ouvrante et fermante des tiges.

Méthodes à base de descripteurs abstraits. Les méthodes d'alignements structure/séquence décrites précédemment permettent de retrouver des séquences similaires à un ensemble de séquences structurées homologues. Pour ce faire, elles tirent parti d'un alignement fiable annoté par une structure commune. Toutefois, cet alignement n'est pas toujours disponible, ou ne contient pas suffisamment de séquences pour former un échantillon statistiquement représentatif de la famille de séquences ciblée. La structure d'une famille peut également n'être connue que partiellement auquel cas les informations capturées dans le modèle de covariance ne reflètent pas la structure réelle. Dans ces cas de figure, on pourra alors se tourner vers les méthodes à base de descripteurs abstraits à base de *descripteurs*. Le principe général de ses méthodes est de rechercher toutes les régions d'une séquence qui répondent à un ensemble de contraintes formalisées sur la formation d'éléments structuraux ou de séquence. Par exemple, la formation d'une tige de trois nucléotides dont les parties ouvrante et fermante sont séparées d'au plus dix nucléotides contenant la séquence `ACGU`. Les différences entre les méthodes basées sur des descripteurs proviennent du pouvoir d'expression du formalisme adopté. Nous nous focaliserons donc principalement sur cette caractéristique dont l'évolution suit l'ordre chronologique d'apparition des méthodes.

RNAMot [GMC90, LGC94] est le premier véritable outil à base de descripteurs applicable à la recherche de séquences structurées. Trois types d'éléments sont définissables dans les descripteurs de RNAMot : les mots, les espaceurs et les tiges. Les mots servent à décrire des régions de taille fixe, tandis que les espaceurs décrivent des régions de taille variable. Les tiges ont une longueur variable bornée, peuvent contenir un nombre variable de mésappariements

et des boucles internes symétriques ou non, et peuvent former des pseudo-œuds. Les mots sont systématiquement décrits par une séquence, contrairement aux espaceurs et aux tiges dont la description du contenu en nucléotides est facultative. Quelque soit le type d'élément décrit, le contenu en nucléotides peut être approximatif, c'est-à-dire que les erreurs sont tolérées dans une certaine mesure variable d'un élément à un autre. De plus, pour les tiges il est possible d'autoriser certaines liaisons bancales et d'en préciser la quantité maximale autorisée. Pour trouver les séquences qui satisfont un descripteur, RNAMot ordonne les éléments du descripteur en fonction de leur probabilité marginale d'apparition estimée de manière empirique, puis tente de placer les éléments récursivement, du moins probable au plus probable. Lorsque deux occurrences chevauchantes satisfont un descripteur, RNAMot calcule un score qui permet de déterminer la meilleure occurrence, c'est-à-dire celle qui remplie au mieux les éléments décrits : les tiges plus longues sont favorisées, les mésappariements et les erreurs défavorisées, ...

Inspiré de RNAMot, Palingol [BKV96] reprend les mêmes types d'éléments descriptifs. Toutefois, Palingol est beaucoup plus expressif car il offre la possibilité de définir des expressions logiques et de construire des branchements. Par exemple, il devient possible de décrire : « si cette tige n'est pas présente ou qu'elle contient plus d'un mésappariement, alors appliquer une pénalité ». Le programme s'avère également plus souple dans la gestion des mots ; un mot peut être décrit par une matrice poids-position, comme celle utilisée pour modéliser les sites d'épissage (section 2.1.2). Pour trouver les séquences qui satisfont un descripteur, l'algorithme de Palingol comporte une phase préliminaire d'indexation de toutes les tiges présentent dans la séquence afin de ne pas rejouer plusieurs fois inutilement des calculs coûteux.

PatScan [DLO97] offrent les mêmes possibilités que Palingol, bien que la gestion des erreurs diffère : les substitutions, les insertions et les délétions sont considérées comme des erreurs différentes, sans aucun moyen simple de les banaliser. Cette caractéristique peut rendre l'élaboration du descripteur particulièrement difficile puisque qu'il faut alors explicitement écrire les expressions conditionnelles qui simulent ce comportement. PatSearch [PLD00, GLL+03] est le descendant de PatScan. Outre certains aspects syntaxiques différents dans la définition des descripteurs, PatSearch estime le nombre d'occurrences attendues par hasard avec un descripteur. Cette valeur est obtenue par simulation, en exécutant PatSearch avec le même descripteur sur des séquences aléatoires de même composition en mono-nucléotides. L'espérance du nombre d'occurrences attendues donne une indication sur la "qualité" d'un descripteur afin de relativiser le nombre d'occurrences réellement observées.

RNAMotif [MEG+01], le descendant de RNAMot, est actuellement le logiciel le plus utilisé. La principale nouveauté apportée dans RNAMotif est la totale liberté laissée à l'utilisateur pour définir son propre système de score. La définition de ce système peut très sophistiquée et faire appel à des structures algorithmiques relativement évoluées, telles que des expressions conditionnelles et des boucles.

MilPat [TdGSG06] est une méthode récente qui se distingue des autres à deux points de vue : la possibilité de décrire des interactions inter-séquences, et l'algorithme d'énumération des occurrences. La fonction de bon nombre d'ARN non-codants implique une interaction avec une autre séquence nucléique (section 1.3) qui se traduit par la formation d'appariements. MilPat permet de décrire ce type d'interactions comme la formation d'une "tige" dont une partie se trouve sur la séquence cible et l'autre partie sur une séquence fournie en plus du descripteur. L'autre originalité de MilPat est l'algorithme d'énumération des

occurrences. Le problème de trouver les occurrences qui satisfont un descripteur est ici formellement défini comme un *problème de satisfaction de contraintes*, une classe de problèmes mathématiques largement étudiée. Cette formalisation permet ainsi d'hériter des algorithmes de résolution classiques de ce type de problèmes qui font de MilPat l'une des méthodes les plus rapides actuellement. Récemment, MilPat a été décliné en une nouvelle version, Darn! [Zyt07, ZGS08], où le formalisme a été étendu aux réseaux de contraintes pondérés qui permet d'intégrer un système de score évolué. Cette modification a également conduit à l'intégration d'un module de gestion des solutions chevauchantes.

BRAliBase III, benchmark de recherche d'ARN non-codants par homologie

Dans les deux sections précédentes, nous avons vu qu'il existait de nombreux logiciels pour retrouver des ARN non-codants sur la base d'une similarité de séquence et/ou de structure. Nous présentons maintenant un benchmark de référence d'une partie de ces méthodes nommé BRAliBase III [FBG07].

BRAliBase III est un jeu de données qui contient trois familles de séquences : 1114 ARN de transfert, 602 ARN ribosomiques 5S et 235 petits ARN U5. Chaque famille est représentée par un alignement structural extrait de la littérature. Au sein de chaque famille, le pourcentage d'identité entre couples de séquences varie de 40% à 95%. Le protocole expérimental mis en place pour tester les méthodes est le suivant. Pour chaque famille, cinq sous-alignements de cinq et vingt séquences sont extraits aléatoirement et utilisés pour retrouver les séquences homologues restantes. Pour les méthodes qui ne travaillent qu'à partir d'une seule séquence, les séquences extraites sont utilisées successivement et les résultats agrégés. Ensuite, pour mesurer la spécificité, un jeu de données de dix alignements est construit pour chaque famille en mélangeant l'alignement structural selon la procédure utilisée dans AlifoldZ [WH04].

Les résultats de BRAliBase III sont synthétisés, toutes familles confondues, dans les tables 3.4 et 3.5. La sensibilité et la spécificité présentées dans ces tables représentent respectivement la proportion de séquences d'ARN non-codants prédites comme telles et la proportion de séquences aléatoires non prédites comme étant des ARN non-codants. Ces proportions sont calculées en fonction du nombre de séquences uniques prédites par les logiciels, et non à partir du nombre de bases impliquées dans les prédictions. Par conséquent, une séquence d'ARN est considérée comme prédite dès lors que le logiciel prédit au moins une base de cette séquence comme faisant partie de la famille recherchée. La plupart des logiciels n'effectuent pas a proprement parler de prédiction "ARN non-codants homologues"/"autre", mais calculent un score d'alignement selon un système propre. Ainsi, pour chacun de ces logiciels un seuil sur le score qu'ils calculent a été ajusté manuellement afin d'optimiser le cœfficient de corrélation de Matthews (noté MCC) en fonction du jeu de données.

Globalement, l'élément qui ressort de cette étude est que toutes les méthodes testées sont remarquablement spécifiques puisqu'aucune ne descend en dessous de 98% de spécificité. Le deuxième élément important est la grande variabilité de la sensibilité en fonction de la famille d'ARN recherchée. En l'occurrence, les résultats font clairement apparaître qu'Infernal et Rsearch sont les méthodes les plus robustes et les plus stables d'une famille d'ARN à une autre. Les ARN de transfert semblent particulièrement difficiles à détecter, notamment par simple homologie de séquence où la sensibilité ne dépasse pas 32% sur les ensembles de cinq séquences et 62% sur les ensembles de vingt séquences. Toujours sur les ARN de transfert, l'approche par homologie de structure semble plus appropriée bien que les résultats sont très variables d'une méthode à une autre : Infernal et Rsearch surpassent largement leurs

Logiciel	ARN ribo. 5S			Petits ARN U5			ARN de transfert		
	Sens.	Spéc.	MCC	Sens.	Spéc.	MCC	Sens.	Spéc.	MCC
Homologie de séquence									
BLAST $w = 11$	54,64	99,66	0,698	90,45	99,43	0,915	16,54	99,87	0,374
BLAST $w = 7$	85,85	99,91	0,914	95,44	99,77	0,962	29,12	99,98	0,519
FASTA	88,16	99,90	0,927	95,99	99,75	0,964	31,40	99,98	0,540
Homologie de structure et de séquence									
ERPIN	19,30	99,77	0,395	28,47	99,90	0,505	13,88	100,00	0,357
INFERNAL	97,80	99,95	0,985	94,73	99,87	0,964	86,68	100,00	0,925
RAVENNA	88,77	99,80	0,925	95,07	99,58	0,950	47,72	99,90	0,665
RSEARCH	98,78	99,93	0,989	95,37	99,99	0,974	87,13	99,92	0,923
RSMATCH	32,05	99,94	0,542	66,95	99,59	0,778	33,64	99,94	0,556

TAB. 3.4 – Résultats de BRALIBASE III sur les ensembles de cinq séquences. La sensibilité et la spécificité sont exprimées en pourcentage.

Logiciel	ARN ribo. 5S			Petits ARN U5			ARN de transfert		
	Sens.	Spéc.	MCC	Sens.	Spéc.	MCC	Sens.	Spéc.	MCC
Homologie de séquence									
BLAST $w = 11$	71,23	98,86	0,765	94,43	97,60	0,854	48,34	99,48	0,639
BLAST $w = 7$	94,66	99,68	0,953	98,37	98,98	0,938	59,89	99,93	0,753
FASTA	96,07	99,65	0,959	98,61	98,59	0,922	61,98	99,91	0,767
Homologie de structure et de séquence									
ERPIN	24,06	100,00	0,473	40,57	100,00	0,619	15,90	100,00	0,383
INFERNAL	98,54	99,96	0,990	96,71	99,89	0,975	96,60	99,97	0,979
RAVENNA	91,51	99,78	0,940	96,33	99,41	0,948	75,07	99,85	0,847
RSEARCH	92,81	99,97	0,958	92,59	99,95	0,956	81,06	99,99	0,892
RSMATCH	54,38	99,77	0,704	93,10	98,71	0,894	59,39	99,81	0,742

TAB. 3.5 – Résultats de BRALIBASE III sur les ensembles de vingt séquences. La sensibilité et la spécificité sont exprimées en pourcentage.

homologues avec une sensibilité supérieure à 85% contre moins de 50% pour les autres sur les alignements de cinq séquences. A l'opposé des ARN de transfert, les petits ARN U5 sont plutôt bien prédits par les deux types d'approches, à l'exception de ERPIN dont la sensibilité ne dépasse pas 41%. Dans les faits, les ARN U5 comportent des sites de fixation très conservés qui, à eux seuls, constituent un signal suffisant pour les prédire. La figure 3.8 montre la structure d'un ARN U5 où les bases sont colorées en fonction de leur degré de conservation au sein de la famille. Trois sites sont particulièrement conservés : la boucle terminale et une des boucles internes de la tige en 3′ ainsi que la multiboucle.

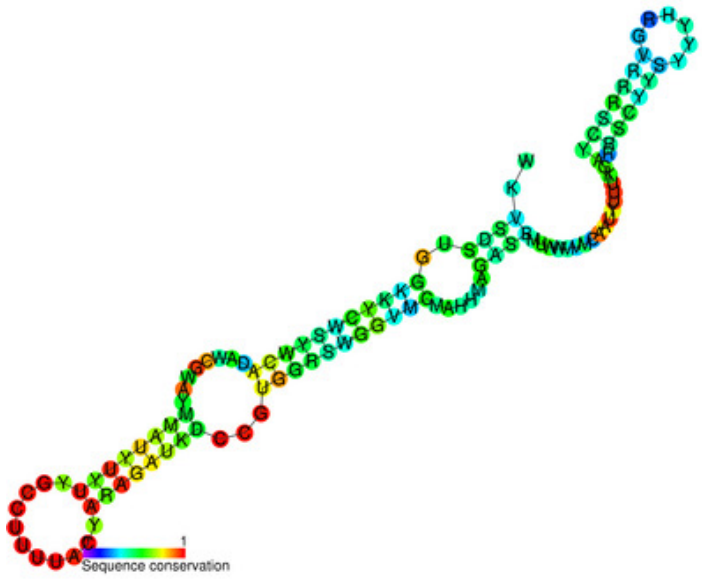


FIG. 3.8 – Structure d'un petit ARN U5 composée de deux tiges juxtaposées. Les couleurs indiquent le degré conservation de chaque base au sein des séquences connues de la famille.

Le nombre de séquences utilisées pour réaliser les prédictions influe particulièrement sur le comportement des méthodes à base d'homologie de séquences : leur sensibilité double lorsque l'on passe de cinq à vingt séquences. Bien qu'on ne dispose pas des résultats en fonction du pourcentage d'identité, les auteurs de BRALIBASE III mentionnent que la faible sensibilité des méthodes par pure homologie de séquences provient de la divergence entre les séquences requêtes et la séquence ciblée. En pratique, ils notent une nette dégradation de la sensibilité de ce type d'approche lorsque l'ensemble des séquences utilisé ne comporte aucune séquence dont le pourcentage d'identité avec la séquence à prédire est supérieur à 65%.

Les résultats en terme d'efficacité sont donnés dans la table 3.6. Sans surprise, les méthodes les plus rapides sont celles qui n'intègrent pas ou peu d'information sur la structure à savoir BLAST, FASTA et ERPIN. Si l'on omet les temps d'initialisation, les méthodes les plus performantes INFERNAL et RSEARCH sont également les plus lentes. A titre de comparaison, INFERNAL et RSEARCH traitent en moyenne entre mille et deux milles fois moins de nucléotides par seconde que ERPIN, BLAST et FASTA. Bien que négligeable pour évaluer des banques de

Logiciel	Temps d'initialisation moyen pour 20 séq. (sec.)	Rapidité (nucléotides/sec.)
Homologie de séquence		
BLAST	0,42	575 440
FASTA	0,15	758 578
Homologie de séquence et de structure		
ERPIN	0,23	363 078
INFERNAL	209	363
RAVENNA	1 479	20 893
RSEARCH	1 380	573
RSMATCH	41,7	3 631

TAB. 3.6 – Efficacité des méthodes testées dans BRALIBASE III.

données conséquentes, le temps d'initialisation qui précède la phase de recherche des méthodes les plus sophistiquées comme RAVENNA, INFERNAL et RSEARCH est relativement élevé.

3.2.4 L'approche comparative, l'existence d'une structure conservée

Bien que les approches par homologie de séquences décrites dans la section précédente peuvent se montrer relativement performantes pour retrouver des séquences appartenant à des familles d'ARN non-codants connues, elles ne permettent pas de réaliser de prédictions *de novo*. Nous nous intéressons maintenant aux méthodes qui tentent de traiter ce problème qui demeure encore à l'heure actuelle un problème ouvert.

Dans la section 3.2.2, nous avons vu que l'existence d'une structure secondaire prédite à partir d'une séquence n'est pas un indice suffisant pour permettre de détecter des ARN non-codants, même lorsque cette structure est significativement stable. Dans le cadre de la prédiction de structure secondaire, le recours à une analyse comparative permet d'améliorer significativement les prédictions en s'appuyant sur les informations évolutives qui relient entre elles des séquences homologues qui partagent une structure commune (section 3.1.2). Au carrefour de ces deux idées se situe la prédiction d'ARN non-codants par analyse comparative. L'idée est de prédire une structure commune à plusieurs séquences puis d'évaluer la "qualité" de cette structure par rapport à ce qui pourrait être attendu par hasard.

La première méthode dédiée à ce problème est QRNA [RE01]. QRNA envisage trois hypothèses pour expliquer la similarité de deux séquences alignées : les séquences sont des séquences codantes homologues ou des séquences non-codantes homologues qui partagent une structure, ou bien leur similarité est fortuite sans rapport avec la préservation d'une fonction commune. Pour évaluer chacune de ces hypothèses, QRNA s'appuie sur trois modèles qui caractérisent chacun un schéma évolutif : le modèle RNA pour les séquences non-codantes homologues où les mutations compensatoires sont privilégiées, le modèle COD pour les séquences codantes homologues où les mutations silencieuses et synonymes sont favorisées, et enfin le modèle OTH où aucun type de mutation n'est favorisé. En pratique, le modèle RNA est une grammaire stochastique profilée, identique à celle de RSEARCH. Les modèles COD et OTH sont quant à eux des modèles de Markov cachés paramétrés par apprentissage. A l'issue de l'évaluation d'un alignement selon les trois modèles, QRNA émet une prédiction sur la nature des séquences en fonction du modèle ayant obtenu la probabilité la plus élevée. Le modèle RNA de QRNA a été repris et étendu dans EVOFOLD [PBS+06] pour traiter des alignements

multiples. EvoFold utilise un type particulier de grammaires stochastiques. Afin d'ajuster les probabilités du modèle, EvoFold s'appuie sur un arbre phylogénétique contenant les distances évolutives qui sépare les organismes dont sont extraites les séquences alignées.

Qrna et EvoFold procèdent à une analyse comparative plutôt fine des mutations entre les séquences. MSARi [CKB04] adopte la même démarche à base de modèles probabilistes mais de manière heuristique. MSARi procède entre trois temps. Premièrement, les probabilités d'appariement de tous les couples de nucléotides sont calculées individuellement pour chaque séquence grâce à la fonction de partition (section 3.1.1). A partir des résultats obtenus, MSARi recherche des tiges conservées grossièrement cohérentes avec l'alignement multiple, c'est-à-dire que les tiges mises en correspondance ne respectent pas nécessairement strictement l'alignement multiple. MSARi suppose en effet que l'alignement peut contenir quelques erreurs et que les bases appariées ne sont pas nécessairement correctement alignées. Enfin, MSARi sélectionne les tiges conservées de manière gloutonne, par nombre de mutations compensatoires décroissant, pour former une structure secondaire commune. La classification est enfin réalisée en fonction de la significativité de la structure obtenue, évaluée en fonction du nombre de mutations compensatoires globalement observées.

Contrairement aux approches précédentes, une autre classe de méthodes s'attache à la stabilité thermodynamique d'une éventuelle structure secondaire commune. Cette approche suit le même schéma que celle présentée dans la section 3.2.2 où l'on évalue la stabilité d'une structure d'énergie minimale prédite sur une seule séquence. La difficulté supplémentaire ici est de construire une distribution de l'énergie libre de structures obtenues non plus sur des séquences individuelles de composition équivalente, mais sur des ensembles de séquences alignées ou non. AlifoldZ [WH04] et ddbRNA [DBDH03] procèdent ainsi à l'évaluation de la stabilité d'une structure commune par rapport à une distribution d'énergie libre construite à partir d'alignements générés en mélangeant les positions de l'alignement multiple original. Par construction, un alignement obtenu par mélange de ses positions respecte deux propriétés : la composition en mono-nucléotides de chaque séquence est préservée et la conservation globale des séquences. Dans ddbRNA, la structure commune est calculée en assemblant de manière gloutonne des tiges conservées, et la procédure de mélange s'efforce en plus de détruire au moins partiellement cette structure commune. Dans AlifoldZ en revanche, la prédiction de la structure commune est déléguée à RNAalifold, et la procédure de mélange est plus complexe car elle respecte le degré de conservation locale, c'est-à-dire que le degré de conservation de chaque position est préservé entre tous les couples de séquences. Cette propriété est particulièrement importante pour préserver les longues insertions/délétions qui pourraient alors être éclatées en plusieurs petites régions et empêcher la formation de tiges lors du repliement commun.

La figure 3.9 montre la distribution du z-score de l'énergie libre de la structure commune prédite par RNAalifold sur des alignements d'ARN de transfert comportant de une à quatre séquences. La distribution de l'énergie libre tracée en trait plein, est comparée à la distribution de l'énergie libre de la structure commune prédite par RNAalifold à partir du même alignement mélangé par la procédure de AlifoldZ. D'après ces résultats, la structure secondaire commune à plusieurs séquences semble significativement plus stable qu'une structure d'ARN non-codant prédite à partir d'une seule séquence. En fait, plus l'alignement comporte de séquences, plus l'énergie libre de la structure commune prédite est faible. RNAalifold intègre en effet dans son calcul de l'énergie libre de la structure commune un bonus négatif pour chaque covariation. Le nombre de covariations observables augmente avec le nombre

de séquences homologues différentes, par conséquent l'énergie libre de la structure commune prédite diminue. En revanche, on ne s'attend pas à trouver de covariations sur les alignements mélangés, quelque soit le nombre de séquences de l'alignement. En fixant un seuil sur la valeur du z-score de l'énergie de la structure prédite à -4, il devient alors possible de distinguer clairement les alignements d'ARN de transfert des alignements de séquences aléatoires. Plus le nombre de séquences alignées est élevé, plus cette classification s'avère efficace : pour quatre séquences, 98,36% des alignements d'ARN de transfert peuvent ainsi être discriminés sans prédire à tort un alignement de séquences aléatoires.

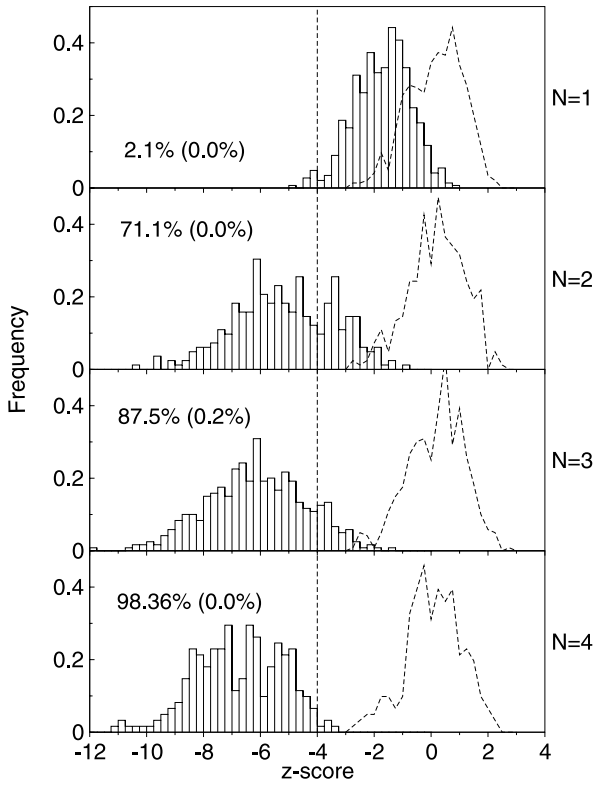


FIG. 3.9 – Distribution du z-score de l'énergie libre de la structure commune prédite par RNAALIFOLD sur des alignements de plusieurs familles d'ARN non-codants évaluée par rapport à une distribution empirique de l'énergie libre des structures prédites par RNAALIFOLD sur des alignements générés par mélange des séquences originales selon la procédure d'ALIFOLDZ. N est le nombre de séquences présentent dans les alignements. Pour $N = 1$, les structures sont prédites par RNAFOLD.

RNAZ [WHS05] est une amélioration de ALIFOLDZ qui intègre une mesure supplémentaire de la stabilité de la structure commune : le SCI (*Structure Conservation Index*). Le SCI évalue la stabilité de la structure commune par rapport aux structures prédites individuellement. Il s'obtient en calculant le rapport $E_A/\bar{E}$, où E_A est l'énergie libre de la structure commune prédite par RNAALIFOLD, et $\bar{E}$ est la moyenne de l'énergie libre des structures individuelles prédites par RNAFOLD. Lorsque le SCI est proche de 0, la structure trouvée par RNAALIFOLD a une énergie libre plus faible que la moyenne de l'énergie libre des structures individuelles : la structure trouvée pour l'alignement n'est pas significative ; les structures sont mal conservées.

Un SCI proche de 1 indique au contraire que les structures sont parfaitement conservées. Un SCI plus grand que 1 indique non seulement que les structures sont parfaitement conservées, mais qu'il existe en plus des mutations compensatoires. Afin d'éviter la construction empirique coûteuse d'une distribution d'énergie libre, les paramètres de cette distribution sont approximés au moyen d'un processus d'apprentissage supervisé, les SVM (*Support Vector Machine*). Ce même type de processus est utilisé pour effectuer la classification de l'alignement en fonction du SCI et du z-score de l'énergie libre de la structure commune. Actuellement, RNAz est la méthode la plus utilisée pour la prédiction d'ARN non-codants.

Comme nous l'avons déjà remarqué dans la section 3.2.2, évaluer la stabilité d'une structure par rapport à une distribution d'énergie libre établie à partir de séquences de même composition en di-nucléotides apporte de meilleurs résultats qu'en ne préservant que la composition en mono-nucléotides. Récemment, Sissiz [GW08] reprend le protocole employé jusqu'ici mais avec une procédure de génération d'alignements multiples qui préserve une composition en di-nucléotides donnée en plus de toutes les propriétés déjà évoquées, notamment la conservation locale. La distribution d'énergie libre obtenue à partir des alignements générés par cette procédure est selon les auteurs plus proche de la distribution réelle.

3.3 Evolution et enrichissement du logiciel caRNAc

Nous présentons maintenant notre contribution en matière de prédiction de structure secondaire, caRNAc, basée sur les travaux initiés en 2003 par Olivier Perriquet. caRNAc [PTD03, TP04, Per03] est une méthode de prédiction de structure secondaire qui suit le paradigme "aligner et replier simultanément" décrit à la page 65. A ce titre, il prédit une structure secondaire conservée entre plusieurs séquences non alignées. Le point fort de caRNAc est d'adapter l'algorithme de Sankoff de manière heuristique pour le rendre praticable, tout en adoptant le parti pris d'éviter la sur-prédiction des appariements, afin de garantir des prédictions sûres. caRNAc intègre également des informations évolutives, en prenant en compte les mutations compensatoires. caRNAc a fait l'objet d'une évaluation sur le benchmark de référence en la matière et a depuis été adopté par la communauté [GG04, Tou07].

Dans cette section, nous présentons les évolutions que nous avons apportées au logiciel. Le but de nos travaux a été de concilier au sein de caRNAc les approches "aligner et replier simultanément" et "aligner puis replier". Pour cela, nous avons utilisé le concept de méta-séquences 1.5.3, à l'image de ce qui a été mis en œuvre dans Protea (section 2.4). Ce faisant, nous avons également cherché à optimiser le cœur algorithmique de caRNAc, afin d'améliorer les temps de calcul et d'ouvrir des perspectives de traitements à grande échelle. Nous commençons par décrire la version initiale de caRNAc, puis nous présentons nos contributions, et enfin nous refermons cette section par des résultats expérimentaux et une illustration de l'utilisation de caRNAc pour la prédiction d'ARN non-codants.

3.3.1 L'existant

caRNAc admet en entrée n séquences d'ARN non alignées et produit pour chaque séquence un structure secondaire sous forme d'une liste de tiges conservées. Cela se fait en deux temps. La première phase de l'algorithme consiste à procéder à tous les repliements deux à deux suivant une adaptation de l'algorithme de Sankoff. Ensuite, ces prédictions sont combinées à l'aide d'une structure de graphe pour obtenir une structure secondaire pour

chaque séquence.

La prédiction d'une structure secondaire commune à deux séquences

L'algorithme déployé dans CARNAC pour la prédiction d'une structure commune à deux séquences produit pour chaque séquence une liste de tiges formant une structure secondaire. Cet algorithme est composé de quatre étapes dont l'enchaînement est décrit schématiquement en figure 3.10 :

1. l'énumération de toutes les tiges potentielles maximales pour chaque séquence selon les paramètres énergétiques du modèle thermodynamique ;
2. la recherche de points d'ancrage entre les séquences, c'est-à-dire des régions très conservées ;
3. l'énumération des couples de tiges compatibles avec les points d'ancrage, et filtrage des couples de tiges *copliables*, c'est-à-dire des tiges entre lesquelles on observe au moins une covariation ;
4. la recherche d'un ensemble de couples de tiges d'énergie minimale selon une adaptation des récurrences de Sankoff pour former une structure secondaire.

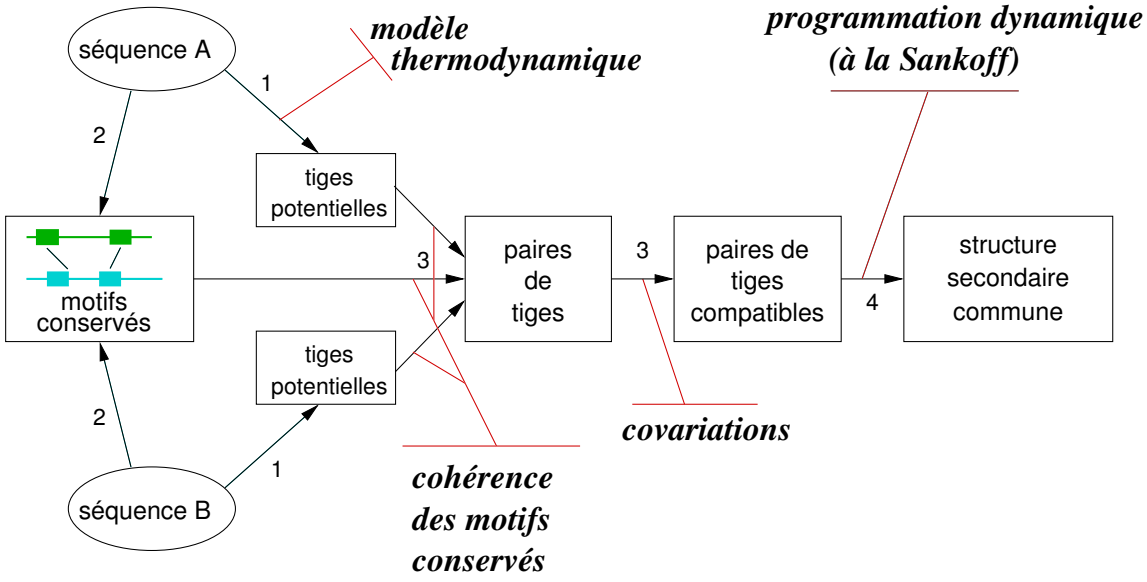


FIG. 3.10 – Déroulement de la prédiction d'une structure secondaire conservée entre deux séquences dans CARNAC

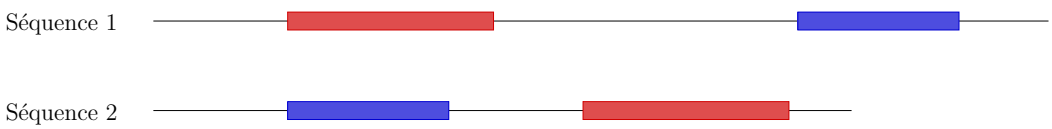
La recherche de tiges Les tiges potentielles énumérées lors de cette étape contiennent aux moins trois appariements canoniques consécutifs, peuvent contenir des mésappariements et sont systématiquement fermées par un appariement canonique `A-U`, `C-G` ou `G-U`. Ces tiges sont dites *maximales* car elles ne peuvent être étendues pour obtenir des tiges d'énergie libre inférieure selon ces règles. L'énergie associée à une tige est calculée en utilisant le modèle de Turner restreint aux empilements d'appariements, aux motifs des boucles terminales et aux mésappariements symétriques. Comme les tiges sont prédites indépendamment les unes des

autres, les règles relatives aux boucles internes et aux embranchements ne peuvent pas être appliquées à cette étape. De même, seules les boucles internes d'une longueur inférieure à huit nucléotides sont évaluées car on peut déjà affirmer à cette étape qu'aucune tige ne pourra être prédite dans cette région non appariée. Toutes les tiges sont énumérées par programmation dynamique avec une complexité spatiale et temporelle quadratique par rapport à la longueur de la séquence, puis filtrées selon leur valeur d'énergie libre grâce à une fonction de seuil. Cette fonction, établie de manière empirique, admet deux paramètres : la longueur de la tige et le taux en `G` et en `C` de la séquence pour tenir compte d'un éventuel biais favorisant la formation de tiges particulièrement stables.

Les points d'ancrage. CARNAC s'appuie sur des points d'ancrage entre les séquences pour guider et accélérer le repliement et l'alignement des séquences. Ces points d'ancrage sont des régions significativement conservées entre les deux séquences, sans insertion ni délétion. L'algorithme se déroule de la manière suivante :

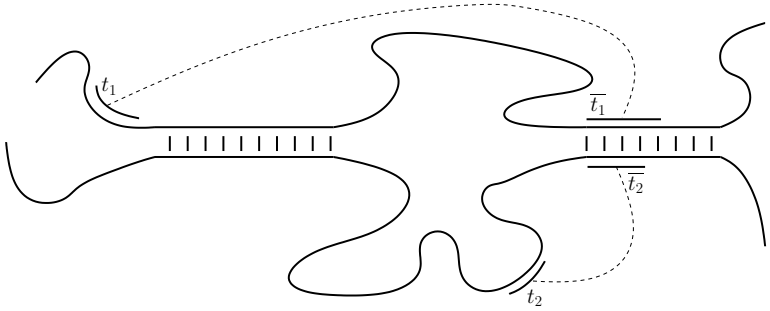
1. la recherche de tous les blocs *maximaux* conservés entre les deux séquences ;
2. le tri des blocs trouvés par score décroissant et filtrage des blocs chevauchants ;
3. la sélection gloutonne des blocs compatibles pour former des points d'ancrage.

Le système de score utilisé pour l'alignement est $+1$ en cas d'identité, -2 en cas de substitution. Un bloc maximal est un bloc qui ne peut être étendu pour atteindre un score plus élevé sans que ce score prenne une valeur négative ou nulle durant l'extension. Tout bloc maximal dont la taille et le score sont supérieurs à 8 est ainsi conservé. Les blocs conservés sont ensuite triés et filtrés pour éliminer les blocs qui impliquent au moins une même base d'une des deux séquences. Bien que drastique, ce critère permet d'éviter de trancher entre deux blocs qui pourraient *a priori* être corrects mais qui pourraient introduire une contrainte erronée dans la suite du déroulement de l'algorithme. Une fois triés, les blocs sont sélectionnés de manière gloutonne par score décroissant pour former des points d'ancrage. Un bloc est ainsi sélectionné s'il est compatible avec l'alignement local déjà construit. Sur l'exemple suivant, les blocs conservés rouges et bleus ne peuvent sélectionnés simultanément comme points d'ancrage car ils introduiraient une incohérence dans l'alignement des deux séquences.

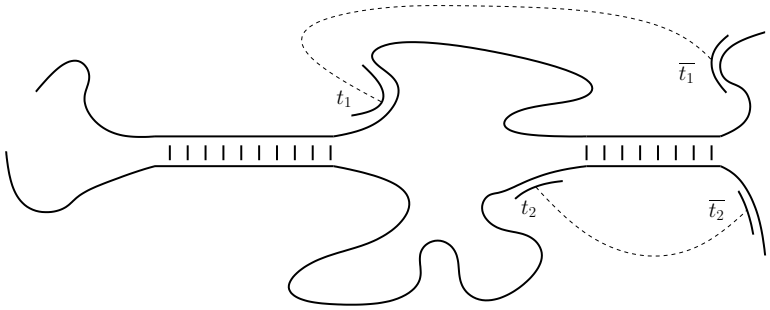


Le filtrage des tiges En fonction des points d'ancrage déterminés à l'étape précédente, les couples de tiges copliables sont énumérés. Deux tiges sont copliables si elles présentent au moins une covariation et si elles respectent les contraintes introduites par les points d'ancrage. Le terme de covariation est ici à prendre au sens fort du terme, c'est-à-dire en présence d'une mutation compensée : une covariation est comptée lorsque les deux bases d'un appariement sont mutées d'une tige à l'autre. La recherche de covariations s'effectue sur les deux tiges alignées sur leur structure primaire. Lorsqu'un couple de tiges présente au moins une covariation, les deux tiges sont dites copliables si elles sont compatibles avec les points d'ancrage, c'est-à-dire si les replier simultanément ne contredit pas l'alignement local des séquences selon les points d'ancrage. Deux tiges ne sont donc pas copliables si elles correspondent à l'un de ces trois cas :

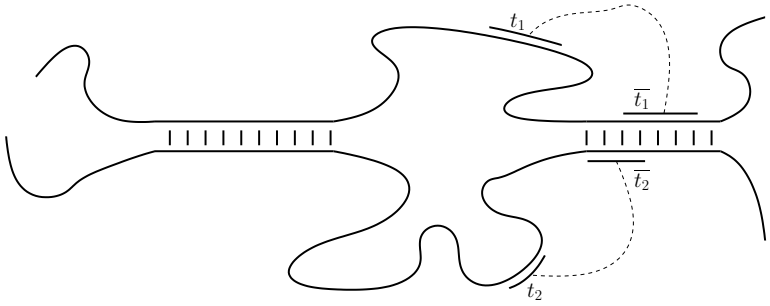
1. violation d'un point d'ancrage : si $(t_1, \overline{t_1})$ et $(t_2, \overline{t_2})$ étaient repliées simultanément, alors l'alignement de t_1 et t_2 contredirait l'alignement local au niveau du point d'ancrage.



2. décalage trop large à l'extérieur d'un point d'ancrage : lorsque l'ouverture ou la fermeture d'une tige tombe entre deux points d'ancrage, un décalage borné est autorisé. Ce décalage est variable selon les zones. Il dépend de la différence de longueur des fragments de séquences entre les points d'ancrage. Sur cet exemple, le décalage entre t_1 et t_2 est trop large.



3. décalage à l'intérieur d'un point d'ancrage : lorsque l'ouverture ou la fermeture d'une tige tombe à l'intérieur d'un point d'ancrage, aucun décalage n'est autorisé. Sur cet exemple, il y a un léger décalage entre $\overline{t_1}$ et $\overline{t_2}$.



A ce niveau, toute tige copliable avec une autre tige est conservée pour l'étape suivante. Une tige qui ne peut est copliée avec aucune tige est supprimée sauf si elle satisfait certains critères :

- il s'agit d'une tige-boucle, c'est-à-dire si la taille de la boucle est d'une longueur inférieure ou égale à huit nucléotides ;
- son énergie libre est relativement faible, en pratique le seuil est fixé à -1500 cal/mol ;
- elle se situe dans une région d'insertion potentielle, c'est-à-dire une région située entre deux points d'ancrage consécutifs significativement plus longue que dans l'autre séquence.

Le corepliement des tiges copliables Le corepliement des tiges est le cœur algorithmique de la prédiction de structure secondaire commune de caRNAc, c'est aussi son originalité. L'algorithme de repliement est une adaptation des récurrences de Sankoff, normalement appliquées au niveau nucléotidique, au repliement de tiges complètes. La complexité de l'algorithme de Sankoff n'est alors plus fonction de la taille des séquences mais du nombre de tiges potentielles. De plus, comme seuls les couples de tiges copliables sont considérés, la taille du problème se retrouve alors encore considérablement réduite.

Une tige t est caractérisée par les positions des extrémités de sa partie ouvrante, $t.leftopen$ et $t.leftclose$, et de sa partie fermante, $t.rightopen$ et $t.rightclose$, comme illustré sur la figure 3.11. Les récurrences de caRNAc reposent sur trois applications next, last et prev permettant à l'algorithme de naviguer entre les tiges, comme illustré sur la figure 3.12. A partir de l'ensemble $\mathcal{A}$ des n tiges potentielles d'une séquence s_a, on définit deux listes $\mathcal{A}_{\rightarrow}$ et $\mathcal{A}_{\leftarrow}$ ordonnées des tiges de $\mathcal{A}$.

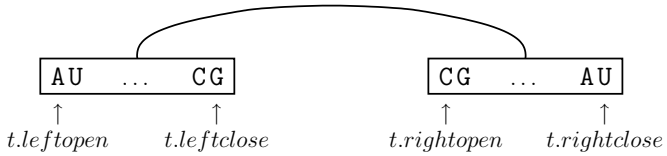


Fig. 3.11 – Chaque tige t est décrite par les positions des extrémités de sa partie ouvrante (gauche), $t.leftopen$ et $t.leftclose$, et de sa partie fermante (droite), $t.rightopen$ et $t.rightclose$.

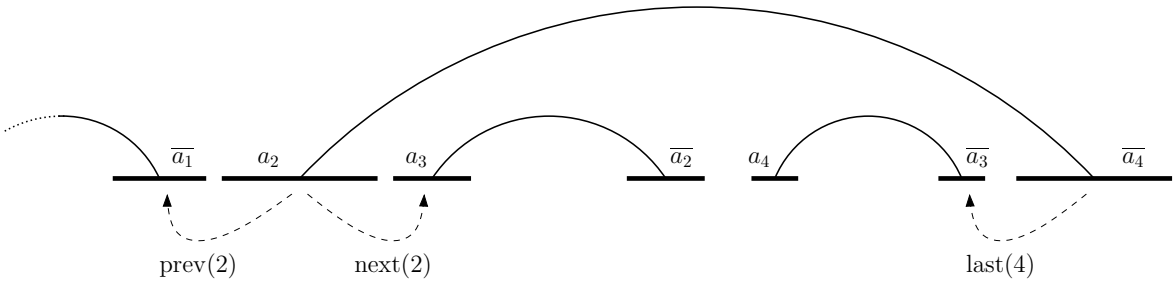


Fig. 3.12 –

$\mathcal{A}_{\rightarrow} = (a_1, a_2, \ldots, a_n)$ désigne la liste des tiges potentielles ordonnées par ordre croissant d'ouverture : $a_i \leq a_j$ si et seulement si $a_i.leftopen \leq a_j.leftopen$. L'application next : $[1..n] \longrightarrow [1..n]$ permet d'obtenir pour une tige t la prochaine tige dont la partie ouvrante ne chevauche pas celle de t :

$$\text{next}(i) = \min\{k \in [i+1..n] \mid a_i.leftclose < a_k.leftopen\}$$

$\mathcal{A}_{\leftarrow} = (\overline{a_1}, \overline{a_2}, \ldots, \overline{a_n})$ désigne la liste des tiges potentielles réordonnées par ordre croissant de fermeture : $\overline{a_i} \leq \overline{a_j}$ si et seulement si $\overline{a_i}.rightclose \leq \overline{a_j}.rightclose$. L'application last : $[1..n] \longrightarrow [1..n]$ permet d'obtenir pour une tige t la la dernière tige dont la partie fermante ne chevauche pas celle de t :

$$\text{last}(j) = \max\{k \in [1..j-1] \mid \overline{a_k}.rightclose < \overline{a_j}.rightopen\}$$

L'application prev : $[1..n] \longrightarrow [1..n]$ permet d'obtenir pour une tige t la tige précédente dont la partie fermante ne chevauche pas la partie ouvrante de t :

$$\text{prev}(i) = \max\{k \in [1..n] \mid \overline{a_k}.rightclose < a_i.leftopen\}$$

Les listes $\mathcal{B}_{\leftarrow}$ et $\mathcal{B}_{\rightarrow}$ sont respectivement analogues à $\mathcal{A}_{\leftarrow}$ et $\mathcal{A}_{\rightarrow}$ pour l'ensemble $\mathcal{B}$ des m tiges d'une seconde séquence s_b. Les applications next et last sont également étendues pour ces listes.

$$S(i,j,k,l) = \min \begin{cases} S(i,j,k,l-1) \\ S(i,j-1,k,l) \\ \min\limits_{1\leq x\leq n} \{S(i,\text{prev}(x),k,l) + S(\text{next}(x),\text{last}(j),0,0) + \text{bind}(a_x = \overline{a}_j, -)\} \\ \min\limits_{1\leq x\leq n} \{S(i,\text{prev}(x),0,0) + S(\text{next}(x),\text{last}(j),k,l) + \text{bind}(a_x = \overline{a}_j, -)\} \\ \min\limits_{1\leq y\leq m} \{S(i,j,k,\text{prev}(y)) + S(0,0,\text{next}(y),\text{last}(l)) + \text{bind}(-, b_y = \overline{b}_l)\} \\ \min\limits_{1\leq y\leq m} \{S(0,0,k,\text{prev}(y)) + S(i,j,\text{next}(y),\text{last}(l)) + \text{bind}(-, b_y = \overline{b}_l)\} \\ \min\limits_{\substack{1\leq x\leq n \\ 1\leq y\leq m}} \{S(i,\text{prev}(x),k,\text{prev}(y)) + S(\text{next}(x),\text{last}(j),\text{next}(y),\text{last}(l)) + \text{bind}(a_x = \overline{a}_j, b_y = \overline{b}_l)\} \end{cases}$$

La complexité spatiale de cet algorithme est en $\mathcal{O}(n^2m^2)$ et sa complexité temporelle en $\mathcal{O}(n^3m^3)$. Cependant, en pratique la complexité spatiale de l'algorithme est réduite à l'hyperdiagonale de la matrice grâce à un examen des tiges qui ne pourront être copliées et des points d'ancrage.

La combinaison des prédictions deux à deux

Pour n séquences, la première étape de caRNAc produit $n(n-1)/2$ couples de prédictions. Ces repliements sont ensuite combinés à l'aide d'un graphe afin d'obtenir une structure unique pour chaque séquence. Cette tâche se déroule en quatre étapes :

1. construction du graphe des tiges ;
2. remaniement et simplification du graphe ;
3. recherche de composantes connexes dans le graphe ;
4. sélection gloutonne pour chaque séquence des tiges pour former la structure finale.

Construction du graphe des tiges Le graphe des tiges est un graphe non dirigé où chaque nœud correspond à une tige apparaissant dans au moins un corepliement, et une arête entre deux nœuds indique que les tiges associées aux nœuds reliés ont été copliées. La figure 3.13 montre un exemple de graphe des tiges obtenu sur cinq ARN de transfert.

Remaniement et simplification du graphe Les tiges prédites par caRNAc ne peuvent pas comporter de renflement ni de boucle interne asymétrique. Une vraie tige peut donc avoir été scindée en deux tiges différentes lors de l'énumération des tiges potentielles. Pour pallier ce problème et par la même simplifier le graphe, les tiges emboîtées sont donc regroupées et les nœuds correspondants du graphe fusionnés, comme illustré sur la figure 3.14.

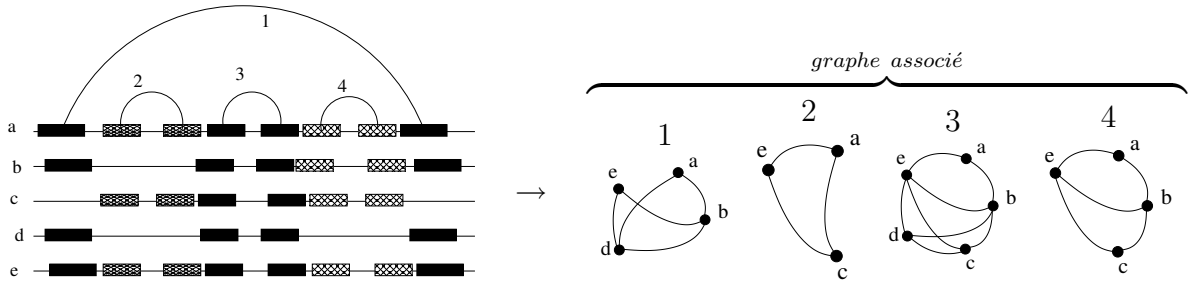


FIG. 3.13 – Graphe des tiges construit après les corepliements de cinq ARN de transfert.

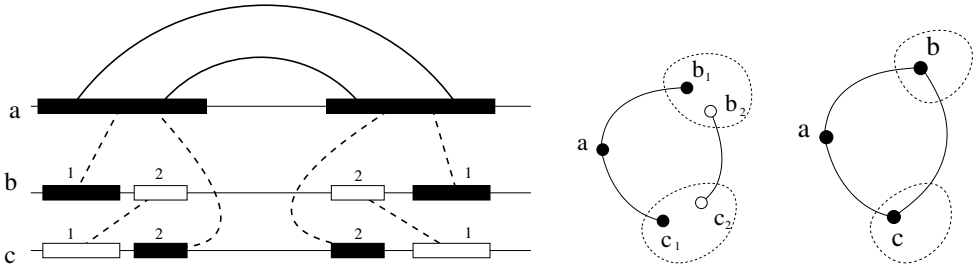


FIG. 3.14 – Regroupement de tiges lorsqu'elles sont emboîtées. La tige de la séquence a s'est repliée avec la tige b_1 et la tige c_2 tandis que les tiges b_2 et c_1 ont été copliées. Toutes ces tiges sont correctes, mais celles des séquences b et c sont considérées comme deux tiges distinctes emboîtées. Dans le graphe, elles sont fusionnées et les arêtes correspondantes sont regroupées.

Pour qu'un couple de tiges soit copliable, il est nécessaire que les tiges présentent au moins une covariation. Lorsque qu'un jeu de données comporte deux séquences proches partageant une structure commune, il est fort probable qu'une partie des tiges communes ne présentent aucune covariation et n'aient donc pas été copliées. Toutefois, ces tiges peuvent avoir été copliées par ailleurs et se retrouvent donc dans le graphe des tiges. Un deuxième type d'arête est donc introduit pour identifier les couples de tiges qui ne présentent pas de covariation : les arêtes étiquetées *identité*. Les arêtes qui correspondent à un corepliement de deux tiges seront étiquetées *coplié*. Cette modification permet d'améliorer la connexité du graphe en présence de tiges qui n'ont pas pu être copliées car elles ne présentaient pas de mutations compensatoires.

La recherche de composantes connexes dans le graphe Les composantes connexes du graphe des tiges correspondent à des ensembles de tiges qui ont pu être copliées et qui sont donc susceptibles de faire partie d'une éventuelle structure commune. Une composante connexe idéale dans le graphe des tiges est alors une clique comportant autant de nœuds que de séquences. Pour évaluer la qualité d'une composante connexe, un indice est calculé pour chacune en fonction du nombre de nœuds qu'elle contient, du nombre de séquences impliquées ainsi que du nombre d'arêtes étiquetées *coplié* et *identité*. L'indice d'une composante est le produit de deux indices : *node_index* qui mesure l'écart en terme de nombre observé de nœuds et le nombre idéal de nœuds, *edge_index* qui mesure l'écart entre terme d'arêtes par rapport au cas idéal.

$$node\_index = \left(\frac{Ns - (N - Ns)}{sq} \right)^2$$

où Ns est le nombre de séquences impliquées dans la composante, N est le nombre de nœuds dans la composante et sq est le nombre initial de séquences.

$$edge\_index = \frac{co}{me - id}$$

où co est le nombre d'arêtes étiquetées *coplié*, id le nombre d'arêtes étiquetées id et me le nombre d'arêtes possibles dans une clique comportant N nœuds, c'est-à-dire $\frac{N(N-1)}{2}$.

La sélection des tiges séquence par séquence Comme chaque tige appartient à une et une seule composante connexe, on attribue à une tige l'indice de la composante qui la contient. Pour chaque séquence, les tiges sont ensuite incorporées de manière gloutonne par indice décroissant jusqu'à un certain seuil. L'incorporation se fait également sous la contrainte de ne pas entrer de conflit avec la structure secondaire en cours de construction : les croisements d'appariements et les chevauchements de tiges sont ainsi interdits. Toutefois, un léger chevauchement est autorisé entre les tiges, et résolu en tronquant la tige la plus longue. Cette liberté par rapport aux contraintes initiales permet de récupérer *a posteriori* des tiges maximales légèrement chevauchantes qui n'auraient pas pu être repliées simultanément.

3.3.2 Introduction des méta-séquences

Le premier but de l'enrichissement de CARNAC est de mieux prendre en compte le schéma évolutif entre les séquences, quelque soit la distance évolutive qui les sépare. Les propriétés

qui ont guidé notre démarche sont que les approches "aligner puis replier" sont très performantes quand les séquences sont proches, alors que les approches "aligner et replier simultanément" sont plus robustes quand les séquences sont plus éloignées. Cela a été clairement établi dans [GG04]. L'idéal serait donc d'avoir une approche tout terrain, qui permette de traiter correctement des jeux de données hétérogènes, contenant des séquences à des distances évolutives quelconques. Pour cela, nous proposons une solution basée sur les méta-séquences, à l'image de ce que nous avons fait dans PROTEA.

Introduction des méta-séquences

Dans l'algorithme original de CARNAC, il existe une contrainte forte au repliement simultané de deux tiges : pour pouvoir être copliées, deux tiges doivent présenter une covariation. Ce critère permet de prédire des tiges pour lesquelles il existe une réelle évidence d'une évolution sous contrainte fonctionnelle. Cependant, ce critère de sélection peut poser problème face à un jeu de données qui comporte un sous ensemble de séquences fortement conservées. La redondance d'information apportée par des séquences proches perturbe ainsi le fonctionnement de l'algorithme. Pour traiter ce problème, on propose de regrouper en amont les séquences ressemblantes sous forme d'un alignement multiple, et d'adapter CARNAC pour ne plus travailler uniquement sur des séquences individuelles, mais sur des méta-séquences, c'est-à-dire des séquences individuelles et/ou des ensembles de séquences représentées par des alignements multiples.

La méta-tige L'introduction des méta-séquences nécessite la définition de la notion de tige sur un alignement multiple. Ceci nous amène à introduire le concept de *méta-tige*. Pour une méta-séquence simple, c'est-à-dire une méta-séquence correspondant à une séquence individuelle, une méta-tige est simplement une tige. Pour une méta-séquence représentée par un alignement multiple de n séquences, une méta-tige correspond à un ensemble de tiges, une sur chaque séquence, comme illustré sur l'exemple de la figure 3.15. Afin de construire un nombre raisonnable de méta-tiges à partir des tiges individuelles prédites sur les séquences qui composent une méta-séquence, on impose que les tiges formant une méta-tige partagent au moins trois appariements contiguës communs. Cette contrainte assure que chaque méta-tige contient au moins une tige par séquence qui répond à la contrainte imposée dans la version originale de CARNAC sur longueur minimale des tiges.

Définition 2 (Méta-tige)**.** *Une méta-tige $T = \{t_1, t_2, \ldots, t_n\}$ d'une méta-séquence $P = \{s_1, s_2, \ldots s_n\}$ est un ensemble comportant exactement n tiges tel que chaque tige t_i est une tige individuelle de la séquence s_i et*

$$\begin{aligned} \forall (t_i, t_j) \in T \times T \qquad & (t_i.leftopen - t_j.leftopen = t_j.rightclose - t_i.rightclose) \\ \wedge \quad & (\min(t_i.leftclose, t_j.leftclose) - \max(t_i.leftopen, t_j.leftopen)) \geq 3 \end{aligned}$$

L'énergie associée à une méta-tige est définie comme la moyenne des énergies des tiges individuelles qu'elle contient. Toutefois, on bonifie cette énergie pour chaque mutation qui préserve un appariement.

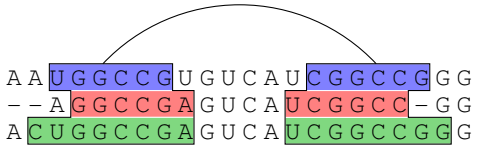


FIG. 3.15 – Exemple d'une méta-tige formée de trois tiges individuelles

La recherche de méta-tiges potentielles La recherche de méta-tiges dans une méta-séquence s'effectue en deux temps : l'identification des tiges potentielles individuellement dans chaque séquence représentée par la méta-séquence selon la procédure originale de CARNAC, puis le regroupement des tiges individuelles pour former les méta-tiges. Etant donné que l'on suppose fiable l'alignement multiple qui représente une méta-séquence, on s'appuie sur cet alignement pour regrouper les tiges individuelles. Une fois les tiges potentielles individuelles identifiées, les positions de ces tiges sont corrigées pour refléter leurs positions effectives dans l'alignement. Le création des méta-tiges se fait de manière progressive en partant de l'ensemble des tiges potentielles identifiées sur la séquence comportant le moins de tiges. Pour chacune de ces tiges on crée une méta-tige la contenant. On complète ensuite, séquence par séquence, les méta-tiges créées en incorporant à chaque méta-tige la tige qui partage un maximum d'appariements identiques en terme de positions sur l'alignement, et au minimum trois appariements communs. A la fin de ce processus, les méta-tiges incomplètes, c'est-à-dire celles qui ne contiennent pas exactement une tige par séquence représentée, sont détruites. Cette procédure a pour intérêt principal d'assurer que le nombre de méta-tiges potentielles n'explose pas avec le nombre de séquences puisque leur nombre est borné par le nombre de tiges d'une séquence.

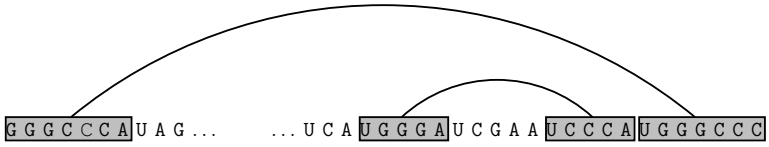
Les points d'ancrage entre méta-séquences La recherche de points d'ancrage est elle aussi adaptée pour traiter les méta-séquences et s'effectue directement entre les alignements multiples. La comparaison intra-méta-séquence n'est pas nécessaire car les séquences sont déjà alignées. Etant données deux méta-séquences représentées par deux alignements multiples $U = \{u_1, \ldots, u_m\}$ composé de m séquences alignées $u_1, \ldots, u_m$ et $V = \{v_1, \ldots, v_n\}$ composé de n séquences alignées $v_1, \ldots, v_n$, le score attribué à la comparaison de deux colonnes i et j respectives de U et V est obtenu en sommant les scores de toutes les comparaisons deux à deux entre la position i d'une séquence alignée de U et la position j d'une séquence alignée de V. Plus formellement, ce score est calculé par la relation suivante

$$s(i,j) = \sum_{1 \leq k \leq m} \sum_{1 \leq l \leq n} \texttt{score}(u_k[i], v_l[j])$$

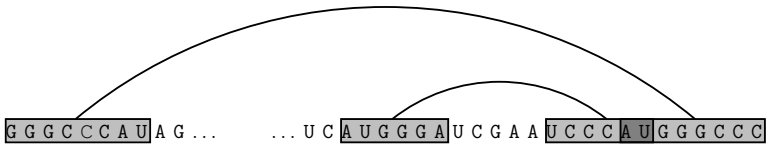
où $\texttt{score}(u_k[i], v_l[j])$ est le score attribué dans la version originale de CARNAC lors de la recherche de blocs conservés, c'est-à-dire $+1$ en cas d'identité entre $u_k[i]$ et $v_l[j]$ et -2 dans le cas contraire. On assimile la comparaison entre un gap et un nucléotide à une substitution. La procédure de sélection gloutonne des points d'ancrage reste identique à ceci près que le seuil sur le score est corrigé. Dans la version originale, le seuil minimal pour retenir un bloc est de 8. Comme le nombre de comparaisons réalisées est ici de $m.n$, ce seuil est maintenant de $8m.n$.

Le filtrage des méta-tiges Pour que deux tiges soient décrétées copliables dans caRNAc, il est nécessaire qu'elles présentent au moins une covariation et que leur repliement simultané soit compatible avec les points d'ancrage. Pour un couple de méta-tiges, ces définitions s'adaptent naturellement. Aucune covariation n'est attendue entre les tiges contenues dans une même méta-tige. Entre deux méta-tiges T_1 et T_2, on considère donc qu'une covariation existe si au moins une tige de T_1 présente une covariation avec au moins une tige de T_2. On pourrait exiger que chaque tige de T_1 présente au moins une covariation avec une tige de T_2, cependant, dans les faits le premier critère est suffisamment sélectif pour retenir les bons couples de méta-tiges. Pour la compatibilité avec les points d'ancrage en revanche, la modification obéit aux mêmes contraintes que si on avait à faire à des séquences individuelles : on impose que tous les couples de tiges (t_1, t_2) issus d'un couple de méta-tiges (T_1, T_2) soient compatibles avec les points d'ancrage.

Le chevauchement de tiges maximales L'algorithme tel qu'il est conçu ne permet pas de prédire simultanément deux tiges qui se chevauchent ne serait-ce que d'une base. Cette restriction peut poser un problème lorsque les vraies tiges d'une structure ne sont pas maximales et que leur extension entraîne un chevauchement comme illustré sur la figure 3.16.



(a) Deux tiges réelles non maximales.



(b) Tiges maximales correspondantes.

Fig. 3.16 – Les tiges réelles ne sont pas nécessairement maximales. Les tiges maximales (b) qui correspondent aux tiges de l'exemple (a) se chevauchent de deux bases et sont donc incompatibles entre elles.

Les tiges potentielles considérées dans caRNAc à l'issue de la première étape sont systématiquement des tiges maximales. Par conséquent, sur l'exemple de la figure 3.16 une seule des deux tiges pourrait donc être prédite.

La gestion des chevauchements est introduite dans l'algorithme en modifiant les définitions des applications next, last et prev (page 91). Soit δ le nombre de bases autorisées à se chevaucher, on redéfinit ces applications de la manière suivante

$$\begin{aligned} \text{next}(i) &= \min\{k \in [i+1..n] \mid a_i.rightopen < a_k.leftopen + \delta\} \\ \text{last}(j) &= \max\{k \in [1..j-1] \mid \overline{a_k}.rightclose < \overline{a_j}.leftclose + \delta\} \\ \text{prev}(i) &= \max\{k \in [1..n] \mid \overline{a_k}.rightclose < a_i.leftopen + \delta\} \end{aligned}$$

En pratique, on fixe $\delta = 2$, ce qui est suffisant pour rattraper les vraies tiges à partir de tiges maximales correspondantes, sans pour autant introduire de fausses tiges.

Le corepliement de méta-séquences L'adaptation de l'algorithme de Sankoff proposée dans caRNAc n'a pas à être modifiée pour pouvoir gérer les méta-séquences. Il est simplement nécessaire de définir une manière d'ordonner les méta-tiges afin de construire les ensembles $\mathcal{A}_{\rightarrow}$ et $\mathcal{A}_{\leftarrow}$, de définir l'énergie associée au repliement d'une méta-tige et au corepliement de deux méta-tiges.

$\mathcal{A}_{\rightarrow}$ désigne la liste des méta-tiges potentielles, rangées par ordre croissant de position d'ouverture. Pour deux méta-tiges $T_i = \{t_i^1, t_i^2, \ldots, t_i^n\}$ et $T_j = \{t_j^1, t_j^2, \ldots, t_j^n\}$ sur un alignement de n séquences, $T_i \leq T_j$ si et seulement si toutes les tiges individuelles vérifient cette relation

$$T_i \leq T_j \Leftrightarrow \forall k \in [1; n] \quad t_i^k.leftopen \leq t_j^k.leftopen$$

De même pour construire la liste $\mathcal{A}_{\leftarrow}$ des méta-tiges potentielles réordonnées par position de fermeture

$$T'_i \leq T'_j \Leftrightarrow \forall k \in [1; n] \quad t_i'^k.rightclose \leq t_j'^k.rightclose$$

L'énergie associée au repliement d'une méta-tige, ou au corepliement de deux méta-tiges, est égale à la somme des énergies des tiges individuelles repliées simultanément. Cette définition pose un problème car elle favorise les repliements individuels dans les méta-séquences qui représentent un grand nombre de séquences surtout lorsqu'elles sont comparées à des séquences classiques. On normalise donc l'énergie associée à une méta-tige en prenant la moyenne des énergies des tiges individuelles plutôt que leur somme.

Révision de l'implémentation de l'algorithme de Sankoff

Le corepliement de deux séquences dans caRNAc est une adaptation de l'algorithme de Sankoff réécrit pour travailler sur des tiges entières et non au niveau nucléotidique. Dans la section précédente, nous avons vu comment enrichir la version existante de cette heuristique pour traiter des méta-tiges. Le corepliement de tous les couples de séquences est l'étape la plus coûteuse de caRNAc. Nous présentons une révision de cet algorithme qui s'avère plus efficace en pratique.

L'un des choix opérés dans caRNAc est de ne prédire que les tiges "sûres", c'est-à-dire les tiges communes qui présentent des covariations. Dans les faits, caRNAc n'autorise donc le repliement de tiges individuelles que pour des petites tiges terminales dans des régions d'insertion, c'est-à-dire entre deux points d'ancrage séparés par des séquences de longueurs très différentes. Partant de cette constatation, nous proposons une implémentation de l'algorithme de corepliement de deux séquences restreint au corepliement de tiges. Cette restriction permet d'optimiser substantiellement l'efficacité de caRNAc sans pour autant diminuer ses performances.

Notre révision de l'algorithme est dérivée de Gardenia [BT06], une méthode d'alignement multiple de structures d'ARN développée dans l'équipe. L'idée est de ne pas stocker tous les calcul intermédiaires et de recalculer au besoin l'énergie optimale de deux fragments de séquences. Sur le papier la complexité spatiale est ainsi diminuée au détriment de la complexité temporelle. En pratique ce choix s'avère cependant plus judicieux car il permet de tirer plus aisément partie des différents mécanismes de mise en cache des machines actuelles.

L'implémentation que nous avons réalisée repose sur deux tables S et S_T de dimension $n \times m$ indexées par les listes des tiges ordonnées $\mathcal{A}_{\leftarrow}$ et $\mathcal{B}_{\leftarrow}$. Chaque cellule $S(j, l)$ contient l'énergie

optimale du repliement simultané des tiges a'_j et a'_l, c'est-à-dire l'énergie du corepliement de b'_j et b'_l ajoutée à l'énergie optimale du repliement des deux séquences entre les parties ouvrantes et fermantes de ces tiges. La table T est une table de travail pour les calculs intermédiaires. Etant donné que seules les parties fermantes des tiges sont indexées dans les tables S et S_T, on défini l'application $\text{open}(i) : [1..n] \longrightarrow [1..n]$ qui permet de localiser la partie ouvrante d'une tige $\overline{a_i}$ dans la liste $\mathcal{A}_{\rightarrow}$.

$$\text{open}(i) = \{k \in [1..n] | a_k = \overline{a_i}\}$$

Cette application est également définie pour l'ensemble des tiges $\mathcal{B}$ de la seconde séquence.

La table S est remplie par position d'ouverture de tige décroissante selon la règle suivante

$$S(j,l) \quad = \quad S_T(\text{last}(j), \text{last}(l)) + \text{bind}(\overline{a_j}, \overline{b_l})$$

où la table S_T est partiellement recalculée pour chaque couple (j,l). Les règles de remplissage de S_T dans le couple d'intervalles $([\text{prev}(\text{open}(j)); \text{last}(l)], [\text{prev}(\text{open}(j)); \text{last}(l)])$ sont les suivantes

$$S_T(i,k) = \min \begin{cases} S_T(i-1,k) & \\ S_T(i,k-1) & \\ S_T(\text{prev}(\text{open}(i)), \text{prev}(\text{open}(k))) + S(i,k) & \begin{array}{l}\text{si } \text{open}(i) \geq \text{next}(\text{open}(j)) \\ \text{et } \text{open}(k) \geq \text{next}(\text{open}(l))\end{array} \end{cases}$$

La dernière étape de l'algorithme consiste à remplir complètement la table S_T sans restriction particulière. A l'issue du remplissage, la cellule $S_T(n,m)$ contient l'énergie optimale du repliement simultané des tiges des deux séquences. Le rebroussement permettant de retrouver les structures s'effectue alors à l'aide d'une pile afin d'utiliser pleinement la table déjà calculée.

3.4 Résultats expérimentaux

Dans la section précédente, nous avons présenté les modifications apportées à caRNAc. Tout d'abord, la gestion de méta-séquences d'un bout à l'autre permet maintenant de traiter les ensembles de séquences hétérogènes en terme de conservation. Ensuite, la tolérance de petits chevauchements entre les tiges permet de pallier au problème inhérent à l'utilisation de tiges maximales alors que les tiges réelles ne le sont pas nécessairement. Enfin, la révision de l'algorithme de corepliement permet de trouver beaucoup plus rapidement une structure commune optimale. Cette dernière modification est particulièrement importante car elle nous offre la liberté de relâcher quelque peu les contraintes de filtrage des tiges potentielles et donc de considérer plus de tiges dans la suite de l'algorithme.

Afin d'apprécier les effets de ces modifications sur le comportement global de la méthode, nous l'avons évaluée sur BRAliBase I, le benchmark de référence des méthodes dédiée à la prédiction de structures communes. En fin de section, nous présentons nos résultats expérimentaux, à caractère plus exploratoire, en matière de prédiction d'ARN non-codants basée sur l'existence d'une structure commune prédite par caRNAc.

3.4.1 Validation sur BRAliBase I

Dans la section 3.1.3, nous avons présenté les résultats de la version 2004 de caRNAc sur le benchmark de référence BRAliBase I. Nous avons repris les données de BRAliBase I afin d'apprécier l'impact des modifications apportées à la version originale de caRNAc, notamment les méta-séquences.

caRNAc 2004 versus caRNAc 2008

Les résultats obtenus sur BRAliBase I par les deux versions de caRNAc sont synthétisés dans les tables 3.7 et 3.8. Globalement, on constate qu'on obtient de meilleurs résultats avec la nouvelle version de caRNAc, quelque soit le jeu de données et le degré de conservation. Le *MCC* est en effet toujours supérieur ou égal à celui atteint par la version 2004. En terme d'efficacité, la version 2008 de caRNAc s'avère beaucoup plus rapide, en particulier sur les séquences longues où le temps de calcul est au minimum divisé par 7.

Pour les ARN de transfert et les RNAse P, la sensibilité et la spécificité des structures prédites sont systématiquement supérieures ou égales aux anciennes valeurs. Pour les ARN ribosomiques en revanche, bien que le compromis sensibilité/spécificité soit globalement amélioré, on constate une légère perte de spécificité. Si l'on observe plus finement les structures prédites pour ces séquences, on remarque que les faux positifs supplémentaires sont des appariements qui appartiennent à la structure tertiaire. caRNAc prédit en réalité une tige de la structure tertiaire au détriment d'une autre tige moins stable de la structure secondaire. La tige prédite par caRNAc n'est pas considérée comme correcte dans le benchmark, bien qu'elle existe dans la structure réelle.

L'évolution des performances varie en fonction du degré moyen de conservation du jeu de données. Sur les jeux de données très conservés en moyenne, les résultats n'évoluent quasiment pas. C'est ici la faible quantité de mutations qui est en cause : une tige ne peut en effet être prédite que si elle présente au moins une covariation. En revanche sur les jeux de données moyennement conservés, les résultats sont nettement meilleurs grâce à l'utilisation des méta-séquences. Ces jeux de données comportent en effet des sous-ensembles de séquences très conservées qui, lorsqu'ils ne font pas l'objet d'un traitement particulier, perturbent la méthode. D'une part ces séquences présentent peu, voire pas du tout, de covariations et d'autre part introduisent une redondance d'information qui fausse les statistiques du graphe des tiges.

Famille	Conservation	caRNAc 2004				caRNAc 2008			
		Sens.	Spé.	MCC	Corrél.	Sens.	Spé.	MCC	Corrél.
tRNA	medium	75,0	93,8	0,836	84,4	100,0	100,0	1,000	100,0
	high	76,2	100,0	0,871	88,1	76,2	100,0	0,871	88,1
RNaseP	medium	59,4	95,0	0,750	77,2	61,5	96,7	0,770	79,1
	high	51,4	100,0	0,716	75,7	51,4	100,0	0,716	75,7
SSU	medium	39,9	93,2	0,610	66,6	53,4	91,5	0,699	72,5
	high	39,3	94,7	0,610	67,0	41,9	94,1	0,628	68,0
LSU	medium	46,8	97,8	0,676	72,3	50,0	95,8	0,692	72,9
	high	43,1	98,6	0,652	70,9	50,9	94,1	0,692	72,5

Tab. 3.7 – Résultats de caRNAc version 2004 et version 2008 sur BRAliBase I.

Famille	Conservation	Longueur	caRNAc 2004	caRNAc 2008	Accélération
tRNA	medium	73	0,125 s	0,052 s	2,40
	high	73	0,515 s	0,043 s	11,98
RNaseP	medium	377	48 s	0,941 s	51,00
	high	377	0,831 s	0,500 s	1,67
SSU	medium	1542	153 s	20 s	7,65
	high	1542	1149 s	22 s	52,23
LSU	medium	2904	2916 s	116 s	25,14
	high	2904	1394 s	97 s	14,37

Tab. 3.8 – Temps d'exécution de caRNAc sur BRAliBase I.

Les résultats obtenus avant et après modification de caRNAc sont donnés dans la table 3.9. Le repliement complémentaire par RNAfold permet d'améliorer les résultats globaux. RNAfold a en effet tendance à compléter les structures prédites par caRNAc par plus d'appariements corrects que de mauvais appariements. Cela se traduit par un meilleur *MCC*, c'est-à-dire un meilleur compromis sensibilité/spécificité, provenant d'une sensibilité qui augmente en moyenne de 22% alors que la spécificité ne diminue que de 15% en moyenne. Par rapport à la version 2004 de caRNAc, tous les résultats vont dans le sens de ces observations. Toutefois, pour les petites sous-unités ribosomiques, la tige de la structure tertiaire prédite par caRNAc induit en erreur le repliement thermodynamique en structure secondaire de RNAfold, ce qui diminue *de facto* la sensibilité et la spécificité.

Famille	Conservation	caRNAc 2004+RNAfold				caRNAc 2008+RNAfold			
		Sens.	Spé.	MCC	Corrél.	Sens.	Spé.	MCC	Corrél.
tRNA	medium	90,0	94,7	0,922	92,4	100,0	100,0	1,000	100,0
	high	100,0	100,0	1,000	100,0	100,0	100,0	1,000	100,0
RNaseP	medium	87,5	83,2	0,852	85,3	89,6	89,6	0,895	89,6
	high	70,8	66,2	0,684	68,5	70,8	66,2	0,684	68,5
SSU	medium	74,4	74,1	0,742	74,3	71,3	71,8	0,715	71,5
	high	78,6	79,5	0,790	79,0	72,7	74,8	0,737	73,8
LSU	medium	83,3	81,2	0,822	82,2	86,3	85,6	0,859	85,9
	high	80,9	79,3	0,801	80,1	83,9	82,5	0,832	83,2

Tab. 3.9 – Résultats de caRNAc dont les structures prédites sont complétées par RNAfold.

caRNAc et les méthodes existantes

Les modifications apportées à caRNAc améliorent ses performances sur BRAliBase I. Quand est-il des résultats de cette nouvelle version face aux autres méthodes ?

Les résultats des méthodes testées dans BRAliBase I sont reportées par jeu de données dans les tables 3.10 et 3.11. Les résultats de caRNAc présentées dans ces tables sont ceux de la version qui intègre les méta-séquences. Globalement, RNAalifold et caRNAc sont les deux méthodes les plus performantes, tous jeux de données confondus, surtout lorsque l'on considère les structures complétées par RNAfold. Bien que Pfold produise de meilleurs

(a) Résultats sur les ARN de transfert

Méthode	Conserv.	Sens.	Spé.	MCC
RNAalifold	medium	77,8	100,0	0,880
	high	90,5	100,0	0,950
Ilm	medium	100,0	75,0	0,863
	high	76,2	69,6	0,722
Pfold	medium	100,0	100,0	1,000
	high	95,2	100,0	0,975
caRNAc	medium	100,0	100,0	1,000
	high	76,2	100,0	0,871
Dynalign	medium	94,3	95,0	0,945
	high	54,8	54,5	0,535
FoldAlign	medium	23,8	33,3	0,268
	high	23,8	31,2	0,259
Structures prédites complétées par RNAfold				
RNAalifold	medium	100,0	100,0	1,000
	high	100,0	100,0	1,000
caRNAc	medium	100,0	100,0	1,000
	high	100,0	100,0	1,000

(b) Résultats sur les ARN de RNase P

Méthode	Conserv.	Sens.	Spé.	MCC
RNAalifold	medium	57,4	57,4	0,571
	high	78,9	77,8	0,782
Ilm	medium	70,4	55,1	0,620
	high	43,7	36,5	0,395
Pfold	medium	87,0	92,2	0,895
	high	66,2	88,7	0,765
caRNAc	medium	61,5	96,7	0,770
	high	51,4	100,0	0,716
Dynalign	medium	32,0	32,8	0,321
	high	40,3	39,6	0,397
FoldAlign	medium	5,2	22,7	0,107
	high	19,7	35,9	0,265
Structures prédites complétées par RNAfold				
RNAalifold	medium	61,1	67,3	0,639
	high	77,5	77,5	0,773
caRNAc	medium	89,6	89,6	0,895
	high	70,8	66,2	0,684

Tab. 3.10 – Résultats de BRAliBase I sur les ARN de transfert et sur les ARN de RNase P.

(a) Résultats sur les petites sous-unités ribosomiques

Méthode	Conserv.	Sens.	Spé.	MCC
RNAalifold	medium	84,4	92,1	0,881
	high	59,8	60,6	0,601
Ilm	medium	59,9	51,5	0,554
	high	51,3	43,0	0,469
Pfold	medium	-	-	-
	high	70,9	92,6	0,810
caRNAc	medium	53,4	91,5	0,699
	high	41,9	94,1	0,628
Dynalign	medium	-	-	-
	high	-	-	-
FoldAlign	medium	-	-	-
	high	-	-	-
Structures prédites complétées par RNAfold				
RNAalifold	medium	88,0	89,8	0,889
	high	59,3	58,3	0,588
caRNAc	medium	71,3	71,8	0,715
	high	72,7	74,8	0,737

(b) Résultats sur les grosses sous-unités ribosomiques

Méthode	Conserv.	Sens.	Spé.	MCC
RNAalifold	medium	75,0	92,1	0,831
	high	79,0	76,3	0,776
Ilm	medium	68,4	58,0	0,630
	high	49,0	39,3	0,438
Pfold	medium	-	-	-
	high	-	-	-
caRNAc	medium	50,0	95,8	0,692
	high	50,9	94,1	0,692
Dynalign	medium	-	-	-
	high	-	-	-
FoldAlign	medium	-	-	-
	high	-	-	-
Structures prédites complétées par RNAfold				
RNAalifold	medium	84,4	89,9	0,871
	high	79,2	77,3	0,782
caRNAc	medium	86,3	85,6	0,859
	high	83,9	82,5	0,832

Tab. 3.11 – Résultats de BRAliBase I sur les ARN des petites et grosses sous-unités ribosomiques.

résultats sur les séquences courtes, c'est-à-dire les ARN de transfert, et de longueur moyenne, les RNase P, il s'avère incapable de traiter les séquences plus longues pour des raisons de complexité et de applications numériques. En effet, pour effectuer une prédiction PFOLD calcule des probabilités qui peuvent être très faibles jusqu'à descendre sous la capacité du type primitif utilisé dans l'implémentation de PFOLD.

Sur les ARN de transfert, RNAALIFOLD et PFOLD sont globalement meilleurs que CARNAC. Pour PFOLD, ses excellents résultats sur ce jeu de données sont biaisés car ces séquences ont été utilisés pour entraîner la méthode. Sur les ARN de RNase P, CARNAC et PFOLD obtiennent les meilleurs résultats sur le jeu de données medium avec un MCC respectif de 0,77 et 0,895. CARNAC est légèrement plus spécifique que PFOLD avec une spécificité de 96,7% contre 92,2% pour PFOLD, mais PFOLD se montre beaucoup plus sensible. Si l'on compare les structures de CARNAC repliées par RNAFOLD à PFOLD, les compromis sensibilité/spécificité des deux méthodes sont strictement équivalents. Sur les ARN de RNase P très conservés (high), les meilleures prédictions sont produites par RNAALIFOLD avec un MCC de 0,782. On note cependant sur ces données que CARNAC est le seul à ne prédire aucun appariement incorrect puisque sa spécificité est de 100%, tout en prédisant plus d'un appariement sur deux de la structure réelle. De plus, sur cette structure prédite par CARNAC les appariements ajoutés par RNAFOLD n'améliore pas le compromis sensibilité/spécificité puisque le MCC passe de 0,716 à 0,684.

Les ARN ribosomiques sont des séquences particulièrement longues qui posent des problèmes de complexité à DYNALIGN et FOLDALIGN. La mémoire requise par ces méthodes pour replier ces séquences dépasse largement les capacités offertes par la machine utilisée pour le benchmark, c'est à dire 1Go. A l'exception du jeu de données ssu high où CARNAC s'avère la méthode la plus performante avec un MCC à 0,628, RNAALIFOLD est la méthode dont le compromis sensibilité/spécificité est meilleur sur les ARN ribosomiques. CARNAC tient ses objectifs puisque la spécificité des structures qu'il prédit ne descend jamais en dessous de 90%. Cette spécificité accrue est un atout important : les appariements prédits par CARNAC constituent des contraintes sûres pour guider un repliement purement thermodynamique. Sur les données très conservées, cette stratégie permet d'atteindre un MCC supérieur à RNAALIFOLD. Notamment dans le cas des petites sous-unités très conservées, le MCC de CARNAC+RNAFOLD vaut 0,737 contre 0,601 pour RNAALIFOLD et 0,588 pour RNAALIFOLD+RNAFOLD .

3.4.2 Vers la prédiction de gènes à ARN

Comme nous l'avons vu dans la section 3.2, l'existence d'une structure secondaire conservée représente une information majeure dans la prédiction de gène à ARN. Nous avons notamment vu comment RNAZ exploite cette information pour tenter de détecter des ARN non-codants homologues. ALIFOLDZ et RNAZ évaluent en effet la stabilité d'une structure commune prédite par RNAALIFOLD par rapport à une distribution d'énergie libre construite de manière empirique dans le cas d'ALIFOLDZ, approximée par apprentissage dans RNAZ. Le principal inconvénient lié à l'emploi de RNAALIFOLD est qu'il procède à une analyse comparative sur un alignement multiple, quelque soit le degré de conservation des séquences. Or, nous avons déjà montré à plusieurs reprises que les alignements de séquences faiblement conservées sont rarement fiables. Comme nous l'avons vu dans la section précédente, CARNAC fait partie des méthodes de prédiction de structure conservée les plus performantes. Nous proposons donc d'adapter le protocole d'ALIFOLDZ en exploitant les prédictions de CARNAC.

Cette expérience nous amène à confronter nos résultats à ceux des méthodes existantes qui procèdent à la prédiction d'ARN non-codants par analyse comparative.

L'existence d'une structure commune prédite par caRNAc

Nous avons dû adapter le protocole d'AlifoldZ pour deux raisons : caRNAc travaille sur des séquences non alignées, et caRNAc ne calcule pas une structure commune consensus mais une structure par séquence globalement partagée par toutes les séquences. Le protocole que nous utilisons est donc le suivant pour un ensemble S de n séquences :

- prédiction par caRNAc de la structure commune des séquences de S qui se traduit par l'obtention de n structures ;
- production d'un alignement A des séquences de S avec ClustalW ;
- mélange des positions de A à l'aide du script `shuffle_aln.pl` d'AlifoldZ afin d'obtenir 100 alignements mélangés ;
- prédiction par caRNAc de la structure commune des séquences de chacun des alignements mélangés ;
- calcul d'un z-score individuel des structures prédites pour les séquences de S.

A l'aide des n z-scores individuels obtenus pour chaque séquence de S, on réalise enfin une prédiction suivant un vote majoritaire : si plus de la moitié des n z-scores associés aux n structures prédites par caRNAc sont inférieurs à un certain seuil α, alors on prédit un ensemble d'ARN non-codants homologues. Dans le cas contraire, on considère que l'on ne se trouve pas en présence d'ARN non-codants homologues.

Afin d'évaluer le potentiel de cette approche, nous avons sélectionné de manière aléatoire quatre séquences de chaque famille présente dans Rfam. Pour chaque famille, nous avons calculé le z-score moyen des structures prédites par caRNAc par rapport à cinquante alignements mélangés de ces séquences. Pour apporter un contrôle négatif, nous avons constitué un jeu de données composé des "familles" de séquences extraites de dix alignements aléatoires générés par famille en mélangeant l'alignement structural des séquences originales. Les résultats obtenus sur ces deux jeux de données sont présentés sur la figure 3.17. Sur les familles d'ARN non-codants homologues, le z-score moyen des structures prédites par caRNAc est en moyenne inférieur à celui calculé sur les ensembles de séquences aléatoires. Les deux distributions du z-score moyen observées ne se détachent toutefois pas complètement, et leur intersection est relativement importante. En fixant à -1 le seuil sur le z-score calculé, environ 20% des séquences aléatoires sont prédits ARN non-codants à tort, et un peu moins de 70% ARN non-codants sont correctement détectés.

Nous avons également envisagé l'utilisation de Sissiz pour obtenir des alignements multiples de même composition en di-nucléotides (section 3.2.4). Toutefois, l'implémentation de Sissiz pose plusieurs problèmes. Sissiz approxime la distribution en di-nucléotides de l'alignement original de manière asymptotique, ce qui ne garantit pas d'obtenir strictement la même distribution, en particulier sur des séquences courtes comme c'est souvent le cas pour les ARN non-codants. D'autre part, Sissiz rencontre quelques problèmes numériques gênants pour une utilisation systématique. L'absence complète d'un di-nuléotide provoque sous certaines conditions une erreur fatale, de même qu'une répartition totalement équiprobable de tous les di-nucléotides qui est considérée comme une "absence de signature" significative. Pour toutes ces raisons, nous n'avons pas poussé nos investigations avec ce logiciel.

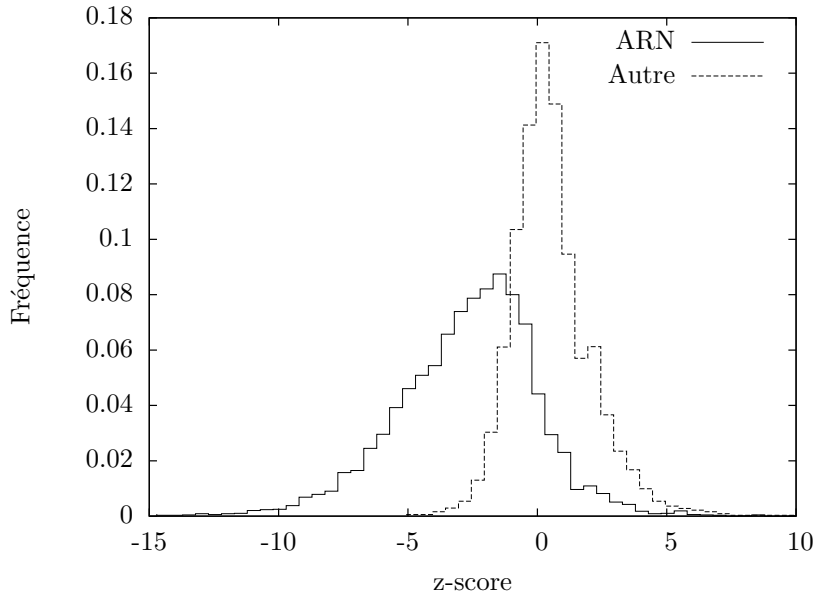


FIG. 3.17 – Répartition du z-score moyen observé de l'énergie libre des structures prédites par CARNAC sur les familles d'ARN non-codants de RFAM (trait plein), et sur des familles de séquences aléatoires (trait discontinu).

Les jeux de données d'évaluation

Nous avons décidé de faire varier plusieurs propriétés susceptibles d'influencer les performances des méthodes : la méthode d'alignement employée, la degré de conservation des séquences, le nombre de séquences et la qualité des structures des ARN non-codants. Les choix des séquences a donc été réalisé avec l'objectif de pouvoir construire des ensembles de séquences dont le pourcentage d'identité moyen varie de 40 à plus de 95%. Afin de mesurer le gain d'information apporté par l'utilisation de plusieurs séquences similaires, nous avons réalisé des alignements comportant de deux à cinq séquences. Nous n'avons pas construit d'alignement de plus de cinq séquences pour une raison simple : lorsque l'on dispose d'autant de séquences similaires, les outils d'inférence de structures communes peuvent suffire à détecter des ARN non-codants homologues. En effet, l'existence d'une structure commune à plus de dix séquences constitue un signal fort pour identifier des ARN non-codants homologues. Dans la section 3.2, nous avons vu que la détection des ARN non-codants à partir d'une séquence dépend de la qualité des structures des ARN à détecter. Pour évaluer l'influence de la qualité des structures sur la détection à partir de plusieurs séquences, nous avons sélectionné des familles d'ARN non-codants dont les structures communes ont une stabilité variable. Nous avons également retenu quelques familles dont les structures communes comportent des pseudonœuds pouvant gêner la prédiction.

A partir de ces propriétés, nous avons recueilli trois types de données provenant de plusieurs organismes : des ARN non-codants homologues pour évaluer la sensibilité, des fragments codants homologues d'ARN messagers et des séquence aléatoires pour évaluer la spécificité. La répartition des données est la suivante : 21 familles d'ARN non-codants, 15 familles d'ARN messagers et 21 "familles" de séquences aléatoires. La majorité des familles d'ARN non-

codants retenues proviennent de RFAM. Afin d'assurer la variabilité des paramètres définies précédemment nous avons ajouté deux familles de micro ARN provenant de miRBase. Dans la section 1.3, nous avons vu que des fragments d'ARN messagers, les introns et les extrémités $3'$ et $5'$ non traduites, sont susceptibles de contenir des structures. Pour produire des séquences qui ne contiennent *a priori* pas de structure, nous avons retiré ces fragments des ARN messagers que nous avons utilisés. Les ARN messagers sont en général composés de plusieurs milliers de bases, contrairement aux ARN non-codants dont la longueur dépasse rarement les 300 bases. Pour constituer un jeu de données comparable au jeu de données positif d'ARN non-codants, certains alignements ont par conséquent été tronqués. Le second jeu de données négatives, les *shuffles*, est composé d'alignements positifs mélangés par la procédure employée dans AlifoldZ [WH04] (section 3.2.4).

Le protocole expérimental

Trois des méthodes les plus récentes ont été testées sur les jeux de données ainsi constitués : QRNA [RE01], ddbRNA [DBDH03] et RNAz [WHS05]. Contrairement à ddbRNA et RNAz qui effectuent une classification binaire "ARN non-codants homologues"/"autre", QRNA effectue une classification en trois classes : "codant", "non-codant" et "autre". L'objectif de nos tests est d'évaluer les performances de détection des ARN non-codants. De notre point de vue, les classes "codant" et "autre" sont équivalentes car elles ne correspondent pas à la prédiction d'ARN non-codants homologues. Ces deux classes sont donc fusionnées. Pour évaluer les performances des méthodes nous utilisons les trois notions classiques, à savoir la sensibilité, la spécificité et le cœfficient de corrélation de Matthews. La sensibilité désigne ici la proportion d'ARN non-codants homologues détectés, la spécificité la proportion d'ARN messagers et d'alignements mélangés non prédits comme des ARN non-codants homologues.

Pour chaque famille, des ensembles de deux, trois et cinq séquences sont créés aléatoirement. Chaque ensemble de séquences est ensuite aligné et soumis aux différentes méthodes, caRNAc étant la seule méthode qui ne nécessite pas d'alignement préalable. Au total, plus de 80 000 alignements ont ainsi été constitués.

L'influence de l'alignement

Nous avons fait appel à cinq méthodes d'alignement largement utilisées : Blast [WBB+08] et Needleman&Wunsch [NW70] pour les alignements deux à deux, ClustalW [THG94], T-Coffee [NHH00] et Dialign2-2 [Mor99] pour les alignements de deux séquences et plus. Selon la méthode utilisée, les résultats varient sensiblement. En moyenne, les meilleurs résultats sont obtenus sur les alignements produits par ClustalW (figure 3.18). Cette observation globale se vérifie sur les résultats moyens de chaque méthode de détection, quelque soit le nombre de séquences utilisées.

RNAz a été entraîné à reconnaître les ARN non-codants sur des alignements produits par ClustalW. Il est donc normal qu'il obtienne de meilleurs résultats sur ce type d'alignements. Les performances de ddbRNA sur les alignements de ClustalW s'expliquent par un nombre moyen de gaps moins important dans ces alignements que dans les alignements de Dialign2-2 et T-Coffee. En effet, les gaps sont une entrave à la recherche de tiges pratiquée dans ddbRNA : les positions qui contiennent au moins un gap sont ignorées et ne contribuent donc pas à la formation des tiges. Les résultats obtenus sur les alignements produits par Blast et Needleman&Wunsch sont en moyenne inférieurs aux alignements de deux séquences

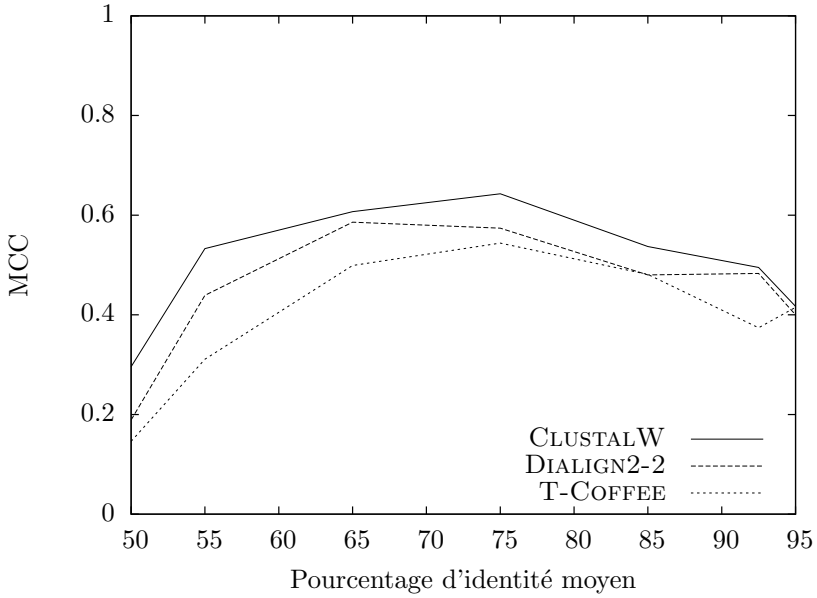


Fig. 3.18 – Résultats moyens des méthodes de détection selon la méthode d'alignement multiple utilisée. Les résultats sont exprimés en fonction du pourcentage d'identité des alignements. Ces résultats sont calculés à partir de tous les alignements de deux, trois et cinq séquences.

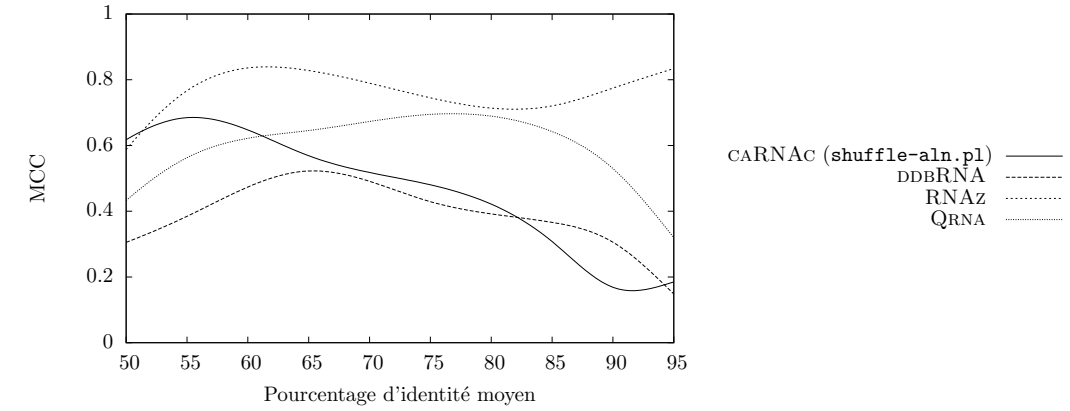
produits par ClustalW (table 3.12). Cette observation est notamment valable pour Qrna qui a pourtant été entraîné à reconnaître les ARN non-codants sur des alignements produits par Blast. Ces résultats nous amènent à nous focaliser sur les alignements générés par ClustalW.

Méthode d'alignement	Sensibilité moyenne (en %)			Spécificité moyenne (en %)		
	ddbRNA	RNAz	Qrna	ddbRNA	RNAz	Qrna
Blast	12,3	42,0	36,1	98,5	93,8	93,4
Needleman&Wunsch	18,3	55,1	23,6	97,5	95,5	98,1
ClustalW	26,7	71,9	41,3	97,2	95,4	98,2

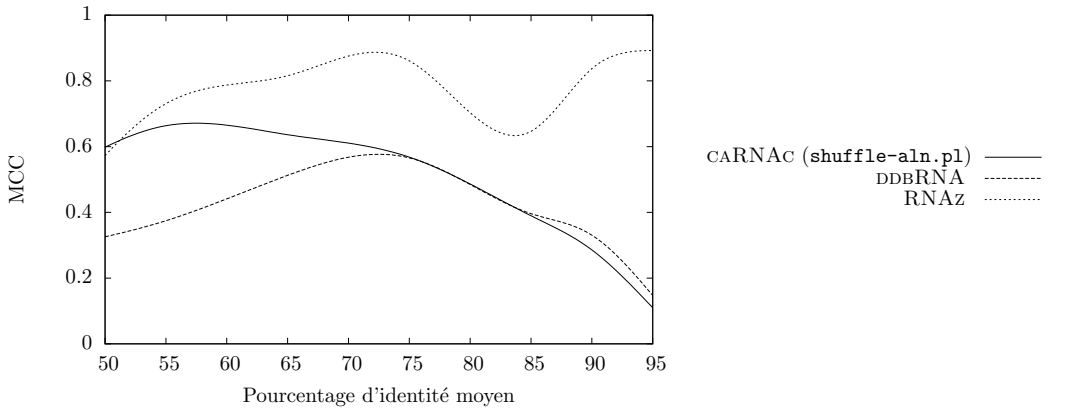
Tab. 3.12 – Résultats moyens obtenus sur des alignements produits par Blast, Needleman&Wunsch et ClustalW avec deux séquences.

L'influence du nombre de séquences

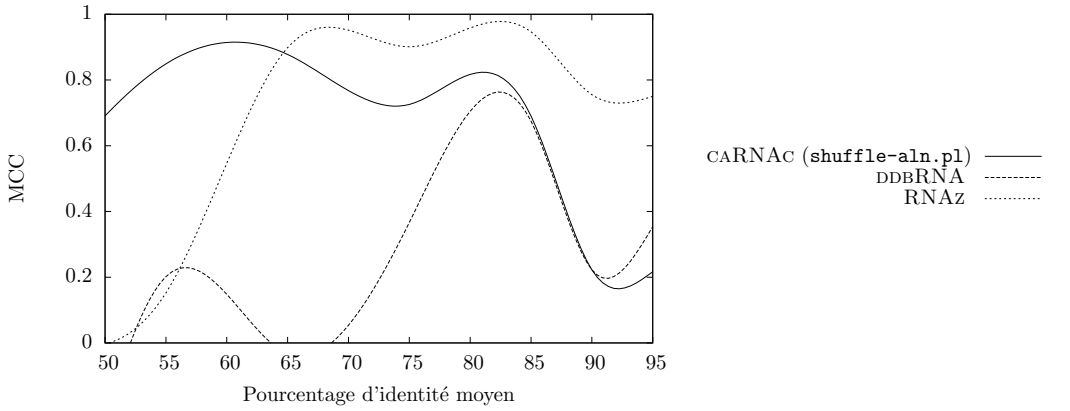
Le table 3.13 et la figure 3.19 donnent les résultats obtenus selon le nombre de séquences utilisées. Les performances de ddbRNA et de RNAz sont en moyenne meilleures sur des alignements de trois séquences. Toutefois, pour RNAz, la sensibilité est bien plus élevée en utilisant cinq séquences. Quant à caRNAc, ses résultats croissent strictement avec le nombre de séquences utilisées.



(a) Alignements de deux séquences.



(b) Alignements de trois séquences.



(c) Alignements de cinq séquences.

FIG. 3.19 – Résultats de QRNA, ddbRNA, RNAz et caRNAc en fonction du nombre de séquences utilisées. Les résultats sont exprimés à l'aide du cœfficient de corrélation de Matthews, en fonction du pourcentage d'identité des alignements.

(a) Résultats de ddbRNA

Nb séq.	Sensibilité (en %)	Spécificité (en %)	MCC
2	27,1	97,1	0,309
3	31,5	97,1	0,335
5	27,0	98,3	0,252

(b) Résultats de RNAz

Nb séq.	Sensibilité (en %)	Spécificité (en %)	MCC
2	78,9	90,4	0,697
3	78,2	93,6	0,704
5	76,6	93,9	0,639

(c) Résultats de caRNAc

Nb séq.	Sensibilité (en %)	Spécificité (en %)	MCC
2	46,4	89,7	0,409
3	63,3	82,9	0,472
5	70,2	90,6	0,628

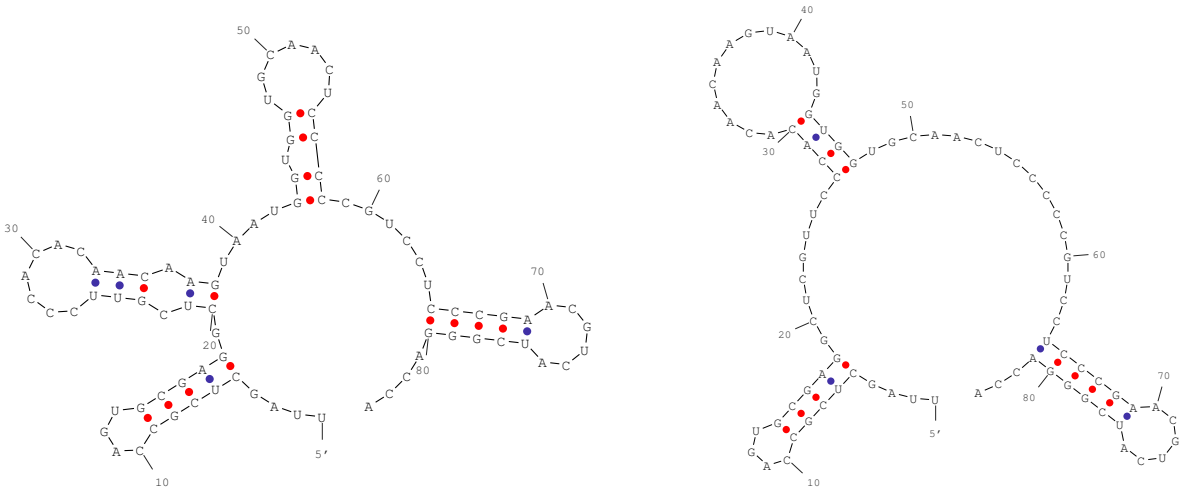
Tab. 3.13 – Résultats selon le nombre de séquences utilisées. Ces résultats sont établis sur les alignements réalisés avec ClustalW, et exprimés en pourcentage pour la sensibilité et la spécificité. La spécificité moyenne est calculée sur les ARN messagers et les shuffles.

L'influence de la conservation

La conservation entre les séquences est un paramètre dont l'influence varie suivant la méthode employée. En moyenne, les meilleurs résultats proviennent des alignements dont le pourcentage d'identité moyen est compris entre 60% et 85%. caRNAc est la seule méthode capable de traiter convenablement des jeux de données dont la conservation moyenne est inférieure à 60% d'identité. Entre 60% et 85% d'identité, toutes les méthodes ont une spécificité moyenne supérieure à 80%. Néanmoins, hors de cet intervalle la spécificité moyenne reste élevée et ne descend pas en dessous de 75%. Par contre, la sensibilité de Qrna et de ddbRNA se dégrade rapidement lorsque l'on dépasse 90% d'identité moyenne.

La spécificité moyenne sur les shuffles est à peu près équivalente à la spécificité moyenne sur les ARN messagers. Lorsque le pourcentage d'identité est inférieur à 80%, la spécificité moyenne sur les shuffles est très légèrement inférieure à celle des ARN messagers ; la situation s'inverse au delà de 80% d'identité.

L'influence de la conservation est en réalité étroitement liée à la qualité d'alignement. En effet, des séquences homologues mal conservées partagent une structure commune que toutes les méthodes ont dû mal à prédire à partir d'un alignement qui n'est pas correct. C'est également la raison pour laquelle caRNAc est la seule méthode qui produise des résultats probants en dessous de 60% d'identité moyenne. Sur la figure 3.20 est représentée à gauche la structure secondaire d'un ARN non-codant présent dans la partie 3′ de l'ARN des virus de la famille des pomovirus. Sur cette figure est également présenté un alignement de trois séquences homologues de ce type d'ARN non-codant produit ClustalW et extrait de notre jeu de données. Bien que plus de 90% des positions de cet alignement soient correctes, seul caRNAc prédit des séquences d'ARN non-codants homologues. La structure prédite par caRNAc sur ces séquences est présentée en partie droite de la figure 3.20.



(a) Structure réelle.

(b) Structure prédite par caRNAc.

```
T91413.1      UUAGCUCGC-CAGUGCGAGGCCUCUUCCUACACAAGAGGUAU---UGG-GGUGCGACUCCCCCGUCUAUCCUGAACGUCAUCAGGACCA
X54354.1      UUAGCUCGC-CAGUGCGAGGCUCGUUCCCACACAACAAGUAA---UGGUGGUGCAACUCCCCCGUCC-UCCCGAACGUCAUCGGGACCA
Y16104.1      UAAUUGAGGACAGUUCCUCUCCCUCUAGCACACAGA-GGUCAAACUGGGUG--CAACUCCCCCC-CCUUCCGUGG-GUAACGGAAACC-
```

(c) Alignement extrait de Rfam.

```
T91413.1      UUAGCUC--GCCAGUGCGAGGCCUCUUCCUACACAAGAGGUAUUGG-GGUGCGACUCCCCCGUCUAUCCUGAACGUCAUCAGGACCA
X54354.1      UUAGCUC--GCCAGUGCGAGGCUCGUUCCCACACAACAAGUAAUGGUGGUGCAACUCCCCCGUCC-UCCCGAACGUCAUCGGGACCA
Y16104.1      -UAAUUGAGGACAGUUCCUCUCCCUCUAGCACACAGAGGUCAAACUGGGUGCAACUCCCCC--CCCUUCCGUGGGUAACGGAAACC-
                        *************************          **********          *************
```

(d) Alignement produit par ClustalW. Le symbole * marque les positions correctes.

Fig. 3.20 – Structure secondaire réelle (à gauche) et structure prédite par caRNAc (à droite) d'un ARN non-codant présent dans l'ARN des pomovirus, ici celle du *Cacao yellow mosaic virus*. En partie inférieure, l'alignement produit par ClustalW de ladite séquence et de deux séquences homologues ainsi que l'alignement structural correspondant extrait de Rfam (RF00233). Ces séquences ont un pourcentage d'identité moyen est de 66%.

Les conclusions

Face à ses concurrents, caRNAc tire son épingle du jeu sur les séquences faiblement conservées où il est le seul à fournir des résultats pertinents. Sur ce type de données, les autres méthodes sont induites en erreur par un alignement incorrect. Sur les séquences relativement bien conservées en revanche, RNAz se dégage nettement de toutes les méthodes existantes en terme de sensibilité. Au niveau spécificité, Qrna, ddbRNA et RNAz sont à peu près équivalentes, bien que Qrna soit nettement plus spécifique sur les ARN messagers grâce à son modèle pour détecter les séquences codantes homologues.

Contrairement à caRNAc, la conception de RNAz repose sur un système d'apprentissage très sophistiqué entraîné sur un très grand nombre de séquences issues d'un large éventail de familles d'ARN non-codants. Cette caractéristique est un point fort pour RNAz lorsque le jeu de données qu'on lui soumet répond aux critères pour lesquels il a été entraîné. Cet atout peut toutefois s'avérer limitant, notamment en ce qui concerne le nombre de séquences. En effet, RNAz ne peut pas traiter de jeux de données comportant plus de dix séquences car son processus d'apprentissage a été limité à des jeux de données contenant au plus dix séquences. caRNAc en revanche n'est pas limité en nombre de séquences. Qui plus est, les différentes expériences que nous avons présentées montre que les performances de caRNAc augmentent avec le nombre de séquences. Cette propriété lui confère un net avantage par rapport aux méthodes existantes dont les performances sont limitées par les difficultés à produire un alignement multiple d'un grand nombre de séquences.

Chapitre 4

Deux exemples d'intégration de Protea et caRNAc

Au cours des chapitres 2 et 3 nous avons présenté deux méthodes que nous avons mis au point pour la prédiction de séquences codantes homologues et de séquences non codantes qui partagent une structure. L'originalité de ces méthodes réside dans le traitement d'ensembles de séquences non alignées par analyse comparative qui permet d'obtenir des résultats significatifs sur des séquences faiblement conservées. De plus, ces méthodes tirent parti du concept que nous avons introduit, les méta-séquences (section 1.5.3), qui permet d'éliminer les redondances de séquences au sein d'un jeu de données et donc de traiter des ensembles de séquences hétérogènes en terme de conservation.

Dans ce chapitre, nous nous intéressons à l'intégration de ces méthodes dans deux projets collaboratifs réalisés au sein de l'équipe. Dans la section 4.1, nous présentons Magnolia, une méthode d'alignement multiple de séquences nucléiques fonctionnelles basée sur les prédictions de Protea et de caRNAc. Dans la section 4.2, nous présentons un pipeline d'annotation par génomique comparative.

4.1 L'alignement multiple de séquences nucléiques

Au cours des chapitres précédents, nous avons fait mention de bon nombre d'outils d'alignement pour identifier des séquences similaires dans un banque de données (sections 2.2 et 3.2.3) ou pour fournir un objet d'étude en vue d'une analyse comparative (section 1.5). On peut regrouper les outils d'alignement de séquences nucléiques en deux groupes. D'une part, les outils génériques qui cherchent à maximiser les ressemblances entre les séquences d'un point du vue syntaxique, c'est-à-dire à mettre en relation un maximum d'acides nucléiques identiques. Ces méthodes, exactes ou heuristiques, ne nécessitent aucune connaissance *a priori* sur les séquences à aligner. D'autre part, les outils plus spécifiques et sophistiqués qui supposent l'existence d'une fonction commune partagée par les séquences à aligner pour proposer un alignement respectueux de cette fonction. C'est ce dernier type de méthode auquel nous nous intéressons ici.

Dans un premier, nous présentons les méthodes dédiées à l'alignement de séquences codantes homologues. Ensuite, nous présentons les méthodes dédiées à l'alignement de séquences qui partagent une structure commune. Puis, nous présentons Magnolia, la méthode d'alignement multiple de séquences codantes homologues et non codantes qui partagent une structure

commune que nous avons développée. Enfin, nous terminons cette section par une évaluation des performances de MAGNOLIA.

4.1.1 L'alignement multiple de séquences codantes homologues

Parmi les méthodes de prédiction de séquences codantes présentées dans la section 2.2, certaines proposent en sortie un alignement qui tient compte des séquences d'acides aminés codées. Toutefois, toutes ces méthodes (PROCRUSTES [GMP96], GENEWISE [BCD04], GENOMESCAN [YLB01], ORFGENE2 [RMK96], PREDICTGENES [GHKB00], GENOMETHREADER [GBSK05]) nécessitent une connaissance *a priori* de la séquence d'acides aminés, ou extraient cette séquence des banques publiques. Leur fonctionnement consiste à réaliser un alignement d'une séquence d'acides aminés de référence contre une séquence nucléique supposée codée une séquence d'acides aminés identique ou similaire.

En ce qui concerne les méthodes de prédiction par analyse comparative (section 2.3), la situation est différente car la quasi totalité de ces méthodes travaillent sur des séquences déjà alignées et ne proposent donc pas d'alignement en sortie. Les seules méthodes qui produisent en sortie des alignements des séquences nucléiques prédites comme des séquences codantes homologues sont les méthodes qui travaillent sur des séquences génomiques complètes. Néanmoins, ces méthodes sont toutes restreintes à l'analyse des couples d'espèces précis et ne savent pas traiter des séquences hors de leur contexte génomique.

A notre connaissance, il n'existe finalement qu'un seul logiciel, DIALIGN2-2 [Mor99] qui réalise l'alignement d'un ensemble de séquences nucléiques en fonction des séquences d'acides aminés potentielles qu'elles peuvent coder. Le principe de DIALIGN2-2 repose sur l'identification de segments conservés, c'est-à-dire des fragments de séquences qui s'alignent correctement sans insertion ni délétion, incorporés de manière gloutonne pour former un alignement complet. Pour un ensemble de n séquences, DIALIGN2-2 commence par rechercher tous les segments conservés entre tous les couples de séquences. Pour chaque couple de séquences, les segments obtenus sont regrouper pour former des ensembles cohérents sans croisement ni chevauchement, c'est-à-dire qu'ils doivent pouvoir faire partie d'un même alignement, appelés des diagonales. Ensuite, les diagonales obtenues pour tous les couples de séquences sont incorporées de manière gloutonne pour former un alignement multiple. En plus de pouvoir identifier les segments entre les séquences nucléiques fournies, DIALIGN2-2 propose d'identifier ces segments au niveau peptidique. Chaque séquence est alors traduite systématiquement selon les trois cadres de lecture possibles pour le brin donné, puis les segments sont identifiés entre les couples de traductions en utilisant la matrice BLOSUM62. La construction des diagonales se fait ensuite à partir de l'ensemble des segments provenant des trente six couples de traductions potentielles obtenus pour un couple de séquences, ce qui permet de supporter la présence de décalages de cadres de lecture. La suite de l'algorithme est inchangée, et le retour aux séquences nucléiques se fait à l'issue de la construction de l'alignement multiple au niveau peptidique.

4.1.2 L'alignement multiple de séquences partageant une structure commune

L'alignement de séquences possédant une structure secondaire commune est un problème qui a beaucoup intéressé la communauté ces cinq dernières années, en partie à cause des nombreuses découvertes de petits ARN non-codants. A l'heure actuelle on dénombre plus

d'une dizaine de méthodes dédiées à ce problème de produire un alignement structural multiple avec inférence de la structure. Parmi ces méthodes, certaines ont déjà été mentionnées dans la section 3.1.2 car en plus d'aligner les séquences, elle réalise une prédiction explicite de la structure secondaire conservée : DYNALIGN, FOLDALIGNM, STEMLOC, PMMULTI, MLOCARNA, STRAL, MURLET et MXSCARNA. Il existe cependant d'autres méthodes qui réalisent un alignement mais qui ne produisent pas en sortie la structure détectée telles que R-COFFEE [WHN08, MWH+08] et LARA [BKR07]. Sans entrer dans les détails, R-COFFEE et LARA sont deux méthodes qui utilisent T-COFFEE pour réaliser un alignement multiple en faisant varier son système de score en fonction des structures prédites sur les séquences individuelles par approche thermodynamique (section 3.1.1).

En marge des méthodes qui intègrent dans leur processus la prédiction de structures individuelles ou communes, il existe une autre famille de méthodes qui s'appuient sur des structures connues ou des prédites : RNAFORESTER [HTGK03, HVG04], MARNA [SB05], MIGAL [AS05] ou encore GARDENIA [BT06]. Contrairement aux autres, MIGAL est limité à l'alignement de deux séquences.

4.1.3 Magnolia, alignement de séquences fonctionnelles homologues

MAGNOLIA [FdMT08] admet en entrée un ensemble de séquences nucléiques non alignées et produit en sortie un alignement multiple de ces séquences en fonction de leur nature. MAGNOLIA est en réalité l'agrégation de trois méthodes développées dans l'équipe : PROTEA (section 2.4), CARNAC (section 3.3) et GARDENIA [BT06]. La figure 4.1 résume de manière schématique le fonctionnement de MAGNOLIA. Dans un premier temps, la fonction commune des séquences est prédite au moyen de PROTEA et de CARNAC+GARDENIA . En fonction des prédictions réalisées, MAGNOLIA produit les alignements multiples correspondants générés par PROTEA pour les séquences codantes et par GARDENIA pour les séquences non codantes qui possèdent une structure commune.

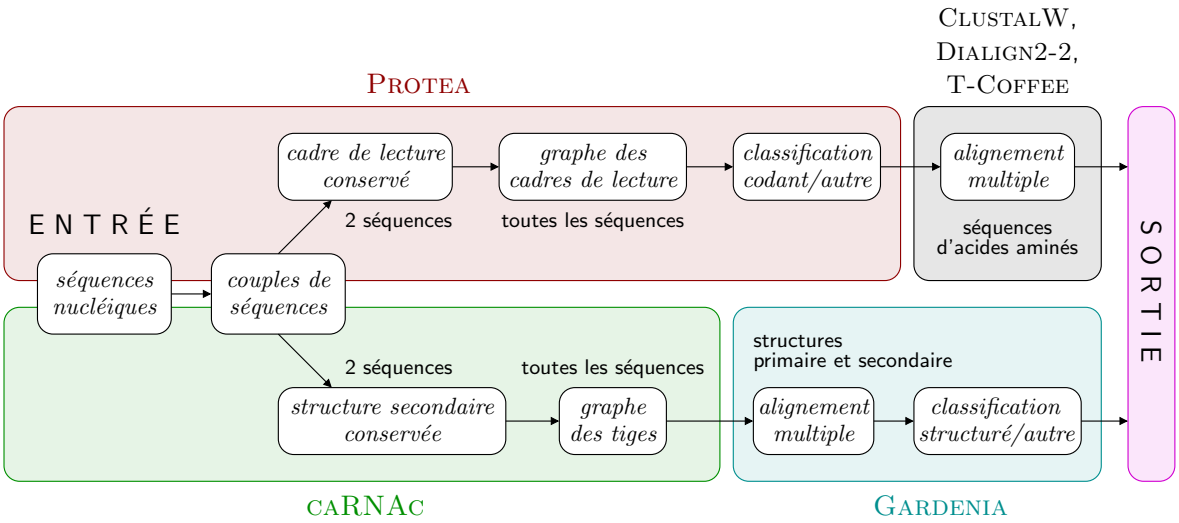


FIG. 4.1 – L'enchaînement des modules qui composent MAGNOLIA.

Prédiction et alignement de séquences codantes

PROTEA (section 2.4) peut être utilisé pour améliorer l'alignement de séquences codantes homologues, en particulier sur des séquences nucléiques divergentes. Pour ce faire, nous avons défini le protocole suivant constitué de trois étapes. La première étape consiste à utiliser PROTEA pour rechercher un cadre de lecture conservé. Si PROTEA détecte des séquences codantes homologues, alors les séquences d'acides aminés codées par les cadres de lectures prédits sont alignées à l'aide d'une méthode d'alignement multiple classique. La méthode d'alignement utilisée ici est indépendante du reste de notre procédure. Enfin, l'alignement multiple des séquences d'acides aminés est rétro-transcrit pour obtenir un alignement multiple des séquences nucléiques initiales. Cette rétro-transcription fait appel à un algorithme *ad hoc* qui permet de gérer les éventuels décalages du cadre de lecture introduits lors de l'analyse par PROTEA. La figure 4.2 présente un exemple de transcription inverse d'un alignement multiple de séquences d'acides aminés. Cet exemple est construit à partir de trois séquences de la famille PF07974 de PANDIT dont le pourcentage d'identité moyen est de 44,3%. Les séquences d'acides aminés prédites par PROTEA ont ici été alignées par CLUSTALW avant d'être rétro-transcrites. Sur cette figure sont également présentés trois alignements multiples produits par CLUSTALW [THG94], DIALIGN2-2 [Mor99] et MULTALIN [Cor88] à partir des séquences nucléiques d'origine. La qualité de chacun de ces alignements par rapport à l'alignement correct fourni dans PANDIT est évaluée grâce à deux mesures : la somme des scores deux à deux (SPS) et la somme des scores par colonne (CS). Soit un alignement multiple A de N séquences comportant M colonnes, on note A_{ij} la base ou le gap présent à la $i$ème colonne de la $j$ème séquence. On définit p_{ijk} tel que p_{ijk} vaut 1 si A_{ij} et A_{ik} sont alignés dans l'alignement de référence, 0 sinon. Les scores S_i et C_i de la $i$ème colonne de A selon l'alignement de référence sont alors donnés par

$$
\begin{aligned}
S_i &= \sum_{j=1, j\neq k}^{N} \sum_{k=1}^{N} p_{ijk} \\
C_i &= \begin{cases} 1 & \text{si } S_i = N(N-1) \\ 0 & \end{cases}
\end{aligned}
$$

Les valeurs du SPS et du CS de A se calculent alors de la manière suivante

$$
\begin{aligned}
SPS &= \frac{\sum_{i=1}^{M} S_i}{\sum_{i=1}^{Mr} S_{ri}} \\
CS &= \frac{\sum_{i=1}^{M} C_i}{M}
\end{aligned}
$$

où M_r est le nombre de colonnes de l'alignement de référence et S_{ri} est le score de la $i$ème colonne de l'alignement de référence. Le SPS et le CS prennent leurs valeurs dans l'intervalle $[0; 1]$, 1 étant la valeur maximale où l'alignement A correspond exactement à l'alignement de référence. Sur l'exemple de la figure 4.2, l'alignement par PROTEA +CLUSTALW est significativement plus proche de l'alignement de référence de PANDIT que les alignements générés par CLUSTALW, DIALIGN2-2 et MULTALIN.

```
O97702_CANFA      TGCAGCCCCCGGGAGGGCCAGCCCGCCTGCAGCCAGCGGGGCGAGTGCCTG------TGTGGCCAATGTGTCTGCCATAGCAGTGACTTTGGCAAGATCACGGGCAAGTACTGC
Q86G85_PSEIC      TGCCGGTCACCTGAAAACAACGAAATCTGCAGTGGAAACGGACAATGTGTA------TGTGGACAATGTATGTGTAACTCTGACGATGACCGCCACTATAGTGGCAAATACTGC
Q19267_CAEEL      TGTTTTGGAAAAGGATCC---------TGTCATGGAGATGGAAGCCGCGAAGGCAGT---GGAAAGTGTAAATGTGAGACTGGA------------TATACTGGAAATCTATGC
                  **          *   *         **          **     *            **  * ***   **  *                    *  ** **    ***
```

(a) Alignement de référence de PANDIT.

```
O97702_CANFA      TGCAGCCCCCGGGAGGGCCAGCCCGCCTGCAGCCAGCGGGGC---GAGTGCCTGTGTGGCCAATGTGTCTGCCATAGCAGTGACTTTGGCAAGATCACGGGCAAGTACTGC
Q86G85_PSEIC      TGCCGGTCACCTGAAAACAACGAAATCTGCAGTGGAAACGGA---CAATGTGTATGTGGACAATGTATGTGTAACTCTGACGATGACCGCCACTATAGTGGCAAATACTGC
Q19267_CAEEL      TGTTTT---------GGAAAAGGATCCTGTCATGGAGATGGAAGCCGCGAAGGCAGTGGAAAGTGTAAATGTGAGACTGGA------------TATACTGGAAATCTATGC
                  **                *      ***          **             ****  * ***   **  *                    *  ** **    ***
```

(b) Alignement de PROTEA. L'alignement au niveau peptidique a été confié à CLUSTALW. Le SPS de cet alignement vaut 0,83 et son CS 0,75.

```
O97702_CANFA      TGCAGCCCCCGGGAGGGCCAGCCCGCCTGCAG-CCAGCGGGGCGAGTGCCTGTGTGGCCAATGTGTCTGCCATAGCAGTGACTTTGGCAAGATCACGGGCAAGTACTGC
Q86G85_PSEIC      TGCCGGTCACCTGAAAACAACGAAATCTGCAG-TGGAAACGGACAATGTGTATGTGGACAATGTATGTGTAACTCTGACGATGACCGCCACTATAGTGGCAAATACTGC
Q19267_CAEEL      ----TGTTTTGGAAAAGGATCCTGTCATGGAGATGGAAGCCGCGAAGGCA---GTGGAAAGTGTAAATGTGAGACTGGATATACTGGAAATCTATGC------------
                            *            ** **         *  *  *     ****  * ***   **  *        *     *  *
```

(c) Alignement de CLUSTALW des séquences nucléiques. Le SPS de cet alignement vaut 0,524 et son CS 0,286.

```
O97702_CANFA      TGCAGCCCCCGGGAGGGCCAGCCCGCCTGCAGCCAGCGGGGCGAGTGCCTGTGTGGCCAATGTGTCTGCCATAGCAGTGACTTTGGCAAGATCACGGGCAAGTACTGC
Q86G85_PSEIC      TGCCGGTCACCTGAAAACAACGAAATCTGCAGTGGAAACGGACAATGTGTATGTGGACAATGTATGTGTAACTCTGACGATGACCGCCACTATAGTGGCAAATACTGC
Q19267_CAEEL      ----TGTTTTGGAAAAGGATCCTGTCATGGAGATGGAAGCCGCGAAGGCAGTGGAAAGTGTAAATGTGAGACTGGA--------------TATACTGGAAATCTATGC
                            *            *****             *    **       *   ****  *                       *  ** **    ***
```

(d) Alignement de MULTALIN des séquences nucléiques en utilisant des informations au niveau peptidique. Le SPS de cet alignement vaut 0,476 et son CS 0,210.

```
O97702_CANFA      TGCAGCCCCCGGGAGGGCCAGCCCGCCTGCAGCCAGCGGGGCGAGTGCCTGTGTGGCCAATGTGTCTGCCATAGCAGTGACTTTGGCAAG--------------------
Q86G85_PSEIC      TGCCGGTCACCTGAAAACAACGAAATCTGCAGTGGAAACGGACAATGTGTATGTGGACAATGTATGTGTAACTCTGACGATGACCGCCAC--------------------
Q19267_CAEEL      ----------------------------TGTTTTGGAAAAGGATCCTGTca------------------------TGGAGATGGAAGCCGcgaaggcagtggaaagtgtaa
                                           **          **    **                                **     **

O97702_CANFA      -------------ATCACGGGCAAGTACTGC
Q86G85_PSEIC      -------------TATAGTGGCAAATACTGC
Q19267_CAEEL      atgtgagactggaTATACTGGAAATCTATGC
                            *  ** **    ***
```

(e) Alignement de DIALIGN2-2 des séquences nucléiques en utilisant des informations au niveau peptidique. Le SPS de cet alignement vaut 0,476 et son CS 0,210.

FIG. 4.2 – Un exemple de trois séquences de la famille PF07974 de PANDIT alignées par PROTEA +CLUSTALW, CLUSTALW, DIALIGN2-2 et MULTALIN. Le pourcentage d'identité moyen de ces séquences est de 44,3%. Le pourcentage d'identité moyen au niveau peptidique est de 30,3%. Le SPS et le CS de chaque alignement est calculé en utilisant l'alignement fourni dans PANDIT comme référence.

Prédiction et alignement de séquences non codantes partageant une structure commune

A l'image de la démarche mise en œuvre à partir des prédictions de PROTEA pour aligner des séquences codantes homologues, la combinaison de CARNAC (section 3.3) et de GARDENIA [BT06] permet de produire un alignement de séquences partageant une structure putative. A partir des structures prédites par CARNAC, GARDENIA réalise un alignement multiple des séquences en utilisant à la fois les informations des structures primaires et secondaires. Les structures secondaires de chaque séquence sont représentées sous forme arc-annotée. L'alignement multiple d'un ensemble de séquences arc-annotées est alors une super-séquence commune incluse. Le schéma d'édition adopté intègre des opérations d'évolution entre les bases non appariées et entre les paires de bases appariées originellement définies dans [JLMZ02]. La construction de la super-séquence est un problème NP-dur. Dans GARDENIA, une approche heuristique est donc mise en œuvre. Dans un premier temps, GARDENIA procède au calcul de la super-séquence de chaque couple de séquences. Ensuite, l'incorporation des super-séquences se fait de manière progressive en utilisant un clustering hiérarchique ascendant en fonction du degré de similarité des couples de séquences. L'alignement des super-séquences est réalisé avec le même algorithme que pour la construction des super-séquences deux à deux. Enfin, l'espace de recherche de l'algorithme d'alignement est contraint à chaque étape par les points d'ancrage trouvés par CARNAC (section 3.3). Ces contraintes permettent d'accélérer de manière significative l'alignement. La prédiction de séquences structurées homologues est réalisée en fixant un seuil déterminé de manière empirique sur la valeur du score d'alignement calculé par GARDENIA.

Implantation de Magnolia

MAGNOLIA est développé sous forme d'un site Web qui fait appel aux différentes composantes et synthétise les résultats. Lorsqu'une prédiction "codant" est réalisée par PROTEA, deux alignements sont produits : l'alignement multiple des séquences d'acides aminées prédites, et la rétro-transcription de cet alignement. Plusieurs méthodes sont proposées à l'utilisateur pour l'alignement au niveau peptidique : CLUSTALW, DIALIGN2-2 et T-COFFEE. La mise en couleur des acides aminés du premier alignement et des codons correspondants dans le second alignement est inspirée des couleurs de RASMOL[2]. La figure 4.3 montre un exemple des alignements produits par MAGNOLIA pour une famille de séquences codantes homologues.

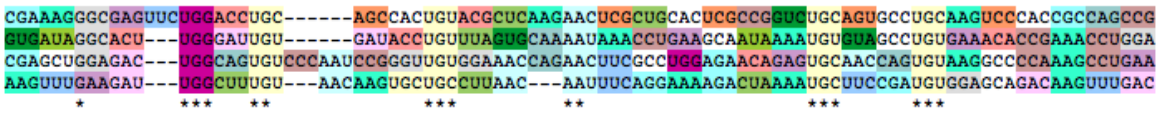


FIG. 4.3 – L'alignement par MAGNOLIA (PROTEA) du domaine Zn-finger des protéines Ran (PFAM PF00641). La longueur moyenne des séquences est de 92 nucléotides et leur pourcentage d'identité moyen de 45,1%. Les triplets de bases sont coloriés en fonction de l'acide aminé codé. L'alignement de référence est quasiment identique à celui fourni dans PANDIT.

Lorsqu'une prédiction "structuré" est produite par CARNAC+GARDENIA , un alignement multiple des séquences annoté par la structure est généré. Chaque tige prédite par CARNAC

[2] `http://www.rasmol.org`

est coloriée. La figure 4.4 montre un exemple d'alignement pour une famille de séquences non codantes homologues partageant une structure commune. Les structures secondaires de chaque séquences sont également représentées sous forme graphique à l'aide du programme NaView [BH88].

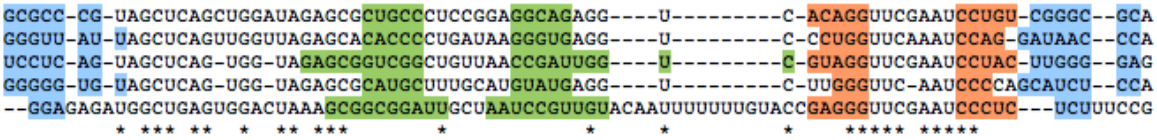


FIG. 4.4 – L'alignement par MAGNOLIA (caRNAc+GARDENIA) de cinq séquence d'ARN de transfert (RFAM RF00005). La longueur moyenne des séquences est de 76 nucléotides et leur pourcentage d'identité moyen de 51,0%. Trois des quatre tiges sont retrouvées et coloriées en bleu, en vert et en orange. Tous les appariements prédits sont corrects et l'alignement multiple est consistent avec l'alignement de référence de RFAM.

Dans MAGNOLIA, les prédictions de PROTEA et de caRNAc+GARDENIA sont considérées comme indépendantes. MAGNOLIA peut donc réaliser une double prédiction, "codant" et "structuré", pour un même ensemble de séquences, auquel cas les alignements de PROTEA et de GARDENIA sont produits. Certaines familles d'ARN non codants présentent en effet à la fois une structure commune conservée et un cadre de lecture conservé. C'est le cas notamment des tmRNA (RFAM RF00023), aussi connus sous les noms ARN 10Sa et SsrA, qui ont des propriétés d'ARN de transfert et d'ARN messager. Le rôle des tmRNA est de libérer un ribosome "bloqué" en cours de traduction. A cet effet, les tmRNA comportent un court cadre de lecture terminé par un codon STOP et comporte une région structurée imitant partiellement la structure d'un ARN de transfert. L'ARN messager sur lequel le ribosome est bloqué est remplacé par un tmRNA et la traduction du tmRNA conduit à l'ajout d'un signal de protéolyse au peptide incomplètement synthétisé afin qu'il soit dégradé. Les tmRNA existent dans tous les génomes bactériens séquencés, et ont récemment été identifié dans des chloroplastes, mais semblent absents chez les eucaryotes. Outre des ARN non codants, d'autres types de familles de séquences partageant une structure commune sont réputées pour contenir un cadre de lecture conservé ou à l'inverse apparaître dans des cadres de lecture. Par exemple, l'élément cis-régulateur des rhinovirus humains (RFAM RF00220) se situe dans le cadre de lecture codant pour la protéine *capside*, c'est-à-dire la protéine qui constitue l'enveloppe du virus (section 1.4.2), ou encore la tige boucle numéro VII du virus de l'hépatite C (RFAM RF00468) dans la région codante du gène NS5B.

MAGNOLIA est spécifiquement conçu pour aligner des séquences fonctionnelles divergentes. Sur des séquences trop proches, l'approche comparative des composantes sous-jacentes, PROTEA et caRNAc, n'est alors pas appropriée. De plus, des séquences proches peuvent être alignées avec des outils d'alignement classiques, qui ne présument d'aucune fonction commune. Par conséquent, lorsque le pourcentage d'identité moyen des séquences dépasse 90%, MAGNOLIA bascule sur une méthode d'alignement plus traditionnelle et propose donc automatiquement un alignement multiple réalisé par CLUSTALW, DIALIGN2-2 ou T-COFFEE selon le choix de l'utilisateur.

4.1.4 Les résultats expérimentaux de Magnolia

Pour évaluer les performances de MAGNOLIA nous avons sélectionné deux jeux de données : PANDIT [WdBQ+06] et BRALIBASE 2.1 [WMS06]. PANDIT est un ensemble de familles de séquences codantes déjà utilisé pour évaluer PROTEA (section 2.5). BRALIBASE 2.1 est un ensemble de familles d'ARN non codants construit dans le but d'évaluer les performances des méthodes d'alignement multiple d'ARN structurés. BRALIBASE 2.1 reprend les familles initialement proposées par Gardner dans BRALIBASE II [GWW05], le premier benchmark pour l'alignement structural, et étend ce jeu de données à plus d'une trentaine de familles d'ARN non-codants. De plus, BRALIBASE 2.1 propose des ensembles de séquences contenant deux, trois, cinq, sept, dix et quinze séquences, contrairement à BRALIBASE II qui ne propose que des ensembles de cinq séquences.

Les résultats de Magnolia sur les familles de séquences codantes de Pandit

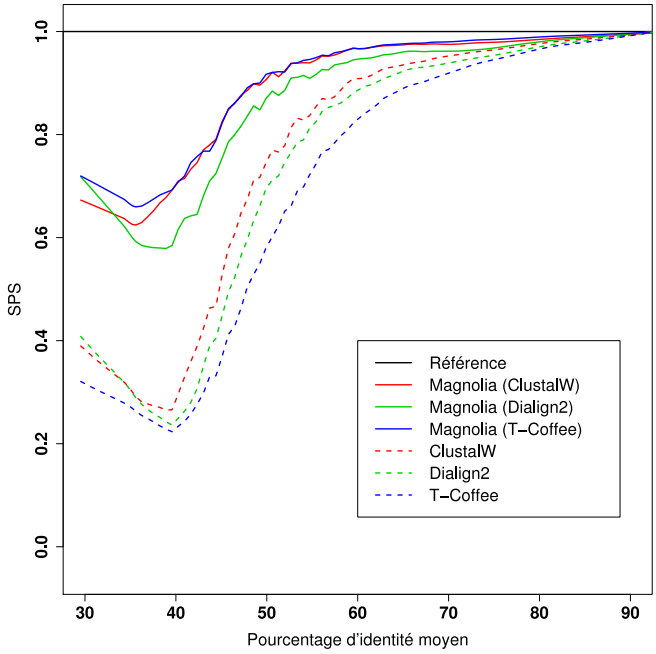


FIG. 4.5 – Comparaison des alignements de MAGNOLIA, CLUSTALW, DIALIGN2-2 et T-COFFEE sur les ensembles de quatre séquences extraites des familles de PANDIT. Le SPS est donné en fonction du pourcentage d'identité moyen des séquences nucléiques. Pour DIALIGN2-2 seul, l'option permettant d'effectuer les comparaisons au niveau peptidique a été activée.

Pour chaque famille de PANDIT, un sous-ensemble de quatre séquences ont été choisies aléatoirement. Sur les 6 491 ensembles ainsi construits, 6 122 (94,3%) sont correctement prédites "codant" par MAGNOLIA, et pour plus de 99% d'entre elles les cadres de lecture prédits sont corrects. Moins de 3% des familles sont prédites "structurés" par MAGNOLIA. Pour estimer la qualité des alignements multiples produits, nous utilisons le SPS décrit à la page 116. Comme MAGNOLIA s'appuie sur une méthode d'alignement multiple externe pour aligner les séquences d'acides aminés prédites, nous avons testé trois méthodes différentes : CLUSTALW, DIALIGN2-2 et T-COFFEE. Les alignements de MAGNOLIA sont comparés aux

alignements produits par ces mêmes méthodes utilisées sur les séquences nucléiques initiales. Les résultats sont présentés en figure 4.5 en fonction du pourcentage d'identité moyen des séquences nucléiques. Quelque soit le degré de conservation des séquences, les alignements de Magnolia sont plus proches des alignements de référence de Pandit que les autres méthodes d'alignement multiple testées. Plus les séquences sont divergentes, plus l'écart se creuse entre les alignements de Magnolia et les autres, quelque soit la méthode d'alignement sous-jacente utilisée.

Les résultats de Magnolia sur le benchmark BRAliBase 2.1

BRAliBase 2.1 contient des ensembles de séquences dont le pourcentage d'identité moyen varie d'environ 30% à 95%. Pour chaque ensemble de séquences, un alignement de référence extrait de la littérature est fourni. Pour nos tests, nous nous sommes focalisés sur les ensembles de séquences faiblement conservées avec un pourcentage d'identité moyen inférieur à 50%. Nos expériences ont donc portés sur 510 ensembles de cinq séquences, 318 de sept séquences, et 174 de dix séquences. Un peu moins de 20% des ensembles testés ont été incorrectement prédits "autre" par Magnolia, et 5% ont été classifiés "codant". Ce taux relativement élevé de prédictions "codant" s'explique par la faiblesse de l'analyse de Protea sur les séquences courtes abondantes et dont la longueur moyenne est ici inférieure à 80 nucléotides. Dans BRAliBase 2.1, deux mesures sont utilisées pour mesurer la qualité des alignements produits : le SPS, déjà utilisé dans la section précédent pour évaluer la qualité des alignements produits par Protea, et l'*index de conservation de structure* (SCI) utilisé dans RNAz (section 3.2.4). Le SCI est un indice qui mesure le degré de conservation en terme d'énergie d'une structure conservée par rapport aux structures optimales individuelles. Les résultats obtenus par Magnolia comparés à ceux des méthodes d'alignement traditionnelles sont présentées sur la figure 4.6 en fonction du pourcentage d'identité moyen et du nombre de séquences. Ces résultats montrent que les alignements produits par Magnolia sont plus proches des alignements de référence que ceux générés par les approches traditionnelles quelque soit le nombre de séquences alignées. Sur la figure 4.7, les résultats de Magnolia sont comparés à ceux produits par les autres méthodes d'alignement structural. Ces graphiques font apparaître que les performances de Magnolia se situent dans la moyenne des autres méthodes quelque soit le nombre de séquences utilisées. Toutefois, les méthodes qui produisent de meilleurs alignements que Magnolia sont aussi beaucoup plus lentes. Ces tests, réalisés sur une machine de bureau classique, ont pris moins d'une demi heure pour Magnolia contre plus de quatre heures pour Mlocarna, Lara et MxscaRNA.

4.2 L'annotation par génomique comparative

Précédemment, nous avons vu comment utiliser les prédictions de caRNAc et de Protea pour produire des alignements multiples. Nous nous intéressons maintenant à la mise en œuvre d'une plate-forme d'annotation par génomique comparative qui combine plusieurs logiciels développés dans l'équipe : caRNAc, Protea et Yass [NK05], un logiciel d'alignement local de séquences. Cette plate-forme se présente sous la forme d'un pipeline logiciel modulaire.

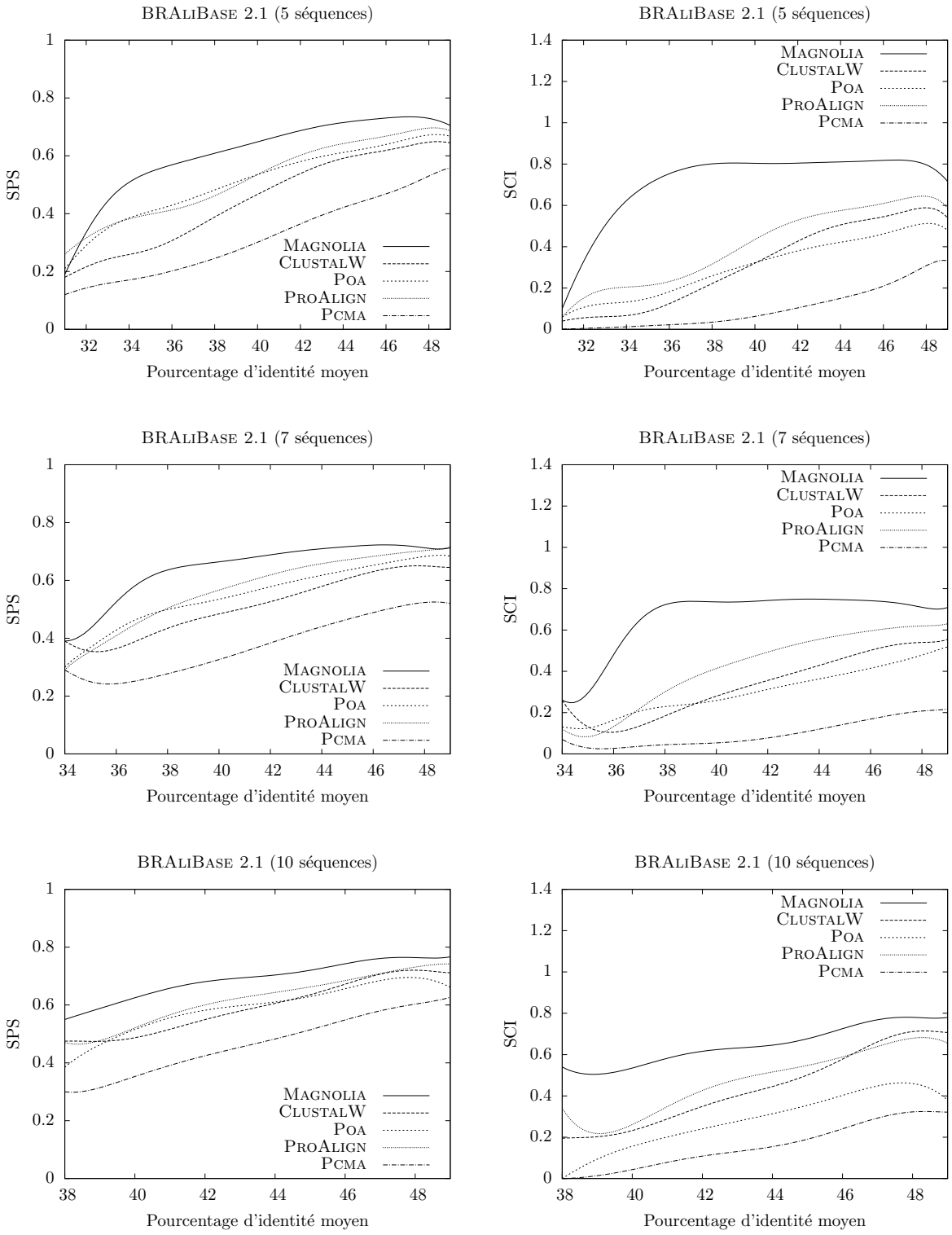


FIG. 4.6 – Résultats de MAGNOLIA sur BRALIBASE 2.1 comparés aux résultats des méthodes d'alignement multiple traditionnelles. Le SPS et le SCI des alignements sont présentés en fonction du pourcentage d'identité moyen des séquences alignées et du nombre de séquences utilisées.

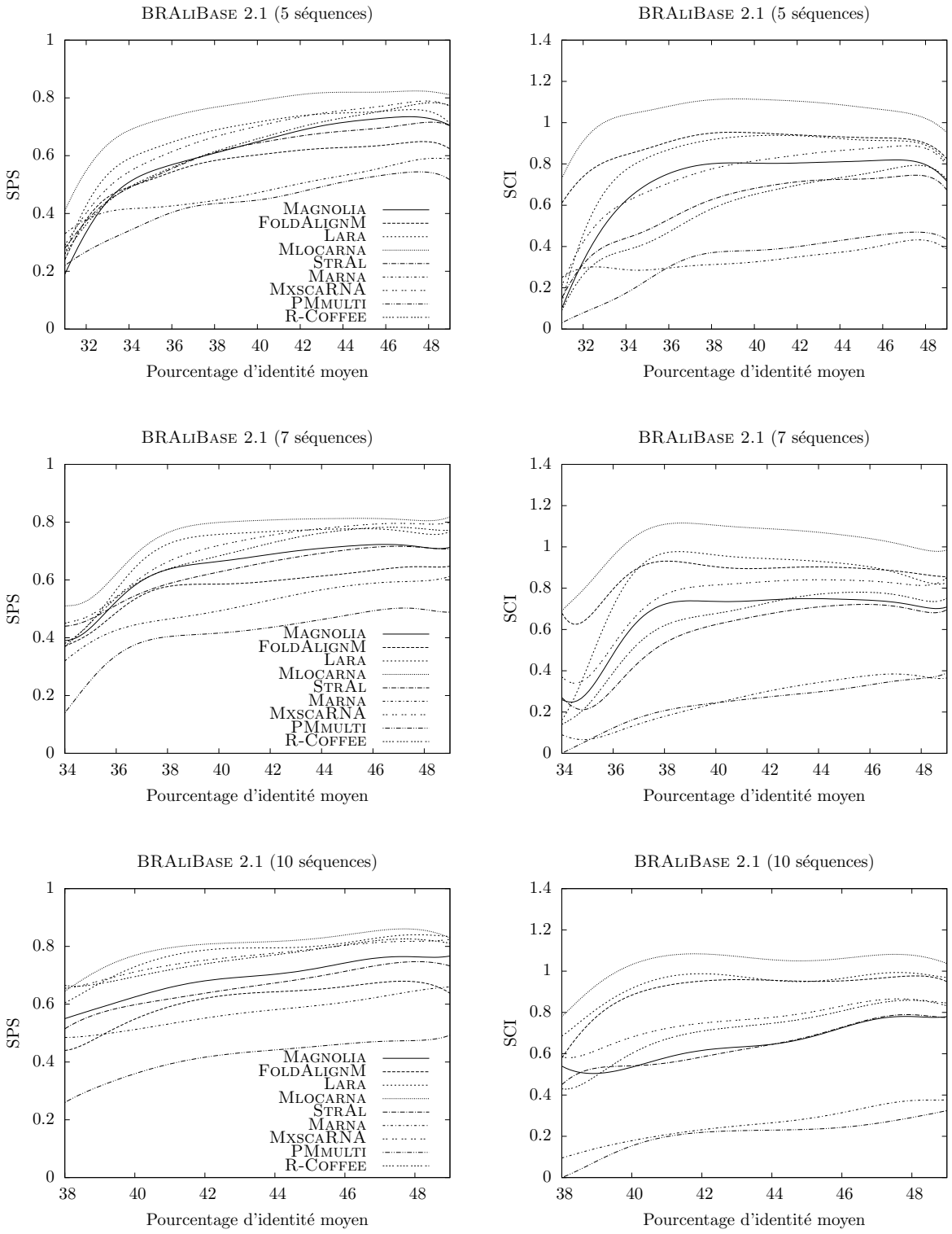


FIG. 4.7 – Résultats de MAGNOLIA sur BRALIBASE 2.1 comparés aux résultats des méthodes qui construisent un alignement multiple structural. Le SPS et le SCI des alignements sont présentés en fonction du pourcentage d'identité moyen des séquences alignées et du nombre de séquences utilisées.

4.2.1 Le pipeline d'annotation

Globalement, le pipeline accepte en entrée une séquence génomique à annoter et une banque de séquences susceptible de contenir des régions homologues avec la séquence à annoter, et produit en sortie des annotations de séquences codantes et d'ARN non-codants hypothétiques. Le fonctionnement global du pipeline est schématisé sur la figure 4.8. Dans un premier temps, la séquence à annoter est comparée à celles présentes dans la banque de données par une méthode d'alignement deux à deux. Ensuite, la totalité des alignements obtenus sont reportés sur la séquence à annoter afin de détecter des régions conservées, c'est-à-dire des régions de la séquence à annoter pour laquelle il existe plusieurs séquences similaires dans la banque. Pour chaque région conservée, les séquences similaires trouvées sont soumises à un ultime traitement afin d'extraire un sous-ensemble pertinent pour un traitement par analyse comparative.

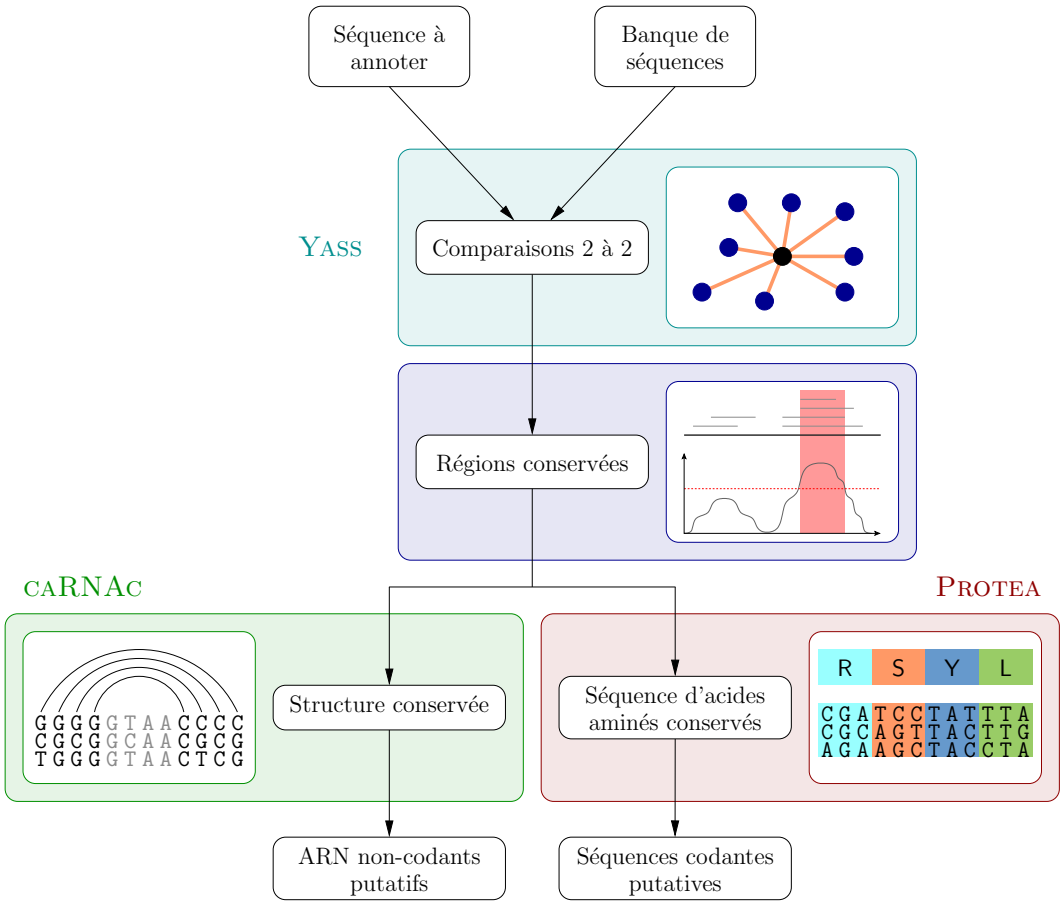


FIG. 4.8 – Pipeline d'annotation automatique de séquences codantes par PROTEA et d'ARN non-codants par caRNAc dans une séquence génomique.

Le choix des génomes à comparer est cruciale car la qualité des prédictions qui peuvent être réalisées dépend pleinement de ces séquences. De manière générale, le facteur déterminant est les distances évolutives qui séparent les organismes dont elles sont issues de l'organisme dont provient la séquence à annoter. Prenons l'exemple du génome d'*Escherichia coli* dans lequel on cherche à identifier de nouvelles séquences codantes ou d'ARN non-codants. La séquence à

annoter sera alors la séquence génomique d'une souche d'*Escherichia coli*, et la banque pourra alors être constituée de séquences génomiques d'autres bactéries plus ou moins éloignées en terme d'évolution. Avec des séquences génomiques très proches de celle de l'organisme ciblé, le risque majeur est que les séquences homologues exhibent trop peu de mutations pour qu'elles puissent à elles seules faire l'objet d'une analyse comparative pertinente. A l'inverse, si l'on choisit des séquences d'organismes trop éloignés, on risque de ne pas être en mesure d'identifier les séquences homologues trop peu conservées, ou pire, que les organismes choisis ne possèdent pas d'homologue pour une séquence fonctionnelle putative du génome à annoter. Bien qu'il n'existe pas de critère absolu pour déterminer les séquences génomiques à choisir, certaines situations font naturellement émerger des contraintes. Par exemple, si on cherche à identifier une séquence liée à un phénotype particulier tel que la production d'un agent pathogène, il apparaît alors nécessaire de considérer d'autres souches du même organisme qui partagent ce même phénotype.

Dans le cas où l'on souhaite découvrir de nouvelles séquences codantes ou d'ARN non-codants, il apparaît naturel de vouloir *masquer* les régions qui comportent déjà des annotations de la séquence à annoter, ou les régions susceptibles de parasiter les comparaisons telles que des régions hautement conservées (plus de 95%) ou des éléments répétés. Le pipeline offre ainsi la possibilité de masquer automatiquement des régions à ignorer durant la comparaison avec la banque. Le masquage des régions déjà annotées permet de diminuer de manière radicale le nombre de régions à comparer et par conséquent le nombre de prédictions à traiter en sortie du pipeline. Toutefois, cette mesure drastique peut également conduire à manquer des séquences codantes ou d'ARN non-codants inconnues qui chevaucheraient ou seraient incluses dans des régions déjà annotées pour une autre fonction.

Comparaison de séquences génomiques

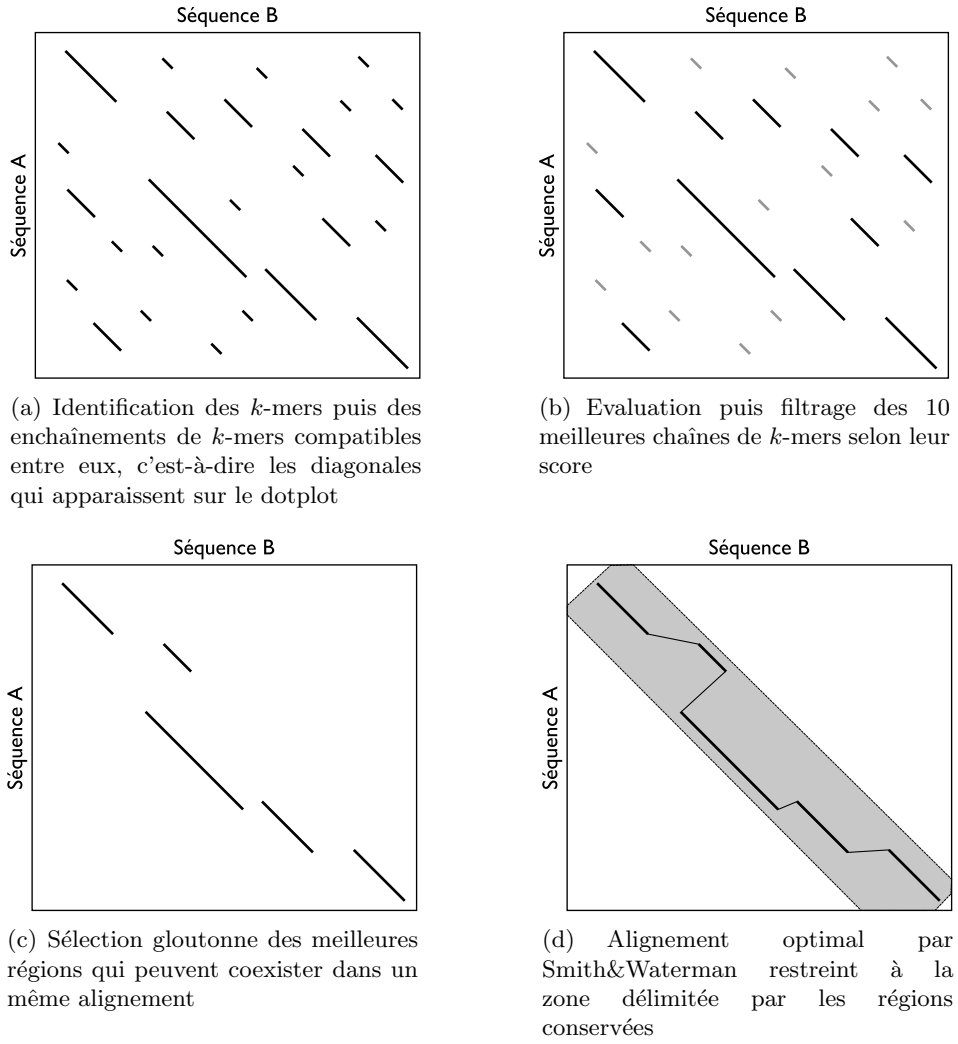
La comparaison de la séquence à annoter avec les séquences présentes dans la banque est la première étape du pipeline. Au cours des chapitres précédents, nous avons déjà fait mention de plusieurs algorithmes permettant de fouiller une banque de séquences à la recherche de séquences similaires tels que l'algorithme de Smith&Waterman, Blast et Fasta. Nous allons nous intéresser plus en détails à ces outils, l'objectif étant de pouvoir apprécier leurs différences par rapport à l'outil que nous utilisons dans le pipeline : Yass.

L'algorithme de Smith&Waterman construit par programmation dynamique l'alignement local optimal de deux séquences. L'équation de récurrence servant à remplir la matrice M de programmation dynamique est

$$M[i][j] = \max \begin{cases} 0 \\ M[i-1][j-1] + s(A_i, B_j) \\ M[i-1][j] + s(A_i, -) \\ M[i][j-1] + s(-, B_j) \end{cases}$$

où $s(A_i, B_j)$ est le coût de substitution du $i$ème nucléotide de la séquence A par le $j$ème nucléotide de la séquence B, $s(A_i, -)$ est le coût d'insertion du $i^{\text{ème}}$ nucléotide de la séquence A dans la séquence B et $s(-, B_j)$ le coût l'insertion du $j^{\text{ème}}$ nucléotide de B dans A. La remontée dans cette matrice à partir de la valeur maximale quelle contient permet d'obtenir l'alignement local optimal des séquences A et B. La complexité en $\mathcal{O}(n^2)$ en espace et en temps de cet algorithme est inadaptée à l'exploration complète d'une banque de données, surtout si celle-ci contient de très longues séquences telles que des génomes entiers ou des

chromosomes eucaryotes. Néanmoins, il existe plusieurs heuristiques pour l'alignement local qui permettent de traiter des banques de données de manière efficace, sans perte significative de sensibilité. Ces heuristiques reposent toutes sur le même principe : les *graines*, c'est-à-dire des séquences ou des sous-séquences d'une longueur donnée parfaitement conservées entre deux séquences. Dans un premier temps, ces heuristiques procèdent à la recherche de ces graines puis les étendent pour former un alignement local complet. Les différences entre ces heuristiques apparaissent à deux niveaux : le type de graine utilisé et la manière de reconstruire un alignement à partir de ces graines. Nous allons aux heuristiques à base de graines contiguës, FASTA et BLAST, puis aux heuristiques à base de graines espacées, PATTERNHUNTER et YASS.



(a) Identification des k-mers puis des enchaînements de k-mers compatibles entre eux, c'est-à-dire les diagonales qui apparaissent sur le dotplot

(b) Evaluation puis filtrage des 10 meilleures chaînes de k-mers selon leur score

(c) Sélection gloutonne des meilleures régions qui peuvent coexister dans un même alignement

(d) Alignement optimal par Smith&Waterman restreint à la zone délimitée par les régions conservées

FIG. 4.9 – Schéma des étapes principales de FASTA pour deux séquences A et B.

Chronologiquement, FASTA [LP85] est la première heuristique d'alignement local qui utilise le principe des graines contiguës. Une graine contiguë est une séquence d'une certaine longueur k, un k-mer, parfaitement conservée entre deux séquences. La valeur de k constitue un paramètre crucial de la méthode qui affecte à la fois sa sensibilité et son efficacité. Plus k est grand, plus la méthode est rapide au détriment de sa sensibilité. L'heuristique de

FASTA se décompose en quatre grandes étapes illustrées par les schémas de la figure 4.9. La première étape consiste à rechercher à l'aide d'une graine contiguë les k-mers conservés. FASTA cherche ensuite à créer des *régions locales similaires*, c'est-à-dire à regrouper des k-mers afin de détecter des séquences conservées plus longues, pouvant contenir des substitutions mais n'introduisant aucun gap. Pour chaque région, FASTA calcule ensuite un score puis filtre les meilleures régions afin de ne garder que les dix meilleures. Le calcul de ce score fait intervenir une matrice de substitution triviale, la matrice identité. Les régions trouvées sont ensuite incorporées de manière gloutonne par score décroissant afin de ne garder que les régions compatibles entre elles, c'est-à-dire des régions qui peuvent faire partie d'un même alignement. L'alignement local est enfin produit par l'algorithme de Smith&Waterman où la matrice de programmation dynamique est réduite à la zone qui englobe les régions retenues à l'étape précédente.

L'heuristique de BLAST [AGM+90] est globalement identique à celle de FASTA. La différence majeure entre BLAST et FASTA se situe au niveau du passage des k-mers à un alignement. Pour chaque k-mer trouvé, BLAST procède à son extension de part et d'autre de façon à trouver une région conservée la plus longue possible. BLAST continue d'étendre la région tant que le score cumulé calculé au fur et à mesure de l'extension ne descend pas en dessous d'un certain seuil. Contrairement à FASTA, BLAST applique des coûts variables aux substitutions : 5 pour un match et -4 pour un mismatch. Les régions ainsi obtenues, appelées des HSP (High-scoring Segment Pairs), sont ensuite filtrées en fonction de l'espérance statistique de leurs scores. Dans la version "avec gap" de BLAST les k-mers trouvés peuvent être groupés pour former une HSP avant leur extension si la distance qui les sépare ne dépasse pas un certain seuil.

PATTERNHUNTER [MTL02] et YASS [NK05] sont des heuristiques qui fonctionnent selon le même schéma global que BLAST. Leur originalité tient à l'utilisation des graines espacées, c'est-à-dire des sous-séquences conservées. Les graines espacées apportent une meilleure sensibilité que les graines contiguës, permettant ainsi de trouver des régions moins conservées, sans pour autant dégrader ni la spécificité ni l'efficacité apportées par les graines contiguës. Le graphique de la figure 4.10 montre le gain de sensibilité théorique apporté par l'utilisation d'une graine espacée par rapport à une graine contiguë de même contenu informationnel, c'est-à-dire où le nombre de nucléotides comparés est identique et appelé *poids* d'une graine. La figure 4.11 montre un exemple d'alignement qui ne pourrait être obtenu avec une graine contiguë de même poids que la graine espacée utilisée. En effet, bien que ces séquences soient similaires, celles-ci ne contiennent pas de mot de longueur 6 qui soit parfaitement conservé.

Quelque soit la méthode d'alignement utilisée, il est nécessaire d'évaluer la significativité des alignements trouvés. Le score d'un alignement n'est en soi pas un critère de décision pour plusieurs raisons : il dépend du système de score utilisé mais également de la longueur des séquences comparées. Pour évaluer la significativité d'un alignement, il est donc nécessaire d'évaluer sa probabilité d'occurrence afin de répondre à la question suivante : quelle était la probabilité de trouver cet alignement "par hasard" en comparant des séquences ne possédant aucune homologie *a priori* ? Cette question suppose que les séquences ont été générées selon un modèle aléatoire. Les travaux de Karlin et Altschul [KO87, KO88, KA90, KB92] ont permis d'établir le modèle actuellement utilisé par toutes les méthodes et selon lequel, pour un système de score fixé, la distribution des scores d'alignements locaux suit une loi de Gumbel [Gum58]. Dès lors, on est en mesure d'évaluer la significativité d'un alignement de score s en calculant sa E-valeur, c'est-à-dire le nombre d'alignements de score supérieur à s attendus

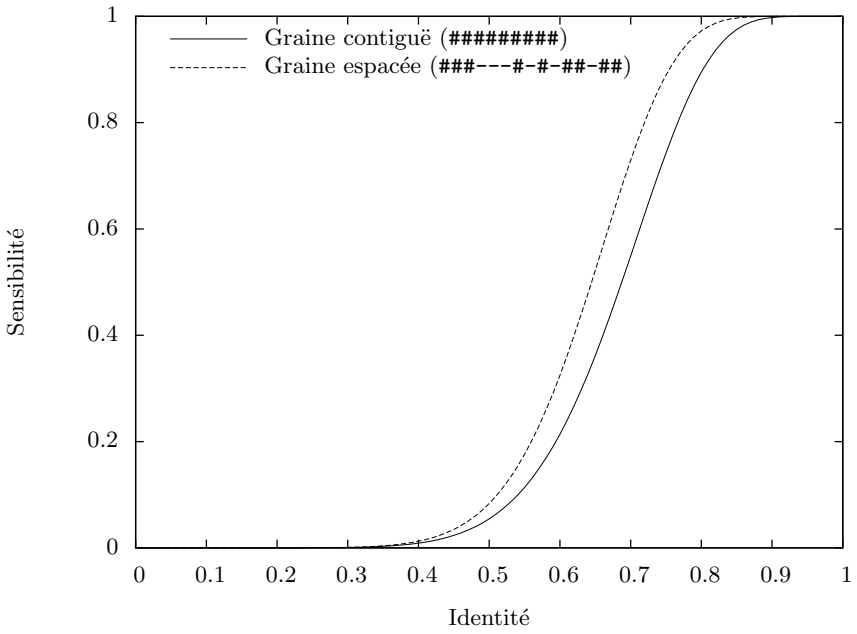


FIG. 4.10 – Comparaison de la sensibilité théorique des graines contiguës et espacées.

```
                    GACTGAACTCAT   TAGACTCGACGA
                    |.||.||||.||   |||.||.|||.|
                    GGCTAAACTAAT   TAGGCTAGACTA
Graine contiguë          ####
 Graine espacée       ##-##           ##-##
```

FIG. 4.11 – Les deux alignements présentent une identité de $9/12 = 75\%$. On considère deux graines de poids 4 : la graine contiguë `####` et la graine espacée `##-##`. La graine contiguë ne détecte que le premier alignement, alors que la graine espacée détecte les deux. Le symbole `#` correspond à une position d'identité et le symbole `-` à une position quelconque.

par hasard lorsque l'on aligne une séquence de longueur m avec une séquence de longueur n, donnée par :

$$\text{E-valeur}(s, m, n) = K.m.n.\exp(-\lambda.s)$$

où K et λ sont des paramètres de la loi de distribution qui proviennent du système de score choisi et de la composition en mono-nucléotides des séquences, m et n sont les longueurs des deux séquences comparées. Dans notre cas où l'on compare une séquence requête contre une banque de données, m est la longueur de la séquence requête et n est la somme des longueurs de toutes les séquences de la banque. Une fois calculée, l'interprétation d'une E-valeur est assez simple : plus la E-valeur associée à un alignement est proche de 0, plus la similarité obtenue est significative.

Afin d'évaluer le gain pratique que peuvent apporter les heuristiques à base de graines espacées, nous avons cherché à comparer systématiquement les résultats de Blast et de Yass. Nous nous sommes restreints à ces deux logiciels pour plusieurs raisons. Blast reporte toutes les similitudes locales détectées sous forme de plusieurs alignements, contrairement à Fasta qui n'en sélectionne qu'une partie pour former un seul et même alignement "global". Fasta peu donc être amené à écarter certaines séquences localement similaires qu'il n'arrive pas à regrouper pour former son alignement, telles que des séquences distantes, répétées, permutées ou inversées. Des différences minimes séparent Yass et PatternHunter. Pour comparer les performances pratiques en terme de sensibilité des approches à base de graines contiguës et espacées, nous avons choisi de travailler sur la comparaison de séquences d'ARN non-codants car leurs séquences tendent à être moins bien conservées que celles des régions codantes. A cet effet, nous avons comparé les performances de Blast et de Yass sur deux jeux de données : les 574 familles d'ARN non-codants de Rfam, et les trois familles d'ARN non-codants de BRAliBase III (page 81). Le protocole expérimental est identique à celui de BRAliBase III : chaque séquence de chaque famille est utilisée comme séquence "requête" pour retrouver ses homologues, soit dans Rfam, soit dans BRAliBase III selon le jeu de données dont elle est issue. Le compromis sensibilité/spécificité de Yass et de Blast sur Rfam est présenté en figure 4.12, toutes familles confondues. Ce compromis est calculé en faisant varier le seuil sur la E-valeur des alignements produits par les deux logiciels. Cette expérience fait clairement apparaître que Yass est plus sensible que Blast à spécificité équivalente. La table 4.1 présente la sensibilité de Blast et de Yass sur chaque famille de BRAliBase III obtenue en fixant à seuil à 10^{-4} sur la E-valeur des alignements produits. Cette seconde expérience confirme les résultats précédents. A spécificité équivalente, Yass détecte quatre fois plus d'ARN de transfert que Blast, presque deux fois plus d'ARN ribosomiques 5S, et 8% d'ARN U5 supplémentaires. En terme de temps d'exécution, Blast et Yass sont équivalents. Cependant, Blast est entre 5 et 10% plus rapide que Yass dans ces expériences simplement car Yass produit plus d'alignements que Blast. Les résultats obtenus par Yass dans ces expériences nous ont conduit à le choisir pour la première étape du pipeline.

Détection de régions conservées

La production des alignements deux à deux par Yass entre la séquence à annoter et les séquences présentes dans la banque constitue la première étape du pipeline. La seconde étape du pipeline consiste à former les groupes de séquences similaires qui serviront par la suite à effectuer les prédictions par analyse comparative sur la séquence à annoter. Chaque groupe de

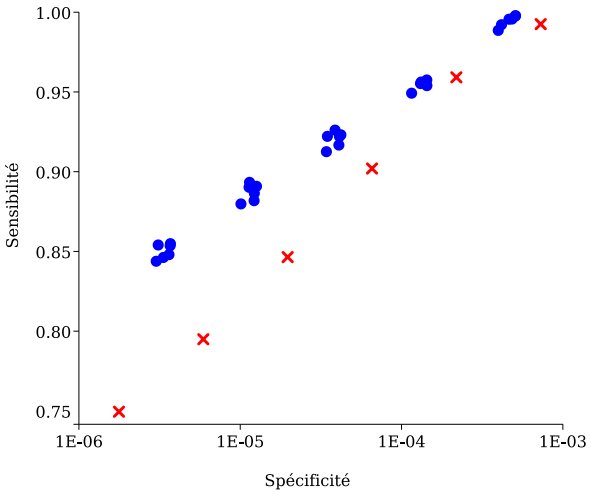


FIG. 4.12 – Compromis sensibilité/spécificité de BLAST (croix) et de YASS (ronds) sur RFAM, toutes familles confondues, avec des graines de poids 9. Chaque rond correspond à une exécution de YASS avec une graine espacée différente de même poids.

Méthode	ARN de transfert	Petits ARN U5	ARN ribo. 5S
BLAST	0,04	0,85	0,32
YASS	0,18	0,93	0,59

TAB. 4.1 – Sensibilité de BLAST et de YASS sur les familles d'ARN non-codants de BRALIBASE III avec des graines de poids 9 . Le seuil sur la E-valeur est ici fixé à 10^{-4}.

séquences doit donc comporter un fragment de la séquence à annoter et un nombre suffisant de séquences similaires. Les alignements deux à deux produits à l'étape précédente impliquent chacun un fragment de la séquence à annoter et une fragment de séquence provenant de la banque de données. La constitution des groupes de séquences à partir de ces alignements revient à identifier sur la séquence à annoter des régions particulièrement conservées, c'est-à-dire des régions pour lesquelles un nombre significatif de séquences similaires ont été trouvées. L'identification de ces régions et la génération des groupes de séquences similaires est réalisée par un logiciel développé dans l'équipe selon un protocole en trois temps. Premièrement, l'ensemble des alignements deux à deux produits à l'étape précédente sont repositionnés sur la séquence à annoter, comme illustré sur la figure 4.13. Ensuite, on calcule pour chaque base de cette séquence son *score cumulé de densité* grâce auquel on délimite les régions les plus conservées. Enfin, on extrait les séquences impliquées dans les régions conservées pour former des groupes de séquences pertinents pour la suite du pipeline.

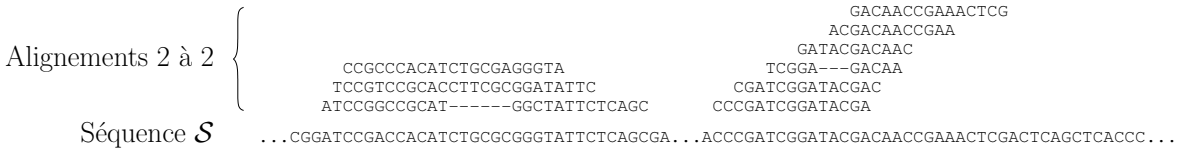


FIG. 4.13 – Repositionnement des alignements deux à deux sur la séquence à annoter.

Chaque position i de la séquence $\mathcal{S}$ à annoter est caractérisée par une *densité en alignements*, notée d_i, donné par la formule suivante

$$d_i = \frac{n_i}{A}$$

où n_i est le nombre d'alignements présents à la position i de la séquence $\mathcal{S}$ et A est le nombre moyen d'alignements par base de $\mathcal{S}$. La valeur de d_i représente la quantité d'alignements qui impliquent la base en position i de la séquence $\mathcal{S}$ par rapport au nombre d'alignements moyens par base pour la séquence $\mathcal{S}$. Par nature, la densité en alignements d'une position ne dépend pas de celle des positions voisines. L'interprétation de sa valeur dépend donc du contexte dans lequel on l'observe et ne représente donc pas un indicateur suffisant pour repérer des régions denses en alignements. Afin de simplifier l'identification des régions conservées, on calcule pour chaque position i de $\mathcal{S}$ son *score cumulé de densité* D_i selon la relation suivante

$$D_i = \max \begin{cases} 0 \\ D_{i-1} + \log(\frac{d_i}{\lambda}) \end{cases}$$

où λ est un paramètre strictement supérieur à 1 qui permet d'ajuster le seuil de détection des régions conservées. La valeur de D_i inclus non seulement la densité en alignements de la position i mais également le score D_{i-1} de la position précédente. Ce dernier élément est très important car il introduit une certaine inertie dans les variations du score qui va nous permettre de détecter simplement les régions denses en alignements. La figure 4.14 illustre la variation du score cumulé sur un exemple fictif. Cet exemple fait apparaître trois pics du score

cumulé qui correspondent chacun à une région particulièrement dense en alignements. On définit une région conservée par un intervalle de positions $[i; j]$ sur la séquence $\mathcal{S}$. Cet intervalle est déterminé de la manière suivante : D_j est un maximum global du score cumulé et i est la plus grande valeur inférieure à j telle que $D_{i-1} = 0$. Sur l'exemple de la figure 4.14, les trois zones grisées correspondent aux trois régions conservées identifiées définies par les intervalles de positions $[11; 26]$, $[50; 83]$ et $[122; 130]$ sur la séquence $\mathcal{S}$. Pour cet exemple, le paramètre λ est fixé à $1,3$. Dans les faits, ce paramètre permet de jouer sur la sensibilité de détection des régions conservées en modifiant l'amplitude de la valeur du score cumulé de densité. Lorsque λ augmente les valeurs du score diminuent, et inversement. Cependant, comme la valeur du score est majorée par zéro, plus λ augmente moins on observe de maximums, et inversement. Augmenter la valeur de λ permet donc de diminuer le nombre de régions conservées détectées ; diminuer sa valeur permet au contraire d'augmenter le nombre de régions détectées.

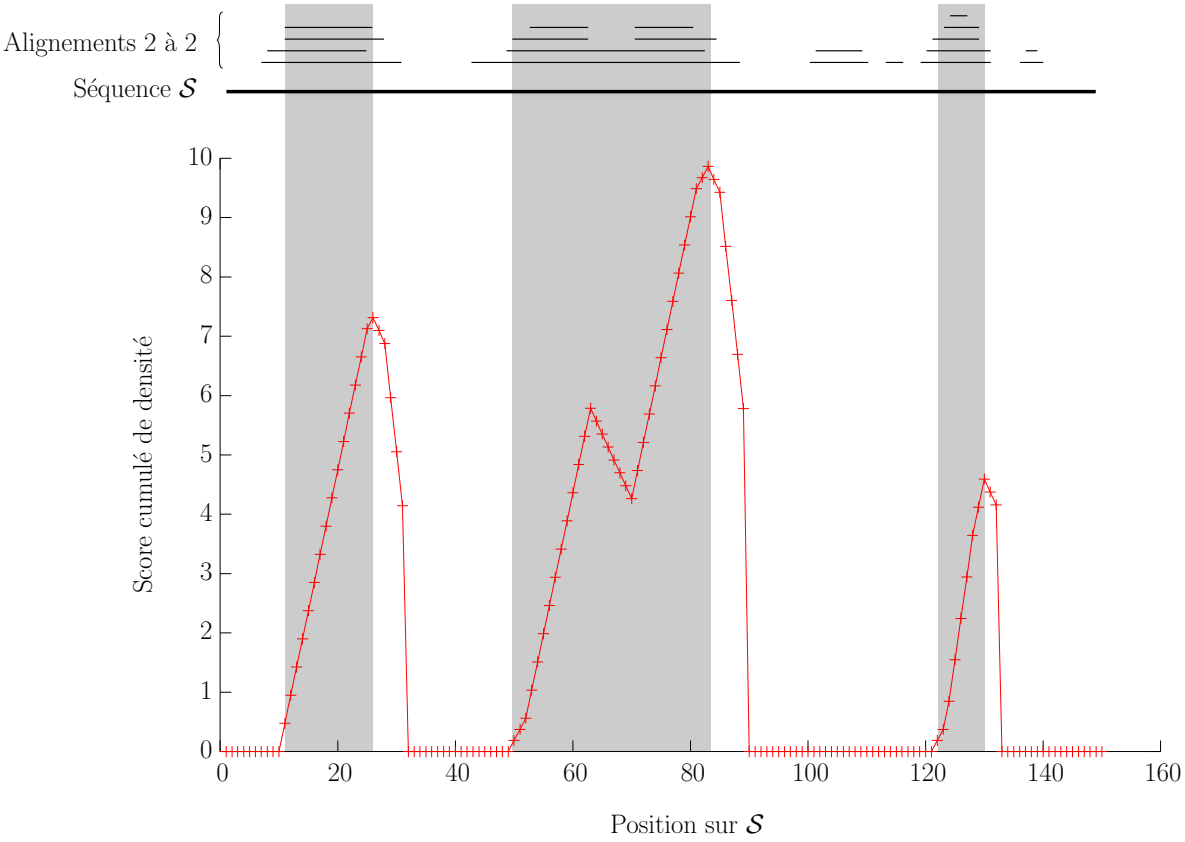


FIG. 4.14 – Variation du score cumulé de densité le long de la séquence à annoter $\mathcal{S}$. Le paramètre λ est ici fixé à $1,3$.

Lorsque les intervalles des régions conservées sont déterminés, les séquences correspondantes des alignements qui intersectent ces intervalles sont extraites. Il arrive que les alignements soient incomplets, notamment lorsque les génomes contiennent des séquences similaires faiblement conservées. Sur l'exemple de la figure 4.14, la seconde région identifiée correspond

à une accumulation de plusieurs alignements dont la majorité ne couvrent que partiellement le fragment de $\mathcal{S}$ en cause. Pour régler ce problème *a posteriori*, on étend de part et d'autre chaque séquence de la région afin d'obtenir une séquence de longueur identique au fragment de $\mathcal{S}$. C'est étape n'est possible que si l'on dispose dans la banque de données du contexte des séquences à étendre.

A ce niveau du pipeline, plusieurs traitements peuvent être appliqués aux groupes de séquences détectés. Ces traitements dépendent essentiellement des méthodes de prédictions auxquelles on souhaite les soumettre. Si l'on souhaite les soumettre à PROTEA et/ou à CARNAC pour prédire des séquences codantes ou d'ARN non-codants homologues, aucun traitement particulier supplémentaire n'est requis. Ces deux logiciels travaillent en effet sur des séquences non alignées et intègrent un pré-traitement pour éliminer les séquences redondantes. Si toutefois l'on souhaite utiliser d'autres méthodes basées sur une analyse comparative, il convient de vérifier certaines propriétés sur les groupes de séquences. Notamment pour les méthodes qui s'appuient sur un alignement, il est nécessaire de s'assurer qu'il est possible de construire un alignement fiable. Dans cette optique, on propose d'épurer chaque groupe de séquences en éliminant les séquences trop divergentes selon un procédé strictement analogue à celui mis en œuvre dans PROTEA et CARNAC pour construire les méta-séquences (section 1.5.3). On propose également de filtrer les groupes de séquences en fonction du nombre de séquences qu'ils contiennent.

4.2.2 Résultats expérimentaux du pipeline

Afin de tester le pipeline, nous sommes partis de l'annotation automatique d'ARN non-codants par analyse comparative conduite par Eddy *et al* [RKJE01] visant à découvrir de nouveaux gènes à ARN dans le génome d'*Escherichia coli*. Dans un premier temps, nous avons repris les données utilisées par Eddy *et al* et adapté l'expérience pour notre pipeline d'annotation. Par la suite, nous avons complété cette expérience en utilisant des séquences provenant d'organismes plus éloignés d'*Escherichia coli* en terme d'évolution que les organismes initiaux.

L'annotation d'ARN non-codants avec Qrna

L'expérience d'Eddy *et al* porte sur la découverte de nouveaux gènes à ARN dans la séquence génomique d'*Escherichia coli* prédits par analyse comparative puis vérifiés expérimentalement. La prédiction d'ARN non-codants est confiée à QRNA (section 3.2.4) sur des alignements obtenus par comparaison des régions inter-géniques du génome d'*Escherichia coli* et de génomes complets de quatre autres organismes.

La séquence génomique d'*Escherichia coli* utilisée est celle d'*Escherichia coli* K12 (MG1655). *Escherichia coli* K12 est une bactérie "modèle" qui se cultive et se manipule facilement en laboratoire. Très étudiée depuis plus de soixante dix ans, son génome fait partie des génomes les mieux annotés en terme de quantité et de qualité d'annotations. Comme la plupart des bactéries, son génome est plutôt court et compact. Composé de moins de cinq millions de bases, il comporte peu de régions *a priori* non fonctionnelles vierges de toute annotation. Toutes ces caractéristiques font d'*Escherichia coli* K12 un sujet idéal pour cette expérience. La quantité de régions inter-géniques à traiter est faible, et par conséquent le nombre de gènes à ARN candidats à vérifier par la suite aussi.

Les régions inter-géniques du génome d'*Escherichia coli* sont déterminées à partir des 115

gènes à ARN et des 4 290 gènes codants annotés. Seules les régions dont la longueur dépasse cinquante nucléotides sont retenues, ce qui représente au total 2 367 séquences couvrant 500 kilobases. La longueur moyenne de ces séquences est de 211 nucléotides, et la séquence la plus longue fait 1 729 nucléotides. Quatre gènes à ARN n'ont volontairement pas été exclus pour fournir un contrôle positif.

Les régions inter-géniques ainsi déterminées sont comparées aux séquences génomiques complètes de quatre organismes proches d'*Escherichia coli* en terme d'évolution :

- *Klebsiella pneumoniae*, souche 342 ;
- *Salmonella enterica enteridis*, souche PT4 ;
- *Salmonella enterica serovar Paratyphi A*, souche AKU_12601 ;
- *Salmonella enterica serovar Typhi*, souche CT18.

Les comparaisons sont réalisées par Blast et les alignements obtenus filtrés pour répondre à trois critères : une E-valeur inférieure à 0,01, une longueur supérieure à cinquante nucléotides, et un pourcentage d'identité supérieur à 65%. Ces critères sont fixés pour fournir des alignements pertinents à Qrna. Au total, 23 674 alignements sont produits par Blast, dont plus de la moitié proviennent de *Salmonella enterica serovar Typhi*.

Pour optimiser le traitement par Qrna, les alignements dont la longueur dépasse deux cents nucléotides sont découpés en fragments de deux cents nucléotides qui se chevauchent de cinquante nucléotides. Parmi ces alignements, Qrna prédit 556 couples de séquences d'ARN non-codants homologues. Ces 556 candidats contiennent les quatre gènes à ARN laissés volontairement. Parmi ces candidats, 281 correspondent à des éléments reconnus *a posteriori* comme n'étant pas des ARN non-codants mais des éléments possédant une structure secondaire caractéristique tels que des terminateurs de gènes, des séquences répétées et des séquences régulatrices. Parmi les 275 candidats restants, 49 ont été choisis manuellement en fonction de l'aspect général de la structure secondaire prédite et de la proximité avec des gènes connus par ailleurs. Après vérification expérimentale par *Nothern Blot*, il apparaît que 11 des 49 candidats retenus sont effectivement transcrits en ARN de longueurs inférieures à quatre cents nucléotides, et 6 semblent faire partie d'ARN plus longs supposés être des ARN messagers. Les 32 candidats restants n'apparaissent pas être transcrits, au moins dans les conditions expérimentales observées.

L'application du pipeline sur des séquences proches

Nous avons repris les données utilisées dans l'expérience précédente afin d'appliquer le pipeline et d'évaluer son potentiel à retrouver des gènes à ARN connus. C'est pourquoi nous avons laissé tous les gènes à ARN connus au moment de notre expérience, c'est-à-dire 171 gènes à ARN car depuis la parution la parution de l'article d'Eddy, les annotations du génome d'*Escherichia coli* K12 ont évolué. Filtrées selon la même procédure qu'Eddy, nous obtenons ainsi 4 353 séquences inter-géniques à comparer aux séquences génomiques des quatre organismes.

A l'issue de la première étape du pipeline, 49 673 alignements sont produits par Yass en filtrant les résultats de la même manière en terme de E-valeur et de longueur, mais en abaissant le seuil sur le pourcentage d'identité à 60%. En effet, contrairement à Qrna, caRNAc traite efficacement les séquences faiblement conservées. Les 171 gènes à ARN présents initialement font partie des alignements retenus. A partir de ces alignements, nous sommes passés à la deuxième étape du pipeline, c'est-à-dire la détection de régions conservées en limitant la taille des régions à 1 000 nucléotides. A l'issue de cette étape, 309 groupes de séquences sont constitués, dont 113 contiennent au moins un fragment des 171 gènes présents à l'origine : 22 ARN ribosomiques, 80 des 89 ARN de transfert et 11 des 60 autres types de gène à ARN restants.

L'application du pipeline sur des séquences éloignées

Pour compléter les résultats obtenus avec l'expérience précédente, nous avons rejoué cette expérience à partir de séquences génomiques d'organismes plus éloignés d'*Escherichia coli*. La figure 4.15 présente l'arbre phylogénétique des quatre organismes sélectionnés dans l'expérience précédente, et des quatre organismes plus éloignés que nous avons choisis :

- *Geobacter sulfurreducens*, souche PCA ;
- *Legionella pneumophila*, souche Corby ;
- *Mycobacterium tuberculosis*, souche F11 ;
- *Rhizobium etli*, souche CIAT 652.

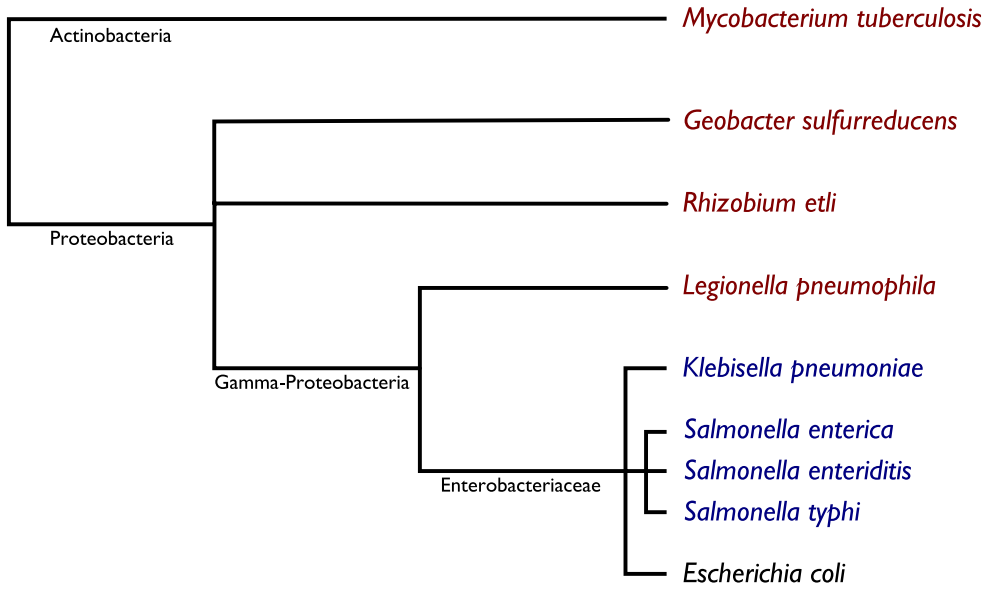


FIG. 4.15 – Arbre phylogénétique reliant les deux groupes de quatre organismes utilisés dans les comparaisons avec *Escherichia coli*.

Les résultats que nous avons obtenus sont les suivants. 1 844 alignements sont générés par YASS à la première étape du pipeline, et 71 groupes de séquences conservées sont détectés. Au sein de ces groupes de séquences on retrouve 87 des 171 gènes à ARN présents dans le génome d'*Escherichia coli* : les 22 ARN ribosomiques, 64 des 89 ARN de transfert et seulement 1 des 60 autres types de gènes à ARN. Parmi les 71 groupes de séquences, 65 présentent au moins une tige conservée par toutes les séquences détectée par caRNAc. Ces 65 groupes intersectent exactement la même quantité de gènes à ARN qu'à l'étape précédente, soit 87 gènes à ARN.

Le volume de données circulant dans le pipeline est drastiquement réduit par rapport à l'expérience précédente, sans pour autant que la quantité de gènes à ARN détectés en sortie soit réduite dans les mêmes proportions. En effet, près de 27 fois moins d'alignements deux à deux sont produits au cours de la première étape et quatre fois moins de régions conservées sont identifiées, mais on retrouve néanmoins 92% des gènes à ARN déjà prédits à partir des séquences proches.

Conclusion

La génomique comparative a connu un essort important au cours de la dernière décennie. En travaillant sur des ensembles de séquences plutôt que sur des séquences isolées, les approches modernes qui prennent part à l'annotation de séquences fonctionnelles s'avèrent particulièrement fécondes. Nos méthodes, Protea et caRNAc, s'inscrivent dans cette dynamique, avec la volonté de réaliser des prédictions de qualité sur des ensembles de séquences hétérogènes en terme de conservation. Protea est dédié à la prédiction de séquences codantes homologues, tandis que caRNAc est dédié à la prédiction de structures secondaires conservées. Protea est une méthode que nous avons développée de bout en bout, alors que caRNAc est le fruit d'un travail initié par Olivier Perriquet durant sa thèse. Nous avons prolongé son travail afin de faire évoluer la méthode et d'y intégrer de nouvelles fonctionnalités.

Protea et caRNAc sont capables de traiter efficacement des ensembles de taille quelconque de séquences dont la longueur peut dépasser plusieurs milliers de bases. Contrairement à la majorité des méthodes existantes qui travaillent sur des séquences alignées, Protea et caRNAc travaillent sur des ensembles de séquences non alignées, évitant ainsi d'être piégés par un alignement de mauvaise qualité. Grâce au concept de méta-séquence que nous avons introduit, les séquences fortement similaires ne constituent plus une redondance d'informations qui pourrait perturber l'analyse comparative et qui est à l'origine de calculs inutiles. Protea et caRNAc procèdent selon un schéma analogue. Dans un premier temps, les séquences sont comparées deux à deux à la recherche d'une séquence d'acides aminés conservée pour Protea, d'une structure secondaire commune pour caRNAc. Puis, chaque méthode combine les résultats obtenus dans un graphe qui lui est spécifique : le graphe des cadres de lecture pour Protea, le graphe des tiges pour caRNAc. Les prédictions sont ensuite réalisées en fonction de certains critères statistiques et propriétés de ces graphes. Les expériences conduites sur les jeux de données de référence ont produit des résultats significatifs, plaçant Protea et caRNAc parmi les méthodes les plus efficaces et les plus performantes dans leurs domaines respectifs, et particulièrement sur les séquences faiblement conservées.

Pistes de recherche à explorer pour Protea

Les comparaisons de séquences d'acides aminés dans Protea sont confiées à un algorithme d'alignement semi-global et aux matrices de substitutions BLOSUM. Il existe d'autres moyens de comparer des séquences d'acides aminés. Dans un premier temps, il serait intéressant de tester d'autres matrices de substitutions, comme les matrices PAM par exemple, et d'autres méthodes de comparaison. Dans l'état actuel, effectuer ces changements n'affecterait en aucun cas le reste de la méthode. A plus long terme, il serait intéressant de proposer un module de comparaison plus fin. Pour cela, on pourrait par exemple avoir recours aux modèles pair-

Markov cachés pour réaliser l'alignement des séquences nucléiques en travaillant au niveau des acides aminés codés et en supportant les changements éventuels du cadre de lecture. Il serait également intéressant de travailler plus finement sur les mutations silencieuses et synonymes attendues et observées entre les séquences. Par exemple, il serait intéressant de proposer système de bonification pour les mutations silencieuses afin de prendre en compte les acides aminés identiques produits par des codons différents.

Pistes de recherche à explorer pour caRNAc

Concernant caRNAc, plusieurs perspectives sont envisageables. La fiabilité des prédictions pourrait être améliorée en fournissant une mesure statistique précise des structures prédites. Pour cela, il faudrait disposer d'une mesure statistique fine pour estimer la qualité des couples de tiges, en fonction de leurs longueurs, de leurs compositions, et surtout de la quantité de mutations simples et compensées qui préservent les appariements. Les améliorations apportées à caRNAc ouvrent également de nouvelles pistes. La version simplifiée de l'algorithme de Sankoff est maintenant suffisamment efficace pour pouvoir envisager un repliement au niveau des nucléotides et non des tiges. Suivre cette voie permettrait notamment de résoudre de manière définitive le problème des tiges maximales chevauchantes qui nous a conduit à modifier l'ordre d'énumération des tiges.

La prédiction des ARN non-codants structurés est un problème complexe. Les méthodes les plus performantes dans ce domaine, telles que RNAz, tirent leur épingle du jeu en ayant recours à des techniques d'apprentissage sophistiquées. Nous avons la conviction qu'il est possible de concevoir une méthode complémentaire des méthodes existantes pour les séquences moyennement et faiblement conservées car les prédictions de caRNAc sont de meilleure qualité que celles de RNAalifold sur lesquelles s'appuie RNAz.

Pistes de recherche communes à Protea et caRNAc

Protea et caRNAc sont bâtis selon des schémas analogues qui reposent sur des comparaisons globales des séquences deux à deux. Dans Protea, on recherche une séquence d'acides aminés conservée à l'aide d'une méthode d'alignement semi-global. L'inconvénient de cette approche est de ne pas prendre en compte la présence d'éventuelles régions non codantes comme les introns ou les extrémités non traduites des ARN. Dans caRNAc, l'adaptation de l'algorithme de Sankoff permet de trouver une structure globalement conservée. caRNAc ne peut donc pas détecter certains éléments de structure conservés dont la localisation varie fortement entre les séquences. Ces choix font que Protea et caRNAc sont bien adaptés à l'analyse de séquences bien "découpées" telles que des séquences provenant de transcriptomes, mais qu'ils ne se sont pas appropriés au traitement de séquences trop "bruitées". Pour dresser un parallèle avec l'alignement classique de séquences, Protea et caRNAc fonctionnent actuellement selon le même procédé que l'alignement global. Il serait intéressant à plusieurs points de vue de les adapter pour qu'ils puissent travailler au niveau local. Protea ne serait ainsi plus sensible au bruit en bordure des séquences ni à la présence de régions non-codantes, potentiellement longues, au sein des séquences. caRNAc ne serait quant à lui plus soumis au phénomène de localisation des éléments structuraux. Pour les deux méthodes, il deviendrait alors envisageable de travailler avec une fenêtre glissante et donc de traiter des ensembles de séquences très longues.

Toujours selon le principe de fenêtre glissante, il serait particulièrement intéressant de proposer un mode de fonctionnement incrémental pour PROTEA et caRNAc. Partant d'un ensemble de séquences homologues, codantes ou structurées, les comparaisons de tous les couples de séquences sont réalisées une seule fois, et le graphe obtenu stocké. A l'aide du fenêtre glissante, on balaye ensuite un génome à la recherche d'une séquence homologue à l'ensemble de séquences pré-traitées, sans refaire de calculs inutiles. Dans l'idée, ce mode de fonctionnement est équivalent aux modèles de covariances utilisés dans des méthodes comme INFERNAL ou ERPIN pour la détection de structures conservées.

Concernant le concept de méta-séquence que nous avons introduit et mis en œuvre dans PROTEA et caRNAc, celui-ci pourrait être affiné. Actuellement, les séquences fortement similaires sont regroupées et représentées par un alignement multiple, tandis que les séquences "uniques" restent inchangées. Ce regroupement réalisé de manière binaire en fonction du pourcentage d'identité pourrait être complété en intégrant des informations phylogénétiques. Ces informations permettraient de pondérer les comparaisons entre méta-séquences dans PROTEA et caRNAc. Ce procédé déjà appliqué dans des méthodes comme EXONIPHY pour la prédiction de séquences codantes ou d'EVOFOLD pour la prédiction d'ARN non-codants semble contribuer à améliorer les résultats notamment entre les séquences qui présentent peu de mutations.

Bibliographie

[AGM+90] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment tool. *Journal of Molecular Biology*, 215(3) :403–410, October 1990.

[AS05] Julien Allali and Marie-France Sagot. A Multiple Graph Layers Model with Application to RNA Secondary Structures Comparison. *String Processing and Information Retrieval*, pages 348–359, 2005.

[BAB+04] Ewan Birney, T. Daniel Andrews, Paul Bevan, Mario Caccamo, Yuan Chen, Laura Clarke, Guy Coates, James Cuff, Val Curwen, Tim Cutts, Thomas Down, Eduardo Eyras, Xose M. Fernandez-Suarez, Paul Gane, Brian Gibbins, James Gilbert, Martin Hammond, Hans-Rudolf Hotz, Vivek Iyer, Kerstin Jekosch, Andreas Kahari, Arek Kasprzyk, Damian Keefe, Stephen Keenan, Heikki Lehvaslaiho, Graham McVicker, Craig Melsopp, Patrick Meidl, Emmanuel Mongin, Roger Pettett, Simon Potter, Glenn Proctor, Mark Rae, Steve Searle, Guy Slater, Damian Smedley, James Smith, Will Spooner, Arne Stabenau, James Stalker, Roy Storey, Abel Ureta-Vidal, K. Cara Woodwark, Graham Cameron, Richard Durbin, Anthony Cox, Tim Hubbard, and Michele Clamp. An overview of Ensembl. *Genome Research*, 14(5) :925–928, 2004. doi:10.1101/gr.1860604.

[BAW+05] Amos Bairoch, Rolf Apweiler, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J. Martin, Darren A. Natale, Claire O'Donovan, Nicole Redaschi, and Lai-Su L. Yeh. The Universal Protein Resource (UniProt). *Nucleic Acids Research*, 33(Suppl 1) :D154–159, 2005. doi:10.1093/nar/gki070.

[BCD04] Ewan Birney, Michele Clamp, and Richard Durbin. GeneWise and Genomewise. *Genome Research*, 14(5) :988–995, 2004. doi:10.1101/gr.1865504.

[BG96] Moisès Burset and Roderic Guigo. Evaluation of gene structure prediction programs. *Genomics*, 34(3) :353–367, June 1996. doi:10.1006/geno.1996.0298.

[BGH+98] Winona C. Barker, John S. Garavelli, Daniel H. Haft, Lois T. Hunt, Christopher R. Marzec, Bruce C. Orcutt, Geetha Y. Srinivasarao, Lai-Su L. Yeh, Robert S. Ledley, Hans-Werner Mewes, Friedhelm Pfeiffer, and Akira Tsugita. The PIR-International Protein Sequence Database. *Nucleic Acids Research*, 26(1) :27–32, 1998. doi:10.1093/nar/26.1.27.

[BH88] Robert E. Bruccoleri and Gerhard Heinrich. An improved algorithm for nucleic acid secondary structure display. *Computational Applications in Biosciences*, 4(1) :167–173, 1988. doi:10.1093/bioinformatics/4.1.167.

[BH00] Vineet Bafna and Daniel H. Huson. The conserved exon method for gene finding. *Proceedings of the 8th International Conference on Intellignet Systems for Molecular Biology ISMB*, 8 :3–12, 2000.

[BHGP94] James W. Brown, Elizabeth S. Haas, Donald G. Gilbert, and Norman R. Pace. The Ribonuclease P Database. *Nucleic Acids Research*, 22(17) :3660–3662, 1994.

[BK97] Christopher B. Burge and Samuel Karlin. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268(1) :78–94, 1997. doi:10.1006/jmbi.1997.0951.

[BK06] Rajnish Bharadwaj and Alex L. Kolodkin. Descrambling Dscam diversity. *Cell*, 125(3) :421–424, May 2006. doi:10.1016/j.cell.2006.04.012.

[BKR+04] Mathieu Blanchette, W. James Kent, Cathy Riemer, Laura Elnitski, Arian F.A. Smit, Krishna M. Roskin, Robert Baertsch, Kate Rosenbloom, Hiram Clawson, Eric D. Green, David Haussler, and Webb Miller. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research*, 14(4) :708–723, April 2004.

[BKR07] Markus Bauer, Gunnar W. Klau, and Knut Reinert. Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization. *BMC Bioinformatics*, 8 :271, 2007. doi:10.1186/1471-2105-8-271.

[BKV96] Bernard Billoud, Milutin Kontic, and Alain Viari. Palingol : a declarative programming language to describe nucleic acids' secondary structures and to scan sequence database. *Nucleic Acids Research*, 24(8) :1395–1403, April 1996.

[BLT93] Mark S. Boguski, Todd M. Lowe, and Carolyn M. Tolstoshev. dbEST – database for "expressed sequence tags". *Nature Genetics*, 4 :332–333, 1993. doi:10.1038/ng0893-332.

[BPM+00a] Serafim Batzoglou, Lior Pachter, Jill P. Mesirov, Bonnie Berger, and Eric S. Lander. Human and Mouse Gene Structure : Comparative Analysis and Application to Exon Prediction. *Genome Research*, 10(7) :950–958, 2000. doi:10.1101/gr.10.7.950.

[BPM+00b] Serafim Batzoglou, Lior Pachter, Jill P. Mesirov, Bonnie Berger, and Eric S. Lander. Human and mouse gene structure : comparative analysis and application to exon prediction. In *Proceedings of the 4th Annual International Conference on Computational Molecular Biology RECOMB*, pages 46–53, 2000.

[Bro99] Michael P.S. Brown. *RNA modeling using stochastic context-free grammars*. PhD thesis, University of California, Santa Cruz, 1999.

[BRS03] Philippe Blayo, Pierre Rouzé, and Marie-France Sagot. Orphan gene finding : an exon assembly approach. *Theoritical Computer Science*, 290(3) :1407–1431, January 2003. doi:10.1016/S0304-3975(02)00043-9.

[BT06] Guillaume Blin and Hélène Touzet. How to Compare Arc-Annotated Sequences : The Alignment Hierarchy. In *String Processing and Information Retrieval (SPIRE)*, volume 4209 of *Lecture Notes in Computer Science*, pages 291–303. Springer Berlin / Heidelberg, 2006. doi:10.1007/11880561_24.

[BWRVdP04] Eric Bonnet, Jan Wuyts, Pierre Rouzé, and Yves Van de Peer. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, 20(17) :2911–2917, 2004.

[BZ04] Vineet Bafna and Shaojie Zhang. FastR : fast database search tool for non-coding RNA. *Proceedings of the IEEE Computer Society Bioinformatics Conference (CSB'04)*, pages 52–61, 2004.

[CBG+05] Liu Changning, Bai Baoyan, Skogerbø Geir, Cai Lun, Deng Wei, Zhang Yong, Bu Dongbo, Zhao Yi, and Chen Runsheng. NONCODE : an integrated knowledge database of non-coding RNAs. *Nucleic Acids Research*, 33(Database issue) :D112–115, 2005.

[CDH01] Richard J. Carter, Inna Dubchak, and Stephen R. Holbrook. A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic Acids Research*, 29(19) :3928–3938, 2001.

[CFKK05] Peter Clote, Fabrizio Ferré, Evangelos Kranakis, and Danny Krizanc. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, 11 :578–591, 2005.

[CK91] David K. Y. Chiu and Ted Kolodziejczak. Inferring consensus structure from nucleic acid sequences. *Computational Applications in Biosciences*, 7(3) :347–352, July 1991.

[CKB04] Alex Coventry, Daniel J. Kleitman, and Bonnie Berger. MSARi : Multiple sequence alignments for statistical detection of RNA secondary structure. *Proceedings of the National Academy of Sciences of the United States of America*, 101(33) :12102–12107, 2004.

[Con08] The UniProt Consortium. The universal protein resource (UniProt). *Nucleic Acids Research*, 36(Database issue) :D190–195, January 2008. doi:10.1093/nar/gkm895.

[Cor88] Florence Corpet. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Research*, 16(22) :10881–10890, 1988.

[CP06] Sourav Chatterji and Lior Pachter. Reference based annotation with GeneMapper. *Genome Biology*, 7(R29), 2006. doi:10.1186/gb-2006-7-4-r29.

[CPL+07] Ségolène Caboche, Maude Pupin, Valérie Leclère, Arnaud Fontaine, Philippe Jacques, and Gregory Kucherov. NORINE : a database of nonribosomal peptides. *Nucleic Acids Research*, 2007. doi:10.1093/nar/gkm792.

[Cri70] Francis Crick. Central Dogma of Molecular Biology. *Nature*, 227 :561–563, August 1970. doi:10.1038/227561a0.

[Cro97] James F. Crow. The high spontaneous mutation rate : is it a health risk ? *Proceedings of the National Academy of Sciences of the United States of America*, 94(16) :8380–8386, August 1997.

[CWC+09] J R Cole, Q Wang, E Cardenas, J Fish, B Chai, R J Farris, A S Kulam-Syed-Mohideen, D M McGarrell, T Marsh, G M Garrity, and J M Tiedje. The Ribosomal Database Project : improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, 37(Database issue) :D141–145, January 2009. doi:10.1093/nar/gkn879.

[DBDH03] Diego Di Bernardo, Thomas Down, and Tim Hubbard. ddbRNA : Detection of conserved secondary structures in multiple alignments. *Bioinformatics*, 19(13) :1606–1611, 2003.

[DBPS07] Arthur L. Delcher, Kirsten A. Bratke, Edwin C. Powers, and Steven L. Salzberg. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, 23(6) :673–679, 2007. doi:10.1093/bioinformatics/btm009.

[DCCC98] John W. Drake, Brian Charlesworth, Deborah Charlesworth, and James F. Crow. Rates of Spontaneous Mutation. *Genetics*, 148 :1667–1686, 1998.

[DCL04] Ye Ding, Chi Yu Chan, and Charles E. Lawrence. Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Research*, 32(Web Server issue) :W135–41, July 2004. doi:10.1093/nar/gkh449.

[DE06] Robin D. Dowell and Sean R. Eddy. Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. *BMC Bioinformatics*, 7 :400, 2006. doi:10.1186/1471-2105-7-400.

[DEKM99] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchinson. *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, July 1999.

[DHK+99] A.L. Delcher, D. Harmon, S. Kasif, Owen White, and S.L. Salzberg. Improved microbial gene identification with GLIMMER. *Nucleic Acids Research*, 27(23) :4636–4641, 1999.

[DL99] Ye Ding and Charles E. Lawrence. A bayesian statistical algorithm for RNA secondary structure prediction. *Computers and Chemistry*, 23(3-4) :387–400, June 1999.

[DL01] Ye Ding and Charles E. Lawrence. Statistical prediction of single-stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond. *Nucleic Acids Research*, 29(5) :1034–1046, March 2001.

[DL03] Ye Ding and Charles E. Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, 31(24) :7280–7301, December 2003.

[DLO97] Mark Dsouza, Niels Larsen, and Ross Overbeek. Searching for patterns in genomic data. *Trends in Genetics*, 13(12) :497–498, December 1997. doi:10.1016/S0168-9525(97)01347-4.

[DS94] Shan Dong and David B. Searls. Gene structure prediction by linguistic methods. *Genomics*, 23(3) :540–551, October 1994. doi:10.1006/geno.1994.1541.

[DWB06] Chuong B. Do, Daniel A. Woods, and Serafim Batzoglou. CONTRAfold : RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22(14) :e90–8, July 2006. doi:10.1093/bioinformatics/btl246.

[DWMS06] Deniz Dalli, Andreas Wilm, Indra Mainz, and Gerhard Steger. STRAL : progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time. *Bioinformatics*, 22(13) :1593–1599, July 2006. doi:10.1093/bioinformatics/btl142.

[ED94] Sean R. Eddy and Richard Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22(11) :2079–2088, June 1994.

[Edd01] Sean R. Eddy. Non-coding RNA genes and the modern RNA world. *Nature Reviews Genetics*, 2(12) :919–929, 2001. doi:10.1038/35103511.

[Eng06] Stefan Engelen. *Algorithmes pour la prédiction de structures secondaires d'ARN*. PhD thesis, Université d'Evry Val d'Essonne, 2006.

[ET07] Stefan Engelen and Fariza Tahi. Predicting RNA secondary structure by the comparative approach : how to select the homologous sequences. *BMC Bioinformatics*, 8 :464, 2007.

[FBG07] Eva K. Freyhult, Jonathan P. Bollback, and Paul P. Gardner. Exploring genomic dark matter : a critical assessment of the performance of homology search methods on noncoding RNA. *Genome Research*, 17(1) :117–125, January 2007. doi:10.1101/gr.5890907.

[FdMT08] Arnaud Fontaine, Antoine de Monte, and Hélène Touzet. MAGNOLIA : multiple alignment of protein-coding and structural RNA sequences. *Nucleic Acids Research*, 36(Web Server issue) :W14–W18, 2008. doi:10.1093/nar/gkn321.

[FHL+07] Laurent Fousse, Guillaume Hanrot, Vincent Lefèvre, Patrick Pélissier, and Paul Zimmermann. MPFR : A multiple-precision binary floating-point library with correct rounding. *ACM Transactions on Mathematical Software*, 33(2) :13, 2007. doi:http://doi.acm.org/10.1145/1236463.1236468.

[FHS00] Martin Fekete, Ivo L. Hofacker, and Peter F. Stadler. Prediction of RNA base pairing probabilities on massively parallel computers. *Journal of Computational Biology*, 7(1-2) :171–182, 2000. doi:10.1089/10665270050081441.

[FHZ+98] Liliane Florea, George Hartzell, Zheng Zhang, Gerald M. Rubin, and Webb Miller. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Research*, 8(9) :967–974, September 1998.

[Fic95] James W. Fickett. ORFs and genes : how strong a connection ? *Journal of Computational Biology*, 2(1) :117–123, 1995.

[FMSB+06] Robert D. Finn, Jaina Mistry, Benjamin Schuster-Bockler, Sam Griffiths-Jones, Volker Hollich, Timo Lassmann, Simon Moxon, Mhairi Marshall, Ajay Khanna, Richard Durbin, Sean R Eddy, Erik L L Sonnhammer, and Alex Bateman. Pfam : clans, web tools and services. *Nucleic Acids Research*, 34(Database issue) :D247–D251, 2006. doi:10.1093/nar/gkj149.

[FSY+99] Yoshifumi Fukunishi, Harukazu Suzuki, Masayasu Yoshino, Hideaki Konno, and Yoshihide Hayashizaki. Prediction of human cDNA from its homologous mouse full-length cDNA and human shotgun database. *FEBS Letters*, 464(3) :129–132, December 1999.

[FT92] James W. Fickett and Chang-Shung Tung. Assessment of protein coding measures. *Nucleic Acids Research*, 20(24) :6441–6450, 1992. doi:10.1093/nar/20.24.6441.

[FT07] Arnaud Fontaine and Hélène Touzet. Computational identification of protein-coding sequences by comparative analysis. In *Proceedings of the 1st IEEE international conference on Bioinformatics and Biomedecine (BIBM), Silicon Valley, California*, pages 95–102. IEEE Computer Society, 2007. doi:10.1109/BIBM.2007.11.

[FT09] Arnaud Fontaine and Hélène Touzet. Computational identification of protein-coding sequences by comparative analysis. *International Journal of Data Mining and Bioinformatics*, 2009. to appear.

[GA96] Raymond D. Gesteland and John F. Atkins. Recoding : dynamic reprogramming of translation. *Annual Review of Biochemistry*, 65 :741–768, 1996. doi:10.1146/annurev.bi.65.070196.003521.

[GB06] Samuel S Gross and Michael R Brent. Using multiple alignments to improve gene prediction. *Journal of Computational Biology*, 13(2) :379–393, March 2006. doi:10.1089/cmb.2006.13.379.

[GBSK05] Gordon Gremme, Volker Brendel, Michael E. Sparks, and Stefan Kurtz. Engineering a software tool for gene structure prediction in higher organisms. *Information and Software Technology*, 47(15) :965–978, December 2005. doi:10.1016/j.infsof.2005.09.005.

[Gel95] Mikhail S. Gelfand. Prediction of function in DNA sequence analysis. *Journal of Computational Biology*, 2(1) :87–115, 1995.

[GG82] Manolo Gouy and Christian Gautier. Codon usage in bacteria : correlation with gene expressivity. *Nucleic Acids Research*, 10(22) :7055–7074, 1982.

[GG04] Paul P. Gardner and Robert Giegerich. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, 5 :140, September 2004. doi:10.1186/1471-2105-5-140.

[GGG80] Richard Grantham, Christian Gautier, and Manolo Gouy. Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic Acids Research*, 8(9) :1893–1912, May 1980.

[GHH+94] Leslie Grate, Mark Herbster, Richard Hughey, David Haussler, I. Saira Mian, and Harry Noller. RNA modeling using Gibbs sampling and stochastic context free grammars. *Proceedings of the 2nd International Conference on Intelligent Systems for Molecular Biology ISMB*, 2 :138–146, 1994.

[GHKB00] Gaston H. Gonnet, Michael T. Hallett, Chantal Korostensky, and L Bernardin. Darwin v. 2.0 : an interpreted computer language for the biosciences. *Bioinformatics*, 16(2) :101–103, February 2000.

[GJ06] Sam Griffiths-Jones. miRBase : the microRNA sequence database. *Methods Molecular BIology*, 342 :129–138, 2006. doi:10.1385/1-59745-123-1:129.

[GJBM+03] Sam Griffiths-Jones, Alex Bateman, Mhairi Marshall, Ajay Khanna, and Sean R. Eddy. Rfam : an RNA family database. *Nucleic Acids Research*, 33(1) :439–441, 2003.

[GJGvD+06] Sam Griffiths-Jones, Russell J. Grocock, Stijn van Dongen, Alex Bateman, and Anton J. Enright. miRBase : microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34(Database issue) :D140–4, January 2006. doi:10.1093/nar/gkj112.

[GJMM+05] Sam Griffiths-Jones, Simon Moxon, Mhairi Marshall, Ajay Khanna, Sean R. Eddy, and Alex Bateman. Rfam : annotating non-coding RNAs in complete genomes. *Nucleic Acids Research*, 33(Database issue) :D121–D124, 2005. doi:10.1093/nar/gki081.

[GL01] Daniel Gautheret and André Lambert. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *Journal of Molecular Biology*, 313(5) :1003–1011, November 2001. doi:10.1006/jmbi.2001.5102.

[GLL+03] Giorgio Grillo, Flavio Licciulli, Sabino Liuni, Elisabetta Sbisa, and Graziano Pesole. PatSearch : A program for the detection of patterns and structural motifs in nucleotide sequences. *Nucleic Acids Research*, 31(13) :3608–3612, July 2003.

[GMC90] Daniel Gautheret, Francois Major, and Robert Cedergren. Pattern searching/alignment with RNA primary and secondary structures : an effective descriptor for tRNA. *Computational Applications in Biosciences*, 6(4) :325–331, 1990.

[GMP96] Mikhail S. Gelfand, Andrey A. Mironov, and Pavel A. Pevzner. Gene recognition via spliced sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*, 93(17) :9061–9066, August 1996.

[GS03] Alison P. Galvani and Montgomery Slatkin. Evaluating plague and smallpox as historical selective pressures for the CCR5-Δ32 HIV-resistance allele. *Proceedings of the National Academy of Sciences of the United States of America*, 100(25) :15276–15279, 2003. doi:10.1073/pnas.2435085100.

[Gui98] Roderic Guigo. Assembling genes from predicted exons in linear time with dynamic programming. *Journal of Computational Biology*, 5(4) :681–702, 1998.

[Gum58] Emil J. Gumbel. *Statistics of extremes*. Columbia University Press, 1958.

[GW08] Tanja Gesell and Stefan Washietl. Dinucleotide controlled null models for comparative RNA gene prediction. *BMC Bioinformatics*, 9 :248, 2008. doi:10.1186/1471-2105-9-248.

[GWW05] Paul P. Gardner, Andreas Wilm, and Stefan Washietl. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Research*, 33(8) :2433–2439, 2005. doi:10.1093/nar/gki541.

[HAZK97] Xiaoqiu Huang, Mark D. Adams, Hao Zhou, and Anthony R. Kerlavage. A tool for analyzing and annotating genomic sequences. *Genomics*, 46(1) :37–45, November 1997. doi:10.1006/geno.1997.4984.

[HBB+02] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater,

J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, and M. Clamp. The Ensembl genome database project. *Nucleic Acids Research*, 30(1) :38–41, 2002.

[HBS04] Ivo L. Hofacker, Stephan H. F. Bernhart, and Peter F. Stadler. Alignment of RNA base pairing probability matrices. *Bioinformatics*, 20(14) :2222–2227, September 2004. doi:10.1093/bioinformatics/bth229.

[HFS+94] Ivo L. Hofacker, Walter Fontana, Peter F. Stadler, Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie*, 125 :167–188, 1994.

[HFS02] Ivo L. Hofacker, Martin Fekete, and Peter F. Stadler. Secondary structure prediction for aligned RNA sequences. *Journal of Molecular Biology*, 319(5) :1059–1066, June 2002. doi:10.1016/S0022-2836(02)00308-X.

[HGK97] Martijn Huynen, Robin R. Gutell, and Danielle Konings. Assessing the reliability of RNA folding using statistical mechanics. *Journal of Molecular Biology*, 267(5) :1104–1112, April 1997. doi:10.1006/jmbi.1997.0889.

[HH92] Steven Henikoff and Jorja G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89 :10915–10919, November 1992. doi:10.1073/pnas.89.22.10915.

[HKC+06] Fan Hsu, W James Kent, Hiram Clawson, Robert M Kuhn, Mark Diekhans, and David Haussler. The UCSC Known Genes. *Bioinformatics*, 22(9) :1036–1046, May 2006. doi:10.1093/bioinformatics/btl048.

[HLG05] Jakob H. Havgaard, Rune B. Lyngsø, and Jan Gorodkin. The FOLDALIGN web server for pairwise structural RNA alignment and mutual motif search. *Nucleic Acids Research*, 33(Web Server issue) :W650–3, July 2005. doi:10.1093/nar/gki473.

[Hol05] Ian Holmes. Accelerated probabilistic inference of RNA structure evolution. *BMC Bioinformatics*, 6 :73, 2005. doi:10.1186/1471-2105-6-73.

[HSM07] Arif Ozgun Harmanci, Gaurav Sharma, and David H Mathews. Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign. *BMC Bioinformatics*, 8 :130, 2007. doi:10.1186/1471-2105-8-130.

[HSP05] Alexander Hüttenhofer, Peter Schattner, and Norbert Polacek. Non-coding RNAs : hope or hype ? *Trends in Genetics*, 21(5) :289–297, May 2005. doi:10.1016/j.tig.2005.03.007.

[HTG07] Jakob H Havgaard, Elfar Torarinsson, and Jan Gorodkin. Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. *PLoS Computational Biology*, 3(10) :1896–1908, October 2007. doi:10.1371/journal.pcbi.0030193.

[HTGK03] Matthias Hochsmann, Thomas Toller, Robert Giegerich, and Stefan Kurtz. Local similarity in RNA secondary structures. *Proceedings of the IEEE Computer Society Bioinformatics Conference*, 2 :159–168, 2003.

[HVG04] Matthias Hochsmann, Bjorn Voss, and Robert Giegerich. Pure multiple RNA secondary structure alignments : a progressive profile approach. *IEEE/ACM*

Transactions on Computational Biology and Bioinformatics, 1(1) :53–62, 2004. doi:10.1109/TCBB.2004.11.

[Ike81a] Toshimichi Ikemura. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. *Journal of Molecular Biology*, 146(1) :1–21, February 1981.

[Ike81b] Toshimichi Ikemura. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes : a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *Journal of Molecular Biology*, 151(3) :389–409, September 1981.

[Ike82] Toshimichi Ikemura. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. *Journal of Molecular Biology*, 158(4) :573–597, July 1982.

[Jac88] Tyler E. Jacks. *Ribosomal frameshifting in retroviral gene expression*. PhD thesis, University of California, 1988.

[JJ98] Jian Jiang and Howard J. Jacob. EbEST : an automated tool using expressed sequence tags to delineate gene structure. *Genome Research*, 8(3) :268–275, March 1998.

[JLMZ02] Tao Jiang, Guohui Lin, Bin Ma, and Kaizhong Zhang. A General Edit Distance between RNA Structures. *Journal of Computational Biology*, 9(2) :371–388, 2002. doi:10.1089/10665270252935511.

[JTZ89] John A. Jaeger, Douglas H. Turner, and Michael Zuker. Improved predictions of secondary structures for RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 86(20) :7706–7710, October 1989.

[JTZ90] John A. Jaeger, Douglas H. Turner, and Michael Zuker. Predicting optimal and suboptimal secondary structure for RNA. *Methods in Enzymology*, 183 :281–306, 1990.

[KA90] Samuel Karlin and Stephen F. Altschul. Methods for assessing the statistical significance of molecular sequence feature by using general scoring schemes. *Proceedings of the National Academy of Sciences of the United States of America*, 87 :2264–2268, 1990.

[KB92] Samuel Karlin and Volker Brendel. Chance and significance in protein and DNA sequence analysis. *Science*, 257 :39–49, 1992.

[KBD+03] D. Karolchik, R. Baertsch, M. Diekhans, T.S. Furey, A. Hinrichs, Y.T. Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, D.J. Thomas, R.J. Weber, D. Haussler, and W.J. Kent. The UCSC Genome Browser Database. *Nucleic Acids Research*, 31(1) :51–54, 2003.

[KC07] Keith Knapp and Yi-Ping Phoebe Chen. An evaluation of contemporary hidden Markov model genefinders with a predicted exon taxonomy. *Nucleic Acids Research*, 35(1) :317–324, 2007. doi:10.1093/nar/gkl1026.

[KE03] Robert J. Klein and Sean R. Eddy. RSEARCH : Finding homologs of single structured RNA sequences. *BMC Bioinformatics*, 4(1) :44, 2003.

[KFDB01] Ian Korf, Paul Flicek, Daniel Duan, and Michael R. Brent. Integrating genomic homology into gene structure prediction. *Bioinformatics*, 17(suppl 1) :S140–S148, 2001.

[KH94] Ben F. Koop and Leroy Hood. Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. *Nature Genetics*, 7(1) :48–53, May 1994. doi:10.1038/ng0594-48.

[KH99] Bjarne Knudsen and Jotun Hein. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 15(6) :446–454, June 1999.

[KH03] Bjarne Knudsen and Jotun Hein. Pfold : RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Research*, 31(13) :3423–3428, July 2003.

[KHF+04] D. Karolchik, A.S. Hinrichs, T.S. Furey, K.M. Roskin, C.W. Sugnet, D. Haussler, and W.J. Kent. The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, 32(Suppl 1) :D493–496, 2004.

[KHRE96] David Kulp, David Haussler, Martin G. Reese, and Frank H. Eeckman. A generalized hidden Markov model for the recognition of human genes in DNA. *Proceedings of the 4th International Conference on Intellignet Systems for Molecular Biology ISMB*, 4 :134–142, 1996.

[KME02] Robert J. Klein, Ziva Misulovin, and Sean R. Eddy. Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proceedings of the National Academy of Sciences of the United States of America*, 99(11) :7542–7547, May 2002. doi:10.1073/pnas.112063799.

[KMH94] Anders Krogh, I. Saira Mian, and David Haussler. A hidden Markov model that finds genes in E. coli DNA. *Nucleic Acids Research*, 22(22) :4768–4778, November 1994.

[KO87] Samuel Karlin and Friedemann Ost. Counts of long aligned word matches among random letter sequences. *Advances in applied probability*, 19 :293–351, 1987.

[KO88] Samuel Karlin and Friedemann Ost. Maximal length of common words among random letter sequences. *Annals of Probability*, 16 :535–563, 1988.

[KS01] Sasivimol Kittivoravitkul and Marek Sergot. PAGAN : Predict and Annotate Genes in genomic sequence based on ANalysis of EST Clusters. In *International Conference on Intellignet Systems for Molecular Biology ISMB*, 2001.

[KTHB02] Peter S. Klosterman, Makio Tamura, Stephen R. Holbrook, and Steven E. Brenner. SCOR : a structural classification of RNA database. *Nucleic Acids Research*, 30(1) :392–394, 2002.

[KTKA07] Hisanori Kiryu, Yasuo Tabei, Taishin Kin, and Kiyoshi Asai. Murlet : a practical multiple alignment tool for structural RNA sequences. *Bioinformatics*, 23(13) :1588–1598, July 2007. doi:10.1093/bioinformatics/btm146.

[KYT+07] Taishin Kin, Kouichirou Yamada, Goro Terai, Hiroaki Okida, Yasuhiko Yoshinari, Yukiteru Ono, Aya Kojima, Yuki Kimura, Takashi Komori, and Kiyoshi

Asai. fRNAdb : a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic Acids Research*, 35(Database issue) :D145–D148, January 2007. doi:10.1093/nar/gkl837.

[LB98] Alexander V. Lukashin and Mark Borodovsky. GeneMark.hmm : new solution for gene finding. *Nucleic Acids Research*, 26(4) :1107–1115, 1998.

[LFL+04] André Lambert, Jean-Fred Fontaine, Matthieu Legendre, Fabrice Leclerc, Emmanuelle Permal, François Major, Harald Putzer, Olivier Delfour, Bernard Michot, and Daniel Gautheret. The ERPIN server : an interface to profile-based RNA motif identification. *Nucleic Acids Research*, 32(Web Server issue) :W160–5, July 2004. doi:10.1093/nar/gkh418.

[LGC94] Alain Laferriere, Daniel Gautheret, and Robert Cedergren. An RNA pattern matching program with enhanced performance and portability. *Computational Applications in Biosciences*, 10(2) :211–212, April 1994.

[LLFG05] Andre Lambert, Matthieu Legendre, Jean-Fred Fontaine, and Daniel Gautheret. Computing expectation values for RNA motifs using discrete convolutions. *BMC Bioinformatics*, 6 :118, 2005. doi:10.1186/1471-2105-6-118.

[LMS+95] Jane E. Lamerdin, Mishcelle A. Montgomery, Stephanie A. Stilwagen, Lisa K. Scheidecker, Robert S. Tebbs, Kerry W. Brookman, Larry H. Thompson, and Anthony V. Carrano. Genomic sequence comparison of the human and mouse XRCC1 DNA repair gene regions. *Genomics*, 25(2) :547–554, January 1995.

[LMZ04] Shu-Yun Le, Jacob V. Jr Maizel, and Kaizhong Zhang. An algorithm for detecting homologues of known structured RNAs in genomes. *Proceedings of the IEEE Computer Society Bioinformatics Conference (CSB'04)*, pages 300–310, 2004.

[LP85] David J. Lipman and William R. Pearson. Rapid and sensitive protein similarity searches. *Science*, 227(4693) :1435–1441, 1985. doi:10.1126/science.2983426.

[LS79] Michael R. Lerner and Joan A. Steitz. Antibodies to small nuclear RNAs complexed with proteins are produced by patients with systemic lupus erythematosus. *Proceedings of the National Academy of Sciences of the United States of America*, 76(11) :5495–5499, November 1979.

[LTHCB05] Alexandre Lomsadze, Vardges Ter-Hovhannisyan, Yury O. Chernoff, and Mark Borodovsky. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research*, 33(20) :6494–6506, 2005. doi:10.1093/nar/gki937.

[LW01] Neocles B. Leontis and Eric Westhof. Geometric nomenclature and classification of RNA base pairs. *RNA*, 7(4) :499–512, April 2001.

[LWHT05] Jianghui Liu, Jason T. L. Wang, Jun Hu, and Bin Tian. A method for aligning RNA secondary structures and its application to RNA motif detection. *BMC Bioinformatics*, 6 :89, 2005. doi:10.1186/1471-2105-6-89.

[LZP99] Rune B. Lyngsø, Michael Zuker, and Christian N. S. Pedersen. Fast evaluation of internal loops in RNA secondary structure prediction. *Bioinformatics*, 15(6) :440–445, June 1999.

[Mar08] Elliott H Margulies. Confidence in comparative genomics. *Genome Research*, 18(2) :199–200, February 2008. doi:10.1101/gr.7228008.

[MB98] Wojciech Makalowski and Mark S. Boguski. Evolutionary parameters of the transcribed mammalian genome : an analysis of 2,820 orthologous rodent and human sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 95(16) :9407–9412, August 1998.

[McC90] John S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7) :1105–1119, 1990. doi:10.1002/bip.360290621.

[MD04] Irmtraud M Meyer and Richard Durbin. Gene structure conservation aids similarity based gene prediction. *Nucleic Acids Research*, 32(2) :776–783, 2004. doi:10.1093/nar/gkh211.

[MDC+04] David H. Mathews, Matthew D. Disney, Jessica L. Childs, Susan J. Schroeder, Michael Zuker, and Douglas H. Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences of the United States of America*, 101(19) :7287–7292, May 2004. doi:10.1073/pnas.0401799101.

[MEG+01] Thomas J. Macke, David J. Ecker, Robin R. Gutell, Daniel Gautheret, David A. Case, and Rangarajan Sampath. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Research*, 29(22) :4724–4735, November 2001.

[MH97] B. Edward H. Maden and John M. Hughes. Eukaryotic ribosomal RNA : the recent excitement in the nucleotide modification problem. *Chromosoma*, 105(7-8) :391–400, June 1997.

[Mor99] Burkhard Morgenstern. DIALIGN 2 : improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, 15(3) :211–218, 1999.

[MSSR02] Catherine Mathe, Marie-France Sagot, Thomas Schiex, and Pierre Rouze. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Research*, 30(19) :4103–4117, 2002.

[MSZT99] David H. Matthews, Jeffrey Sabina, Michael Zuker, and Douglas H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288(5) :911–940, May 1999. doi:10.1006/jmbi.1999.2700.

[MT02] David H. Mathews and Douglas H. Turner. Dynalign : an algorithm for finding the secondary structure common to two RNA sequences. *Journal of Molecular Biology*, 317(2) :191–203, March 2002. doi:10.1006/jmbi.2001.5351.

[MTL02] Bin Ma, John Tromp, and Ming Li. PatternHunter : faster and more sensitive homology search. *Bioinformatics*, 18(3) :440–445, 2002. doi:10.1093/bioinformatics/18.3.440.

[MWH+08] Sebastien Moretti, Andreas Wilm, Desmond G Higgins, Ioannis Xenarios, and Cedric Notredame. R-Coffee : a web server for accurately aligning noncoding

RNA sequences. *Nucleic Acids Research*, 36(Web Server issue) :W10–W13, July 2008. doi:10.1093/nar/gkn278.

[MYH+08] Toutai Mituyama, Kouichirou Yamada, Emi Hattori, Hiroaki Okida, Yukiteru Ono, Goro Terai, Aya Yoshizawa, Takashi Komori, and Kiyoshi Asai. The Functional RNA Database 3.0 : databases to support mining and annotation of functional RNAs. *Nucleic Acids Research*, October 2008. doi:10.1093/nar/gkn805.

[MZB96] Wojciech Makalowski, Jinghui Zhang, and Mark S. Boguski. Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Research*, 6(9) :846–857, September 1996.

[NE07] Eric P. Nawrocki and Sean R. Eddy. Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Computational Biology*, 3(3) :e56, March 2007. doi:10.1371/journal.pcbi.0030056.

[NGM01] Pavel S. Novichkov, Mikhail S. Gelfand, and Andrey A. Mironov. Gene recognition in eukaryotic DNA by comparison of genomic sequences. *Bioinformatics*, 17(11) :1011–1018, 2001.

[NHH00] Cédric Notredame, Desmond G. Higgins, and Jaap Heringa. T-Coffee : A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302 :205–217, 2000.

[NJ80] Ruth Nussinov and Ann B. Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 77(11) :6309–6313, 1980.

[NK05] Laurent Noé and Gregory Kucherov. YASS : enhancing the sensitivity of DNA similarity search. *Nucleic Acids Research*, 33(suppl2) :W540–543, 2005.

[NPGK78] Ruth Nussinov, George Piecznik, Jerrold R. Grigg, and Daniel J. Kleitman. Algorithms for loop matchings. *SIAM Journal on Applied Mathematics*, 35(1) :68–82, July 1978.

[NW70] Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3) :443–453, 1970.

[OSKR06] Aleksey Y. Ogurtsov, Svetlana A. Shabalina, Alexey S. Kondrashov, and Mikhail A. Roytberg. Analysis of internal loops within the RNA secondary structure in almost quadratic time. *Bioinformatics*, 22(11) :1317–1324, June 2006. doi:10.1093/bioinformatics/btl083.

[PAA+03] Genis Parra, Pankaj Agarwal, Josep F. Abril, Thomas Wiehe, James W. Fickett, and Roderic Guigò. Comparative Gene Prediction in Human and Mouse. *Genome Research*, 13(1) :108–117, 2003. doi:10.1101/gr.871403.

[PBG00] Genis Parra, Enrique Blanco, and Roderic Guigò. GeneID in Drosophilia. *Genome Research*, 10(4) :511–515, April 2000. doi:10.1101/gr.10.4.511.

[PBS+06] Jakob Skou Pedersen, Gill Bejerano, Adam Siepel, Kate Rosenbloom, Kerstin Lindblad-Toh, Eric S Lander, Jim Kent, Webb Miller, and David Haussler. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Computational Biology*, 2(4) :e33, April 2006. doi:10.1371/journal.pcbi.0020033.

[Per03] Olivier Perriquet. *Approche algorithmique de la prédiction de structures secondaires*. PhD thesis, Université des Sciences et Technologies de Lille, December 2003.

[PH03] Jakob Skou Pedersen and Jotun Hein. Gene finding with a hidden Markov model of genome structure and evolution. *Bioinformatics*, 19(2) :219–227, 2003.

[PLD00] Graziano Pesole, Sabino Liuni, and Mark Dsouza. PatSearch : a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance. *Bioinformatics*, 16(5) :439–450, May 2000.

[PM08] Marc Parisien and Francois Major. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, 452(7183) :51–55, March 2008. doi:10.1038/nature06684.

[PSD+07] Ken C. Pang, Stuart Stephen, Marcel E. Dinger, Par G. Engstrom, Boris Lenhard, and John S. Mattick. RNAdb 2.0–an expanded database of mammalian non-coding RNAs. *Nucleic Acids Research*, 35(Database issue) :D178–82, January 2007. doi:10.1093/nar/gkl926.

[PSE+05] Ken C. Pang, Stuart Stephen, Par G. Engstrom, Khairina Tajul-Arifin, Weisan Chen, Claes Wahlestedt, Boris Lenhard, Yoshihide Hayashizaki, and John S. Mattick. RNAdb–a comprehensive mammalian noncoding RNA database. *Nucleic Acids Research*, 33(Database issue) :D125–30, January 2005. doi:10.1093/nar/gki089.

[PTD03] Olivier Perriquet, Hélène Touzet, and Max Dauchet. Finding the common structure shared by two homologous RNAs. *Bioinformatics*, 19(1) :108–116, January 2003.

[RAG97] Mikhail A. Roytberg, Tatiana V. Astakhova, and Mikhail S. Gelfand. Combinatorial approaches to gene recognition. *Computers and Chemistry*, 21(4) :229–235, 1997.

[RDM99] Igor B. Rogozin, Dino D'Angelo, and Luciano Milanesi. Protein-coding regions prediction combining similarity searches and conservative evolutionary properties of protein-coding sequences. *Gene*, 226 :126–137, 1999.

[RE99] Elena Rivas and Sean R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of Molecular Biology*, 285(5) :2053–2068, February 1999. doi:10.1006/jmbi.1998.2436.

[RE00] Elena Rivas and Sean R. Eddy. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, 6 :583–605, 2000.

[RE01] Elena Rivas and Sean R. Eddy. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2, 2001.

[RG05] Jens Reeder and Robert Giegerich. Consensus shapes : an alternative to the Sankoff algorithm for RNA consensus structure prediction. *Bioinformatics*, 21(17) :3516–3523, September 2005. doi:10.1093/bioinformatics/bti577.

[RKJE01] Elena Rivas, Robert J. Klein, Thomas A. Jones, and Sean R. Eddy. Computational identification of noncoding RNAs in E. coli by comparative genomics. *Current Biology*, 11(17) :1369–1373, September 2001.

[RMK96] Igor B. Rogozin, Luciano Milanesi, and Nikolay A. Kolchanov. Gene structure prediction using information on homologous protein sequence. *Computational Applications in Biosciences*, 12(3) :161–170, June 1996.

[RMO01] Sanja Rogic, Alan K. Mackworth, and Francis B. F. Ouellette. Evaluation of gene-finding programs on mammalian sequences. *Genome Research*, 11(5) :817–832, 2001. doi:10.1101/gr.147901.

[RSG07] Jens Reeder, Peter Steffen, and Robert Giegerich. pknotsRG : RNA pseudoknot folding including near-optimal structures and sliding windows. *Nucleic Acids Research*, 35(Web Server issue) :W320–4, July 2007. doi:10.1093/nar/gkm258.

[RSZ04a] Jianhua Ruan, Gary D. Stormo, and Weixiong Zhang. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, 20(1) :58–66, January 2004.

[RSZ04b] Jianhua Ruan, Gary D. Stormo, and Weixiong Zhang. ILM : a web server for predicting RNA secondary structures with pseudoknots. *Nucleic Acids Research*, 32(Web Server issue) :W146–9, July 2004. doi:10.1093/nar/gkh444.

[Rus93] Peter J. Russell. *Fundamentals of Genetics and the Biology Place*. Pearson Education, Limited, 1993.

[Ruv01] G Ruvkun. Molecular biology. Glimpses of a tiny RNA world. *Science*, 294(5543) :797–799, October 2001. doi:10.1126/science.1066315.

[San85] David Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM Journal on Applied Mathematics*, 45 :810–825, 1985. doi:10.1137/0145048.

[SB05] Sven Siebert and Rolf Backofen. MARNA : multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics*, 21(16) :3352–3359, August 2005. doi:10.1093/bioinformatics/bti550.

[SC09] Dietmar Schmucker and Brian Chen. Dscam and DSCAM : complex genes in simple animals, complex animals yet simple genes. *Genes & Development*, 23(2) :147–156, January 2009. doi:10.1101/gad.1752909.

[Sch02] Peter Schattner. Searching for RNA genes using base-composition statistics. *Nucleic Acids Research*, 30(9) :2076–2082, 2002.

[SDKW98] Steven L. Salzberg, Arthur L. Delcher, Simon Kasif, and Owen White. Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*, 26(2) :544–548, 1998.

[Sea92] David B. Searls. The Linguistics of DNA. *American Scientist*, 80 :579–591, 1992.

[SG94] David J. States and Warren Gish. Combined use of sequence similarity and codon bias for coding region identification. *Journal of Computational Biology*, 1(1) :39–50, 1994.

[SH04] Adam Siepel and David Haussler. Computational identification of evolutionarily conserved exons. In *Proceedings of the 8th Annual International Conference on Resaerch in Computational Molecular Biology*, pages 177–186, New York, NY, USA, 2004. ACM Press. doi:10.1145/974614.974638.

[SH05] Adam Siepel and David Haussler. Phylogenetic hidden Markov models. In R. Nielsen, editor, *Statistical Methods in Molecular Evolution*, pages 325–351. Springer, New York, 2005.

[SKBA78] Benjamin C. Stark, Ryszard Kole, Emma J. Bowman, and Sidney Altman. Ribonuclease P : an enzyme with an essential RNA component. *Proceedings of the National Academy of Sciences of the United States of America*, 75(8) :3717–3721, August 1978.

[SM82] Rodger Staden and Alan D. McLachlan. Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Research*, 10(1) :141–156, 1982.

[SMR01] Thomas Schiex, Annick Moisan, and Pierre Rouzé. EuGene : An Eucaryotic Gene Finder that combines several sources of evidence. In Olivier Gascuel and Marie-France Sagot, editors, *Computational Biology*, pages 111–125. Lecture Notes in Computer Science 2066, 2001.

[SPD+99] Steven L. Salzberg, Mihaela Pertea, Arthur L. Delcher, Malcolm J. Gardner, and Hervé Tettelin. Interpolated Markov models for eukaryotic gene finding. *Genomics*, 59(1) :24–31, July 1999. doi:10.1006/geno.1999.5854.

[SS95] Eric E. Snyder and Gary D. Stormo. Identification of protein coding regions in genomic DNA. *Journal of Molecular Biology*, 248(1) :1–18, 1995.

[SS00] Asaf A. Salamov and Victor V. Solovyev. Ab initio Gene Finding in Drosophila Genomic DNA. *Genome Research*, 10(4) :516–522, March 2000. doi:10.1101/gr.10.4.516.

[Sta03] Mario Stanke. *Gene Prediction with a Hidden Markov Model*. PhD thesis, Mathematisch-Naturwissenschaftlichen Fakultäten der Georg-August-Universität zu Göttingen, 2003.

[Sto90] Gary D. Stormo. Consensus patterns in DNA. *Methods in Enzymology*, 183 :211–221, 1990.

[SVR+06] Peter Steffen, Bjorn Voss, Marc Rehmsmeier, Jens Reeder, and Robert Giegerich. RNAshapes : an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, 22(4) :500–503, February 2006. doi:10.1093/bioinformatics/btk010.

[SW03] Mario Stanke and Stephan Waack. Gene Prediction with a Hidden-Markov Model and a new Intron Submodel. *Bioinformatics*, 19(suppl 2) :ii215–ii225, 2003.

[TdGSG06] Patricia Thebault, Simon de Givry, Thomas Schiex, and Christine Gaspin. Searching RNA motifs and their intermolecular contacts with constraint networks. *Bioinformatics*, 22(17) :2074–2080, September 2006. doi:10.1093/bioinformatics/btl354.

[TER03] Fariza Tahi, Stefan Engelen, and Mireille Régnier. A Fast Algorithm for RNA Secondary Structure Prediction Including Pseudoknots. In *BIBE*, pages 11–17, 2003.

[TGR02] Fariza Tahi, Manolo Gouy, and Mireille Regnier. Automatic RNA secondary structure prediction with a comparative approach. *Computers and Chemistry*, 26(5) :521–530, July 2002.

[THG94] Julie D. Thompson, Desmond G. Higgins, and Toby J. Gibson. CLUSTAL W : improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22) :4673–4680, 1994.

[THG07] Elfar Torarinsson, Jakob H. Havgaard, and Jan Gorodkin. Multiple structural alignment and clustering of RNA sequences. *Bioinformatics*, 23(8) :926–932, April 2007. doi:10.1093/bioinformatics/btm049.

[TKKA08] Yasuo Tabei, Hisanori Kiryu, Taishin Kin, and Kiyoshi Asai. A fast structural multiple alignment method for long RNA sequences. *BMC Bioinformatics*, 9 :33, 2008. doi:10.1186/1471-2105-9-33.

[TMM07] Leila Taher, Peter Meinicke, and Burkhard Morgenstern. On splice site prediction using weight array models : a comparison of smoothing techniques. *Journal of Physics : Conference Series*, 90 :012004 (8pp), 2007.

[Tou07] Hélène Touzet. Comparative analysis of RNA genes : the caRNAc software. *Methods in Molecular Biology*, 395 :465–474, 2007.

[TP04] Hélène Touzet and Olivier Perriquet. CARNAC : folding families of related RNAs. *Nucleic Acids Research*, 32(Web Server issue) :W142–5, July 2004. doi:10.1093/nar/gkh415.

[TPP99] JD Thompson, F Plewniak, and O Poch. BAliBASE : a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, 15(1) :87–88, 1999. doi:10.1093/bioinformatics/15.1.87.

[TRG+03] Leila Taher, Oliver Rinner, Saurabh Garg, Alexander Sczyrba, Michael Brudno, Serafim Batzoglou, and Burkhard Morgenstern. AGenDA : homology-based gene prediction. *Bioinformatics*, 19(12) :1575–1577, 2003. doi:10.1093/bioinformatics/btg181.

[TSF88] Douglas H. Turner, Naoki Sugimoto, and Susan M. Freier. RNA structure prediction. *Annual Review of Biophysics and Biophysical Chemistry*, 17 :167–192, 1988.

[TTKA06] Yasuo Tabei, Koji Tsuda, Taishin Kin, and Kiyoshi Asai. SCARNA : fast and accurate structural alignment of RNA sequences by matching fixed-length stem fragments. *Bioinformatics*, 22(14) :1723–1729, July 2006. doi:10.1093/bioinformatics/btl177.

[TW06] Herbert H. Tsang and Kay C. Wiese. SARNA-Predict : A Simulated Annealing Algorithm for RNA Secondary Structure Prediction. In *Computational Intelligence and Bioinformatics and Computational Biology, 2006. CIBCB '06. 2006 IEEE Symposium on*, pages 1–10, 2006. doi:http://dx.doi.org/10.1109/CIBCB.2006.330973.

[TW07] Herbert H. Tsang and Kay C. Wiese. SARNA-Predict : A Study of RNA Secondary Structure Prediction Using Different Annealing Schedules. In *Computational Intelligence and Bioinformatics and Computational Biology, 2007. CIBCB '07. IEEE Symposium on*, pages 239–246, 2007.

[Usp37] James V. Uspensky. *Introduction to Mathematical Probability*, pages 23–24. NewYork :McGraw-Hill, 1937.

[Vit67] Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2) :260–269, April 1967.

[Vos06] Bjorn Voss. Structural analysis of aligned RNAs. *Nucleic Acids Research*, 34(19) :5471–5481, 2006. doi:10.1093/nar/gkl692.

[WBB+08] David L. Wheeler, Tanya Barrett, Dennis A. Benson, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M. Church, Michael DiCuccio, Ron Edgar, Scott Federhen, Michael Feolo, Lewis Y. Geer, Wolfgang Helmberg, Yuri Kapustin, Oleg Khovayko, David Landsman, David J. Lipman, Thomas L. Madden, Donna R. Maglott, Vadim Miller, James Ostell, Kim D. Pruitt, Gregory D. Schuler, Martin Shumway, Edwin Sequeira, Steven T. Sherry, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Roman L. Tatusov, Tatiana A. Tatusova, Lukas Wagner, and Eugene Yaschenko. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 36(Suppl 1) :D13–21, 2008. doi:10.1093/nar/gkm1000.

[WdBQ+06] Simon Whelan, Paul I. W. de Bakker, Emmanuel Quevillon, Nicolas Rodriguez, and Nick Goldman. PANDIT : an evolution-centric database of protein and associated nucleotide domains with inferred trees. *Nucleic Acids Research*, 34 :Database issue D327–331, 2006. doi:10.1093/nar/gkj087.

[WFHS99] Stefan Wuchty, Walter Fontana, Ivo L. Hofacker, and Peter Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2) :145–165, February 1999.

[WFM+04] Woj M. Wojtowicz, John J. Flanagan, S. Sean Millard, S. Lawrence Zipursky, and James C. Clemens. Alternative splicing of Drosophila Dscam generates axon guidance receptors that exhibit isoform-specific homophilic binding. *Cell*, 118(5) :619–633, September 2004. doi:10.1016/j.cell.2004.08.021.

[WGJMOG01] Thomas Wiehe, Steffi Gebauer-Jung, Thomas Mitchell-Olds, and Roderic Guigo. SGP-1 : Prediction and Validation of Homologous Genes Based on Sequence Alignments. *Genome Research*, 11(9) :1574–1583, 2001. doi:10.1101/gr.177401.

[WH04] Stefan Washietl and Ivo L. Hofacker. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *Journal of Molecular Biology*, 342(1) :19–30, September 2004. doi:10.1016/j.jmb.2004.07.018.

[WHN08] Andreas Wilm, Desmond G. Higgins, and Cédric Notredame. R-Coffee : a method for multiple alignment of non-coding RNA. *Nucleic Acids Research*, 36(9) :e52, May 2008. doi:10.1093/nar/gkn174.

[WHS05] Stefan Washietl, Ivo L. Hofacker, and Peter F. Stadler. Fast and reliable prediction of noncoding RNAs. *Proceedings of the National Academy of Sciences of the United States of America*, 102(7) :2454–2459, 2005.

[WMS06] Andreas Wilm, Indra Mainz, and Gerhard Steger. An enhanced RNA alignment benchmark for sequence alignment programs. *Algorithms for Molecular Biology*, 1 :19, 2006. doi:10.1186/1748-7188-1-19.

[WPK+07] Stefan Washietl, Jakob S. Pedersen, Jan O. Korbel, Claudia Stocsits, Andreas R. Gruber, Jorg Hackermuller, Jana Hertel, Manja Lindemeyer, Kristin Reiche, Andrea Tanzer, Catherine Ucla, Carine Wyss, Stylianos E. Antonarakis, France Denoeud, Julien Lagarde, Jorg Drenkow, Philipp Kapranov, Thomas R. Gingeras, Roderic Guigo, Michael Snyder, Mark B. Gerstein, Alexandre Reymond, Ivo L. Hofacker, and Peter F. Stadler. Structured RNAs in the ENCODE selected regions of the human genome. *Genome Research*, 17(6) :852–864, June 2007. doi:10.1101/gr.5650707.

[WPVdP04] Jan Wuyts, Guy Perriere, and Yves Van de Peer. The European ribosomal RNA database. *Nucleic Acids Research*, 32(Suppl 1) :D101–103, 2004. doi:10.1093/nar/gkh065.

[WR04] Zasha Weinberg and Walter L. Ruzzo. Exploiting conserved structure for faster annotation of non-coding RNAs without loss of accuracy. *Bioinformatics*, 20 Suppl 1 :i334–41, August 2004. doi:10.1093/bioinformatics/bth925.

[WR06] Zasha Weinberg and Walter L. Ruzzo. Sequence-based heuristics for faster annotation of non-coding RNA families. *Bioinformatics*, 22(1) :35–39, January 2006. doi:10.1093/bioinformatics/bti743.

[WRH+07] Sebastian Will, Kristin Reiche, Ivo L Hofacker, Peter F Stadler, and Rolf Backofen. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Computational Biology*, 3(4) :e65, April 2007. doi:10.1371/journal.pcbi.0030065.

[XMU94] Ying Xu, Richard J. Mural, and Edward C. Uberbacher. Constructing gene models from accurately predicted exons : an application of dynamic programming. *Computational Applications in Biosciences*, 10(6) :613–623, December 1994.

[XU97] Ying Xu and Edward C. Uberbacher. Automated gene identification in large-scale genomic sequences. *Journal of Computational Biology*, 4(3) :325–338, 1997.

[YLB01] Ru-Fang Yeh, Lee P. Lim, and Christopher B. Burge. Computational Inference of Homologous Gene Structures in the Human Genome. *Genome Research*, 11(5) :803–816, 2001. doi:10.1101/gr.175701.

[YLLL04] Xiaomin Ying, Hong Luo, Jingchu Luo, and Wuju Li. RDfolder : a web server for prediction of RNA secondary structure. *Nucleic Acids Research*, 32(Web Server issue) :W150–3, July 2004. doi:10.1093/nar/gkh445.

[YWR06] Zizhen Yao, Zasha Weinberg, and Walter L Ruzzo. CMfinder–a covariance model based RNA motif finding algorithm. *Bioinformatics*, 22(4) :445–452, February 2006. doi:10.1093/bioinformatics/btk008.

[ZGS08] Matthias Zytnicki, Christine Gaspin, and Thomas Schiex. DARN ! A Weighted Constraint Solver for RNA Motif Localization. *Constraints*, 13(1–2) :91–109, February 2008. doi:10.1007/s10601-007-9033-9.

[ZS81] Michael Zuker and Patrick Stiegler. Optimal computer folding of large RNA sequences using thermodynamic and auxiliary information RNA sequences using thermodynamic and auxiliary information. *Nucleic Acids Research*, 9(1) :133–148, 1981.

[ZS84] Michael Zuker and David Sankoff. RNA secondary structures and their prediction. *Bulletin of Mathematical Biology*, 46 :591–621, 1984.

[Zuk89] Michael Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 244 :48–52, 1989.

[Zyt07] Matthias Zytnicki. *Localisation d'ARN non-codants par réseaux de contraintes pondérées*. PhD thesis, Université de Toulouse III - Paul Sabatier, 2007.