

**UNIVERSITE DE STRASBOURG  
ECOLE DOCTORALE DES  
SCIENCES DE LA VIE ET DE LA SANTE**

**THESE**

*pour obtenir le grade de*

**DOCTEUR DE L'UNIVERSITE DE STRASBOURG**

**Discipline: Sciences du Vivant  
Spécialité: Bioinformatique**

*présente et soutenue publiquement par*

**José António ALMEIDA COSTA da CRUZ**

*Le 20 Septembre 2011*

*Titre:*

**Development of Evolutionary Models for  
Non-Coding RNAs**

**Directeur de Thèse: Prof. E. WESTHOF**

**Membres du jury:**

<b>Prof. Dino MORAS</b>	<b>Examineur</b>
<b>Prof. Alain DENISE</b>	<b>Rapporteur Externe</b>
<b>Dr Christine GASPIN</b>	<b>Rapporteur Externe</b>
<b>Prof. Rolf BACKOFEN</b>	<b>Membre Invité</b>
<b>Prof. Eric WESTHOF</b>	<b>Directeur de Thèse</b>

À Elisabete,  
por acreditar

# Acknowledgements

Thank You

To my dear Elisabete who trusted me more than myself.

To Vicente and Matilde for patiently sharing their father.

To my parents Manuela and Bernardino for always nurturing our hard work, critical spirit, free thought and joy of life.

To Rodrigo who crossed the fingers for me.

To Conceição and Joaquim for always being there.

To Eric Westhof for his guidance, deep insight and patience. It was a privilege to be your student!

To Jorge Carneiro for taking the risk and making this thesis possible.

To Pascal Auffinger for trying to teach me. . . I hope I learned something.

To Fabrice Jossinet for sharing his office and his good (caustic) humor.

To the present and past members of the team Aya Kitamura, Benoît Masquida, Bertrand Beckert, Jiro Kondo, Lisa Al-Shikhley, Mélanie Meyer, Rym Kachouri-Lafond, Valerie Fritsch and Yaser Hashem for their warm support.

To Jean-Luc Souciet, Veronique Leh-Louis, Laurence Despons, Bernard Dujon and all the members of Gènevures for being such an inspiring and nice community.

To my PDBC colleagues: Afonso Guerra, Assunção Senra, Bruno Martins, Hugo Fernandes, Hugo Martins, Paula Freire, Ricardo Pinho, Rodrigo Liberal, Sónia Martins and Susana Barbosa for making the first year of the PDBC such an enjoyable time and to Manuela Cordeiro always there to help!

To our dear friends Marco e Joana for unconditional friendship. Everything was so much easier thanks to you.

To Gorete and Zé Arlindo for the precious help.

To our team: Eugénia & Tó, Fátima & Zé, Guida & Raul, Guigui & Bruno, Marília & Carlos, Sandra & Zé Manel and Pe. Zé Cruz for your presence.

# Agradecimentos

Obrigado

À minha querida Elisabete que confia mais em mim do que eu próprio.

Ao Vicente e à Matilde por partilharem, pacientemente, o seu pai.

Aos meus pais Manuela e Bernardino por alimentarem o sentido do trabalho, espírito crítico, o livre pensamento e a alegria de viver.

Ao Rodrigo que torçe sempre por mim.

À Conceição e ao Joaquim por estarem sempre lá quando é preciso.

A Eric Westhof pela orientação, *deep insight* e paciência. Foi um privilégio ser seu estudante!

A Jorge Carneiro por correr o risco e tornar esta tese possível.

Ao Pascal Auffinger por me tentar ensinar, espero ter aprendido.

Ao Fabrice Jossinet por partilhar o seu espaço e o seu bom humor.

Aos membros, passados e presentes, da equipa: Aya Kitamura, Benoît Masquida, Bertrand Beckert, Jiro Kondo, Lisa Al-Shikhley, Mélanie Meyer, Rym Kachouri-Lafond, Valerie Fritsch and Yaser Hashem pelo vosso caloroso apoio.

A Jean-Luc Souciet, Veronique Leh-Louis, Laurence Despons, Bernard Dujon e todos os membros do Gènevures por serem um grupo tão inspirador e simpático.

Aos colegas do PDBC: Afonso Guerra, Assunção Senra, Bruno Martins, Hugo Fernandes, Hugo Martins, Paula Freire, Ricardo Pinho, Rodrigo Liberal, Sónia Martins and Susana Barbosa por fazerem do primeiro ano do PDBC um momento tão agradável e à Manuela Cordeiro sempre pronta a ajudar!

Aos nosso queridos amigos Joana & Marco pela amizade incondicional. Tudo foi tão mais fácil graças a vocês.

À Gorete and ao Zé Arlindo pela ajuda preciosa.

À nossa equipa: Eugénia & Tó, Fátima & Zé, Guida & Raul, Guigui & Bruno, Marília & Carlos, Sandra & Zé Manel a ao Pe. Zé Cruz pela presença constante.

## Abstract

In this thesis I applied bioinformatics techniques to three open problems on ncRNA studies: (i) how to compare meaningfully three-dimensional RNA models; (ii) how to automate the annotation of ncRNA in eukaryotic genomes in an accurate fashion and (iii) how to detect structural ncRNA modules using sequence information alone.

I developed two new structural comparison metrics that take into account the structural specificities of structured ncRNAs: The Deformation Index (*DI*) and the the Deformation Profile (*DP*). The *DP* enriches the Root Mean Square Deviation (*RMSD*) with base pair prediction accuracy measurements. The *DP* provides multi-scale information on the differences between target and reference models at local, intra and inter-domain scales. These metrics can be used to evaluate the quality of predicted ncRNA models and can help to improve structure prediction tools. Following this work on structural comparison the first ncRNA structure prediction assessment experiment was developed: **RNAPuzzles**. Three first rounds of evaluation with the participation of seven research groups representative of the RNA structure prediction community were performed.

To answer the need for a fast and reliable ncRNA annotation in the context of large scale genome sequencing projects (Génolevures and Dikaryome projects), I implemented two automatic annotation pipelines, integrating publicly available tools, for homology and *de novo* ncRNA search in genomes. Both pipelines were applied to 15 yeast genomes and 1051 ncRNA genes were found, corresponding to more than 80% of the expected ncRNAs (assuming the number of ncRNAs from *S. cerevisiae* (86) as reference). Additionally I identified: (i) several new potential ncRNAs; (ii) several new synteny relationships between ncRNA loci; and (iii) new examples of extended structural domains in well known essential ncRNAs. These results show the feasibility of automatic search for ncRNAs in full genomes and the utility of such approaches in large genome annotation projects.

Finally, I developed a new algorithm to detect structural RNA modules in sequences: **RMDetect**. It was designed to identify 3D structural modules in RNA sequences. It uses a Bayesian Network to represent the searched modules and the joint base pair probability estimation to select candidates. Four modules can be searched for: G-bulges, Kink-turns, C-loop and Tandem-GAs. In test sequences all of the known modules were found with a false discovery rate of 0.23. In 1444 publicly available alignments 21 yet unreported and 141 known modules were identified. **RMDetect** is a step to bridge the gap between sequence analysis and 3D RNA studies. It can be used in the refinement of RNA 2D structures, the assembly of RNA 3D models, and the search of structured ncRNAs in genomic data.

# Résumé étendu en Français

## Introduction

Les acides ribonucléiques (ARN) sont des biopolymères essentiels présents dans toutes les formes de vie. Jusqu'à très récemment, on pensait que les ARN étaient presque exclusivement liés à la synthèse des protéines, puisque, d'un côté, l'ARN messager, l'ARN de transfert et l'ARN ribosomal jouent un rôle essentiel dans l'expression génétique, et que les ARN nucléaires sont essentiels pour l'épissage des gènes eucaryotes. Ce point de vue restreint de l'ARN, en tant que transporteur passif de l'information génétique, a commencé à changer avec la découverte de ses propriétés catalytiques et de son rôle de régulateur génétique par le biais de l'interférence génétique. Le fait que les molécules catalytiques et régulatrices de l'ARN ne codent pas de protéines – contrairement à l'ARN messager – a conduit à la notion d'“ARN non codant” (ARNnc) pour désigner de manière générique cette classe de molécules.

L'ARNnc participe à de nombreuses fonctions cellulaires aussi diverses que la synthèse des protéines, l'épissage des gènes, l'élongation des télomères, la régulation de l'expression génique (riboswitches et miARN) ou l'inactivation des gènes (inactivation des chromosomes), parmi beaucoup d'autres. Bien que la proportion précise des transcrits fonctionnels dans la cellule soit encore une question ouverte, il a été montré qu'au moins 93% du génome humain (et un pourcentage similaire chez la souris et d'autres eucaryotes) est transcrit. Comme moins de 2% de celui-ci correspond à de la séquence codante, le nombre potentiel d'ARNnc fonctionnel est vaste, même si l'on considère une partie de ces transcrits comme le résultat du bruit de transcription. Trouver des gènes d'ARNnc dans des génomes représente un défi en soi. Les régions non codantes sont souvent mal conservées au niveau de la séquence, à l'inverse des régions codantes qui présentent pour leur part une conservation de séquence importante; des substitutions synonymes de codons et biais de codons, en raison de leur potentiel de codage de protéines. L'annotation des gènes d'ARNnc exige d'importants efforts humains dans les phases de recherche et de validation du processus d'annotation. Ainsi, des outils pour accélérer et simplifier l'effort d'annotation sont indispensables, principalement pour les projets de séquençage génomique à grande

échelle.

Les molécules d'ARNnc structurées présentent une structure tridimensionnelle formée par des hélices double brin et des régions en simple brin qui se combinent en une architecture complexe par des interactions à distance et des empilements d'hélices. La structure tertiaire fait toujours intervenir des appariements non-Watson-Crick entre des nucléotides non contigus de la séquence. Les axes de rotation dans les liaisons covalentes du squelette sucre-phosphate donnent également une grande flexibilité lors du repliement. Toutes les combinaisons valides de rotations et de paires de base, pour une séquence donnée peuvent générer un nombre incalculable de structures tridimensionnelles potentielles. Dans la quantité de structures géométriquement possibles, comment détecter celles qui présentent des fonctions biologiques observables? La prédiction de la structure tridimensionnelle d'ARN est par conséquent un domaine de recherche en expansion dans lequel plusieurs outils prometteurs, récemment publiés, visent à produire des modèles tridimensionnels d'ARN à partir d'information de séquences et ceci de façon plus ou moins automatique. Aujourd'hui, aucun modèle ou outil de prédiction n'est capable de produire des modèles d'ARN suffisamment précis par rapport aux structures natives. Ainsi plusieurs questions sont ouvertes: Qu'est-ce qu'une prédiction de structure biologiquement significative? Comment évaluer un modèle prédit? Comment savoir si un outil ou une approche de prédiction est efficace de façon consistante? Ces outils produisent-ils des échantillonnages de l'espace de tous les repliements significatifs? Dans quels scénarios (type de molécule, taille, complexité, ...), réussissent ou échouent-ils? Une façon pratique de répondre à ces questions est de comparer systématiquement les modèles prédits avec des structures à la résolution atomique obtenus par cristallographie rayons X et d'étudier les similitudes et les différences observées. Avoir des méthodes et des outils appropriés pour évaluer les modèles prédits et comparer les outils de prédiction est un besoin pressant de ce champ de recherche.

L'étude des structures d'ARNnc révèle de nombreuses sous structures récurrentes, nommés "modules structuraux", avec des fonctions structurales très spécifiques comme la boucle en épingle à cheveux, les tétraboucles, la boucle "G-bulged", les motifs en A-mineur ou le "Kink-Turn". L'identification des modules structuraux donne des indices importants pour découvrir un ordre derrière les très complexes structures tridimensionnelles d'ARN.

Dans la présente thèse, nous avons appliqué l'approche informatique et bioinformatique à trois questions ouvertes dans l'étude de l'ARNnc: (i) comment comparer des modèles tridimensionnels d'ARN, (ii) comment automatiser l'annotation de gènes d'ARNnc dans les génomes eucaryotes de façon aussi précise que possible, (iii) comment détecter des modules structuraux d'ARN en utilisant uniquement des informations de séquence.

## Comparaison des modèles tridimensionnels d'ARN

La métrique la plus couramment utilisée, pour comparer des structures tridimensionnelles de molécules, tant pour les protéines que pour les ARNnc, est l'écart quadratique moyen (RMDS). Bien que le RMSD soit simple à formuler et à calculer, il lui manque une interprétation fonctionnelle claire. Pour fournir une mesure significative de similarité des structures, la métrique de comparaison structurale devrait prendre en compte la nature des molécules comparées et leurs caractéristiques structurales les plus pertinentes. Elle devrait également fournir des indications sur les caractéristiques qui contribuent ou qui pénalisent la valeur de similarité. Beaucoup de mesures de comparaison structurale ont été proposées pour les structures des protéines. Malheureusement, les outils spécifiquement développés pour la comparaison des protéines ne s'adaptent pas à comparaison de l'ARN et la quantité de travail développé par la communauté de l'ARN est nettement plus faible. Une motivation supplémentaire pour rechercher des bonnes métriques de comparaison est que les outils automatiques de prédiction de structures ont tendance à produire des centaines voire des milliers de modèles qui sont impossibles à analyser manuellement un par un, nécessitant ainsi des méthodes de comparaison automatique.

Nous avons développé deux nouveaux indicateurs de comparaison structurale qui prennent en compte les spécificités structurales des molécules d'ARN: l'indice de déformation (ID) et le profil de déformation (PD). L'ID enrichit le RMSD avec des mesures de précision de la prédiction de paires de bases. Le PD vise à fournir des informations multi-échelles sur les différences entre la cible et les modèles prédits au niveau local, intra-domaine et inter-domaine. Ces métriques peuvent être utilisées pour évaluer les modèles prédits d'ARN contre les structures d'ARN observées et aider la communauté de prédiction de structures à évaluer la qualité de leurs modèles et améliorer ses outils. Suite à notre travail sur la comparaison structurale, nous avons développé la première expérience d'évaluation de prédiction de structures d'ARN: **RNAPuzzles**. Nous avons effectué les trois premières rondes de l'évaluation de prévisions avec la participation de sept groupes de recherche représentatifs de la communauté de prédiction de structures d'ARN. Une poursuite immédiate de ces travaux de comparaison structurale serait d'analyser comment les mesures ID et PD se comportent dans l'espace conformationnel de l'ARN. Cette ligne de travail bénéficierait grandement de nouveaux modèles d'échantillonnage aléatoire de structure d'ARN qui serait une contribution essentielle à la compréhension théorique de l'espace conformationnel de l'ARN. Après **RNAPuzzles**, des nouveaux défis de prédiction de structures devraient être publiés dans un futur proche. L'effort de traitement et de comparaison de structures sera considérablement simplifié dans ces prochaines étapes grâce aux développements déjà accomplis. L'amélioration de la mécanique de fonctionnement de **RNAPuzzles**, à savoir avec des outils



d'évaluation meilleurs et plus complets, devrait aussi faire partie de l'évolution future.

## Annotation d'ARNnc

Pour répondre à la nécessité d'un pipeline d'annotation d'ARNnc rapide et fiable dans le contexte des projets de séquençage génomique de grand envergure, tels que les projets Génolevures et Dikaryome, nous avons mis au point deux pipelines d'annotation automatique, intégrant des outils disponibles publiquement, de recherche d'ARNnc par homologie et *de novo*. Les deux pipelines ont été appliqués à 15 génomes de levures et ont permis de trouver et d'annoter 1051 gènes d'ARNnc, correspondant à plus de 80% des ARNnc attendus pour ces génomes si on prend comme référence le nombre d'ARNnc chez *S. cerevisiae* (86). En outre, plusieurs nouveaux ARNnc, encore inconnus chez les *Saccharomycotinae*, ont été détectés. De plus, nous avons mis en évidence un ensemble de nouvelles observations sur la synténie de gènes d'ARNnc et de nouveaux exemples de domaines supplémentaires dans certains ARNnc essentiels. Les résultats montrent la faisabilité de la recherche automatique des ARNnc dans les génomes complets et l'utilité de telles approches dans les grands projets de séquençage et d'annotation génomique. L'intégration complète, dans le pipeline de développement, de nouveaux outils tels que ceux de prédiction de gènes d'ARNnc *de novo* ainsi que la possibilité de traiter des données provenant d'autres sources, comme les expériences de séquençage profond, sont les prochains défis à court terme dans cette ligne de travail. La confirmation expérimentale de ces observations, qui est au-delà de l'approche bioinformatique, doit être le prolongement naturel du projet d'annotation. Dans le strict domaine bioinformatique, le développement de nouvelles approches pour détecter les gènes d'ARNnc insaisissables tels que la composante ARN de la télomerase seraient des ajouts utiles à notre pipeline.

## Détection de modules structuraux d'ARN

Enfin, j'ai développé un algorithme original pour détecter les modules structuraux d'ARN uniquement à partir des informations de séquence (RMDelect). L'algorithme a été conçu pour identifier les modules structuraux connus dans les séquences simples et dans les alignements multiples en l'absence de toute autre information. L'algorithme utilise un réseau bayésien pour la représentation des modules couplé à l'estimation de la probabilité conjointe des paires de bases Watson-Crick participant à des modules. Actuellement, quatre modules peuvent être recherchés: la boucle "G-bulge", le "Kink Turn", la boucle C et la boucle "tandem GA". Dans des séquences de test de contrôle, nous avons trouvé l'ensemble des modules connus avec un taux

de fausse découverte de 0.23. En cherchant les 1444 alignements publiquement disponibles, nous avons identifié 21 modules encore non détectés et 141 modules connus. RMDetect est une étape utile pour combler le fossé entre l'analyse pure de séquences et l'étude structurale de l'ARN. De plus, il peut être utilisé dans l'affinement des structures 2D d'ARN, dans l'assemblage de modèles 3D, et dans la recherche et l'annotation de gènes d'ARN structurés dans les génomes. Nous espérons améliorer l'approche actuelle par l'ajout de nouveaux modèles structuraux. La recherche de modules structuraux dans des génomes complets serait la prochaine étape dans cette ligne de recherche.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	RNA Structure . . . . .	2
1.2	The RNA Folding Problem . . . . .	4
1.3	Non Coding RNAs . . . . .	7
1.4	ncRNA Evolution . . . . .	7
1.5	Article – The Dynamic Landscapes of RNA Architecture . . .	12
<b>2</b>	<b>RNA Structure Comparison</b>	<b>19</b>
2.1	Structure Comparison . . . . .	20
2.1.1	Root Mean Square Deviation . . . . .	21
2.2	Deformation Index . . . . .	23
2.3	Deformation Profile . . . . .	26
2.4	Article – New metrics for comparing and assessing discrepancies between RNA 3D structures and models . . . . .	34
<b>3</b>	<b>RNA Puzzles</b>	<b>46</b>
3.1	RNA Tertiary Structure Prediction . . . . .	47
3.2	RNA Puzzles . . . . .	49
3.2.1	Structure Comparison Pipeline . . . . .	51
3.2.2	RNA Puzzles Web Site . . . . .	51
3.3	Conclusions . . . . .	54
3.4	Article – A CASP-like Evaluation of <i>de novo</i> structure predictions . . . . .	56
<b>4</b>	<b>Annotation of ncRNA genes</b>	<b>95</b>
4.1	Introduction . . . . .	96
4.2	The Pipeline . . . . .	97
4.2.1	Terminology . . . . .	97
4.2.2	Data Collection . . . . .	99
4.2.3	Homology search . . . . .	100
4.2.4	<i>de novo</i> Search . . . . .	101
4.2.5	Hit processing and candidate generation . . . . .	103
4.2.6	Automatic Candidate Validation . . . . .	105

4.3	Manual candidate validation . . . . .	107
4.4	Technical implementation . . . . .	108
<b>5</b>	<b>Annotation of ncRNAs in Budding Yeasts</b>	<b>112</b>
5.1	Budding yeasts . . . . .	113
5.2	The Annotated Genomes . . . . .	114
5.2.1	Data sources . . . . .	114
5.3	<i>Saccharomyces cerevisiae</i> – Reference Genome . . . . .	114
5.4	The annotation process . . . . .	116
5.4.1	Phase 1: Génolevures 2, Génolevures 3 and <i>E. gossypii</i>	116
5.4.2	Phase 2: Nakaseomycetes . . . . .	121
5.5	Some Noticeable Annotation Results . . . . .	121
5.5.1	The telomerase ncRNA . . . . .	121
5.5.2	The synteny between RNase MRP and snR66 . . . . .	122
5.5.3	The TPP riboswitch . . . . .	123
5.5.4	A snoRNA candidate in <i>Yarrowia lipolytica</i> . . . . .	126
5.5.5	<i>de novo</i> Candidates . . . . .	126
5.6	The extremely large ncRNAs in the Nakaseomycetes family .	126
5.7	Conclusions . . . . .	129
5.8	Article – Identification and Annotation of Non-Coding RNAs in Saccharomycotina . . . . .	134
<b>6</b>	<b>Detection of Structural Modules</b>	<b>143</b>
6.1	Introduction to Structural Modules . . . . .	143
6.2	Search for Structural Modules . . . . .	145
6.2.1	Interaction Networks . . . . .	145
6.2.2	Position Weight Matrices . . . . .	146
6.3	RMDetect Approach . . . . .	148
6.3.1	Modeling with Bayesian Networks . . . . .	148
6.3.2	Adding Base Pair Probability Information . . . . .	151
6.3.3	Adding Alignment Information . . . . .	151
6.3.4	Automatic Construction of Models . . . . .	151
6.4	Search Algorithms . . . . .	155
6.4.1	Single Sequence Search Algorithm . . . . .	155
6.4.2	Multiple Sequence Clustering Algorithm . . . . .	155
6.4.3	Candidate Evaluation . . . . .	157
6.5	Results . . . . .	157
6.6	Conclusions . . . . .	157
6.7	Article – Sequence-based Identification of 3D Structural Mod- ules in RNA with RMDetect . . . . .	160
<b>7</b>	<b>Conclusions and Perspectives</b>	<b>170</b>

<b>A Useful Concepts</b>	<b>173</b>
A.1 Score . . . . .	173
A.2 E-value . . . . .	173
A.3 Quantity of Information . . . . .	174
A.4 Entropy . . . . .	174
A.5 Mutual Information . . . . .	175
A.6 Radius of gyration . . . . .	175
A.7 Receiver Operating Characteristic . . . . .	176
A.8 Z-Score . . . . .	179
<b>B Base Pairs</b>	<b>180</b>
B.1 Isostericity . . . . .	180

# List of Tables

1.1	List of ncRNA families . . . . .	8
2.1	<i>RMSD</i> , <i>INF</i> and <i>DI</i> values for three predicted models of rat 28S E-loop . . . . .	26
2.2	Deformation Profile ( <i>DP</i> ) values and ratios for intra and in- terdomain regions. . . . .	33
3.1	RNA structure prediction tools. . . . .	55
4.1	ncRNA homology search tools. . . . .	101
4.2	Low homology example . . . . .	102
5.1	List of budding yeast complete genomes . . . . .	113
5.2	ncRNAs in <i>S. cerevisiae</i> and Génolevures genomes . . . . .	117
5.3	<i>de novo</i> search results. . . . .	120
5.4	Phase 1 ncRNA annotations . . . . .	120
5.5	Phase 2 ncRNA annotations (Nakaseomycetes genomes) . . .	121
5.6	Telomerase (TER) vs. snR161 syntenic region . . . . .	123
5.7	RNase MRP vs. snR66 syntenic region . . . . .	124
6.1	Results of RMDetect on public database alignments. . . . .	158
A.1	Example of a list of gene candidates. . . . .	176
A.2	List of gene candidates ordered by <i>Score1</i> . . . . .	177
A.3	List of gene candidates ordered by <i>Score2</i> . . . . .	179

# List of Figures

1.1	RNA structural hierarchy . . . . .	5
1.2	Some major roles of ncRNAs . . . . .	9
1.3	ncRNAs evolutionary constraints. . . . .	10
1.4	Telomerase ncRNA conservation. . . . .	11
2.1	Base-base interactions proportion . . . . .	24
2.2	Distribution of <i>RMSD</i> vs. <i>INF</i> values . . . . .	26
2.3	Native and predicted models of rat 28S E-loop. . . . .	27
2.4	Building steps of the Deformation Profile . . . . .	29
2.5	Native hammerhead ribozyme and predicted Models . . . . .	30
2.6	Deformation profile of model <i>S1</i> . . . . .	31
2.7	Deformation profile of model <i>S2</i> . . . . .	32
3.1	Number of published X-Ray structure of RNA molecules. . . . .	50
3.2	RNAPuzzles processing pipeline. . . . .	52
3.3	RNAPuzzles web site. . . . .	53
4.1	The main components of the annotation pipeline . . . . .	98
4.2	Hit cluster of the U1 snRNA of <i>C. nivariensis</i> candidate . . . . .	106
4.3	Hit location map . . . . .	108
4.4	Multiple sequence alignment example . . . . .	109
4.5	Anootation pipeline - Full description . . . . .	110
5.1	Annotated genomes . . . . .	115
5.2	ROC analysis of E-value for <i>INFERNAL</i> and <i>BLAST</i> . . . . .	118
5.3	TPP Riboswitch . . . . .	125
5.4	<i>Y. lipolytica</i> snoRNA candidate . . . . .	127
5.5	Examples of three <i>de novo</i> candidates . . . . .	128
5.6	RNase P extensions . . . . .	130
5.7	U1 snRNA extensions . . . . .	131
5.8	Lengths of the yeast specific extensions. . . . .	132
6.1	Interaction network of a kink-turn . . . . .	146
6.2	Consensus interaction network of the kink-turn module . . . . .	147

6.3	Kink-turn search using a simple position independent model .	149
6.4	Simple Bayesian Network . . . . .	150
6.5	Bayesian Networks for the four studied modules . . . . .	152
6.6	Predicted modules in the Lysine riboswitch secondary structure	153
6.7	Module clustering on a lysine riboswitch alignment . . . . .	154
A.1	Example of a ROC curve. . . . .	178
B.1	The four bases and their respective edges. . . . .	181
B.2	Leontis-Westhof classification of the twelve base pair types. .	182
B.3	Examples of a <i>cis</i> and <i>trans</i> base pairs. . . . .	183
B.4	Examples of four isosteric base pairs. . . . .	184



# Chapter 1

## Introduction

Ribonucleic acids (RNAs) are essential biopolymers present in all forms of life. Until very recently RNAs were thought to be almost exclusively related with protein synthesis. The messenger, transfer and ribosomal RNAs participate in the core pathways of gene expression and the small nuclear RNAs in the splicing pathway of eukaryotes. This restricted view of RNAs as passive carriers of genetic information started to change less than 30 years ago with the discovery of the catalytic property of RNAs by Cech and Altman (Kruger et al., 1982; Guerrier-Takada et al., 1983). It became clear that RNA was not only a simple intermediate between DNA and proteins. In fact, RNA shares both characteristics: like DNA, RNA is able to code genetic information and, like proteins, RNA is able to catalyze chemical reactions. This double role prompted the “RNA World” theory proposing that autonomous RNA molecules would have been the precursors of the DNA and proteins in ancient life forms (Gilbert, 1986). A not less surprising finding was the gene regulation role of RNA through genetic interference (Fire et al., 1998). The fact that small RNA sequences could silence genes by hybridizing with complementary regions of the mRNA genes was a major perspective change and paved the way for the discovery of a plethora of RNA mediated regulatory mechanisms (Waters and Storz, 2009; Carthew and Sontheimer, 2009).

The fact that the newly discovered catalytic and regulatory RNA molecules do not code for proteins, contrary to the more abundant and well known mRNAs, led to the term of “non coding RNAs” (ncRNA) as a general designation of this class of molecules and this term will be used to generally refer to this class of molecules here.

Our knowledge of the number and variety of RNA roles in the cell still continues to grow (Ponting et al., 2009), as well as the number of long RNAs that are found to be transcribed in a tightly regulated fashion but for which no function is known (Jung et al., 2010). In reality the total number of ncRNA genes in either eukaryote or prokaryote genomes is still a rough

estimate and the discovery of new ncRNA genes as well as of known ncRNAs in newly sequenced genomes is an active research topic.

Many of the diverse biological functions performed by ncRNAs depend on its intricate three-dimensional shape and determining the precise structure of ncRNAs is an important research subject.

It is important to notice that although some ncRNA molecules correspond to self contained structured molecules – like the ribosome, the RNase P, the tRNAs, . . . – many other ncRNAs correspond to parts of larger transcripts and, in some cases, like the riboswitches, co-occur with protein coding regions in the same transcript.

The work presented in this thesis is at the intersection of the two research topics described above: ncRNA gene discovery and ncRNA structure determination.

This first chapter presents a short introduction on ncRNA functions and structure emphasizing the ncRNA families more relevant to this thesis. The following chapters describe, each, a self-contained part of the developed work. The development of a set of metrics for three-dimensional RNA structures comparison (Chapter 2) and the application of these metrics in a newly developed RNA structure prediction benchmark: `RNAPuzzles` (Chapter 3). The implementation of an annotation pipeline for ncRNA genes in yeast (Chapter 4) and the application of this pipeline to the annotation of yeast genomes from the Génolevure’s consortium (Chapter 5). Finally, in a first attempt to close the gap between pure sequence analysis and structural RNA studies, I developed an algorithm for structural modules detection based on sequence information (Chapter 6). Finally, Chapter 7 presents a general conclusion and some future perspectives for future work.

## 1.1 RNA Structure

RNAs consist of long chains made of four nucleotides (adenine, cytosine, guanine and uracil) joined along a sugar phosphate backbone formed by covalent bonds between the O3’ atom of the ribose moiety and the P atom of the phosphate group of the subsequent nucleotide. RNA is chemically very close to deoxyribonucleic acid (DNA). The differences between RNA and DNA are: The additional oxygen atom O2’ on the sugar group (ribose instead of deoxyribose) and the presence of the uracil instead of the thymine nucleotide (Saenger, 1984). Figure 1.1 illustrates some of the concepts discussed in this section.

The chemical differences between RNA and DNA are accompanied by large differences in properties, structure and function. Contrary to DNA which is present in the cell nucleus<sup>1</sup> as a, chemically stable, long (from a few thousands to many millions of base pairs) double stranded helix, RNA

---

<sup>1</sup>In eukaryotic cells.

is present in all cellular compartments in the form of short or very long (from tens to several thousands of bases) single stranded molecule that can fold on itself forming shorter helical regions<sup>2</sup> organized in intricate three-dimensional architectures.

Each of the four nucleobases that constitute a RNA chain consists of one (for the pyrimidines, cytosine and uracil) or two (for the purines, adenine and guanine) planar cyclic rings their three edges exposed for base-base interaction through hydrogen bonds: the Watson-Crick edge, the Sugar edge and the Hoogsteen edge.

RNA helices are formed by stacks of G=C, A=U and (wobble) G≠U base pairs commonly known as Watson-Crick (WC) or canonical base pairs as they involve the WC edges of the intervening bases. Base pairs involving other combinations of edges are also common and are generically known as non Watson-Crick (non WC) or non canonical pairs (see Appendix B). Long range base-base interactions through non WC base pairs play key roles in RNA architecture as they help to stabilize the three-dimensional RNA fold.

On the RNA backbone, two contiguous phosphate groups are connected through six covalent bonds between sugar atoms (5' -P-O5'-C5'-C4'-C3'-O3'-P- 3'). Each of these covalent bonds establishes a torsion angle pivot which allows the full rotation of the connected atoms (except for the C4'-C3' bond which allows only a partial rotation that changes the puckering of the sugar ring). The freedom of rotation of these six bonds confers a great flexibility to the backbone and contributes to the huge conformational space of possible RNA structures. This conformational space, however, is constrained by some well known phenomena:

- **Backbone conformers:** It has been observed that backbone torsion angles tend to occupy some discrete positions of the potentially infinite angle space (Yathindra and Sundaralingam, 1973). Recently, Richardson and co-authors (Richardson et al., 2008) published a set of 46 torsion angle combinations (conformers) as the most commonly observed torsion angle values clusters.
- **Hydrogen bonds:** Hydrogen bonds between nucleotide groups are responsible for base pairing. The length and directionality of H-bonds are one of the main constraints for RNA base pairing.
- **Base stacking:** Weaker, but much more numerous, than H-bonds, Van der Waals interactions between the aromatic rings of the nucleotides contribute to the stability of helices and are often observed in extra helical stacking of either contiguous nucleotides (as in single stranded regions) or long range staking interactions.

---

<sup>2</sup>The geometry of the RNA and DNA helices is, generally, not the same. Naturally occurring DNA normally presents the B-DNA geometry, while naturally occurring RNA helices present the A-DNA geometry (Saenger, 1984).

The RNA structure description is usually systematized according to an hierarchical model that sets up three structural levels:

- **Primary structure:** The linear sequence of nucleotides as they are transcribed from the template DNA.
- **Secondary structure:** The set of Watson-Crick base pairs forming helical regions connected by single stranded regions.
- **Tertiary structure:** The three-dimensional structure of the RNA molecule, which comprises the set of non Watson-Crick base pairs, the co-axial stacking of helices, and the architecture of the N-way helical junctions.

In this model each level describes a specific degree of structural complexity and reorganizes the features at the level immediately below at a higher dimension – e.g. the secondary structure organizes the complementary sequences described in the primary structure in helical regions and the tertiary structure organizes the relative position of helices in space.

The conceptual separation between secondary and tertiary structures is supported by a number of observations: The secondary structure prediction is a more tractable problem than tertiary structure prediction; Covariation analysis of sequence alignments allows the prediction of secondary structure elements; The data provided by experimental techniques, such as chemical probing or UV-melting, can be interpreted as stemming either from secondary structure or tertiary structure depending on the conditions (presence or absence of divalent ions, temperature, ...).

## 1.2 The RNA Folding Problem

Most of the RNA structure prediction tools (with few exceptions) (Xayaphoumine et al., 2005) deal exclusively with the “final” or “native” conformation of the RNA molecule. They assume that the “native” conformation represents a thermodynamic equilibrium of the molecule corresponding to the most energetically favorable folding state (McCaskill, 1990).

Currently, the most reliable way of determining the structure of biomolecules at atomic scale is through X-ray crystallography. Although invaluable for the information it conveys and the insight it provides about the molecular function, a crystallographic structure of an RNA molecule remains a static snapshot of a dynamic reality and provides little information on how the newly transcribed, linear chain of nucleotides folds into a fully functional 3D architecture.

The kinetic perspective of the RNA folding assumes that a linear RNA transcript will follow several pathways on the conformational space but also

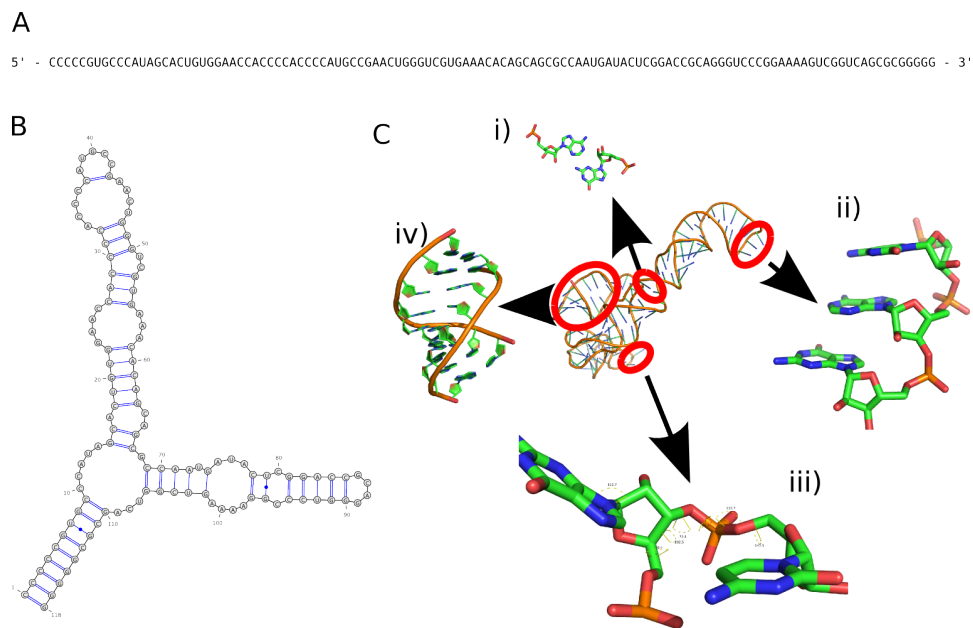


Figure 1.1: RNA structural hierarchy. (A) Primary structure: sequence. (B) Secondary structure: helical and single stranded regions. (C) Tertiary structure: three-dimensional coordinates of atoms (PDB: 3JQ4). In crystallographic model are depicted some structural components: (i) A non WC base pair (A $\circ$ G sugar-sugar in *trans*); (ii) Three stacked bases; (iii) The backbone torsion angles and (iv) An helical region.

that the final state depends on the speed of transcription compared with the folding rate. The thermodynamic perspective assumes that the more energetically favorable conformations will be always proportionally more frequently occupied. At present there is no complete model of RNA folding which could predict either the folding pathways or the equilibrium conformations, but it is well known that the native conformation observed through crystallography is often one among many possible conformations and that many factors affect the folding of an RNA model such as:

- **Ions:** The electrostatic environment induced by positively charged ions plays a major role in RNA folding by counteracting the repulsive force of the negatively charged backbone. Monovalent ions only are required to stabilize the secondary structure while divalent ions, like magnesium, are necessary for stabilizing the three-dimensional structure.
- **Co-transcriptional folding:** RNA folds at a higher rate than it is transcribed and the emerging strand of RNA can start immediately to establish base pair interactions. Thus, all factors affecting transcription order, speed, pauses... will have a potential effect on the final structure (Wong et al., 2007; Nechooshtan et al., 2009).
- **Induced fit:** In several cases of RNA molecules, with catalytic activity, the active conformation is achieved by structural adjustments induced by conformational changes in domains distant from the active site (Martick and Scott, 2006; Toor et al., 2008)
- **Ligands:** The presence of ligands can drastically change the final outcome of the folding. Riboswitches are good examples of alternative conformations arising depending on the presence of small metabolites (Breaker, 2008).
- **Proteins and chaperones:** The presence of interacting proteins during the folding modulates the folding process. For example, in the ribosome the RNA bases that firstly interact with proteins are those involved in three way junctions organization (Adilakshmi et al., 2008). Additionally, several RNA chaperones are known to unfold and target to degradation misfolded RNAs.
- **Quasi hierarchical folding:** Most of the folding models assume a strict hierarchical folding in which secondary structure folds first and tertiary interactions form after that to organize the secondary structures in the final three-dimensional shape (Tinoco and Bustamante, 1999). New data suggests that tertiary interactions participate early

in the folding process and guides the folding process along the conformational space (Greenleaf et al., 2008; Noeske et al., 2007; Chauhan and Woodson, 2008).

In (Cruz and Westhof, 2009) we present an overview of the many factors affecting RNA folding and the current theoretical models and experimental techniques used to approach the problem. This review can be found at section 1.5.

### 1.3 Non Coding RNAs

As referred in the introduction, ncRNAs are generically defined as RNA elements that do not code for proteins. This apophatic designation of ncRNA, naming it not for what it is, but for what it is not, is a disturbing choice and, in this case, it is also reductionist – ncRNAs do not constitute an homogeneous group of molecules but a rather diverse set of families of different functions, sizes and structures – and somehow misleading – certain ncRNAs are not complete molecules by themselves but are parts of coding RNA molecules (mRNAs) outside the coding sequence (5' and 3' UTRs) usually with regulatory functions such as the riboswitches (Breaker, 2008). In this work when I refer to ncRNAs I mean all RNA molecules or part of molecules that do not code for proteins and which function depends on its three-dimensional structure.

The big number of annotated ncRNA families (Gardner et al., 2009) and their very diverse functions make it difficult to present a complete picture of the role of ncRNAs in the cell. In this section we enumerate, from the currently known ncRNAs families, those with particular importance for our work. Figure 1.2 represents the classic “Central Dogma of Molecular Biology” with the respective ncRNA families roles annotated and Table 1.1 lists these families with a short description.

### 1.4 ncRNA Evolution

Part of the difficulty in the bioinformatics study of ncRNAs is due to the particular evolutionary constraints of ncRNA molecules and their effects on ncRNA sequences. Non coding structural RNAs are subject to structural constraints characterized by:

- Ability to sustain compensatory mutations in helical regions by the permutation between AU, CG, GC, GU, UA and UG base pairs.

Name <sup>1</sup>		Occur. <sup>2</sup>	Function <sup>3</sup>
rRNA	Ribosomal RNA	all	Protein synthesis in all cellular life forms.
tRNA	Transfer RNA	all	Transport of amino acids to protein synthesis in the ribosome.
RNase P	Ribonuclease P	all	tRNA maturation by cleaving the 5' leader sequence of the pre tRNA.
RNase MRP	Ribonuclease mitochondrial RNA processing	E	rRNA maturation by cleaving the pre rRNA. Initiation of mitochondrial DNA replication.
snRNA	Small nuclear RNA	E	Participate in the spliceosome complex.
snoRNA	Small nucleolar RNA	E	rRNA maturation through position specific methylation (C/D Box) and pseudouridylation (H/ACA box).
Group I and II introns		all	Self splicing ribozymes. Mobile elements with no known biological function in the host.
Riboswitches		all	Gene expression regulation by transcription, translation and splicing modulation through ligand recognition.
TLC	RNA component of Telomerase	E	Elongation of telomeres.

Table 1.1: List of ncRNA families. This table displays the most relevant ncRNA families to our work. A generally complete and updated list of ncRNAs can be found in (Gardner et al., 2009).<sup>1</sup> Name and current abbreviation of the ncRNA family. <sup>2</sup> Indicates if the family is observed in eukaryotes (E) or in all domains of life (all). <sup>3</sup> Known function of the family.



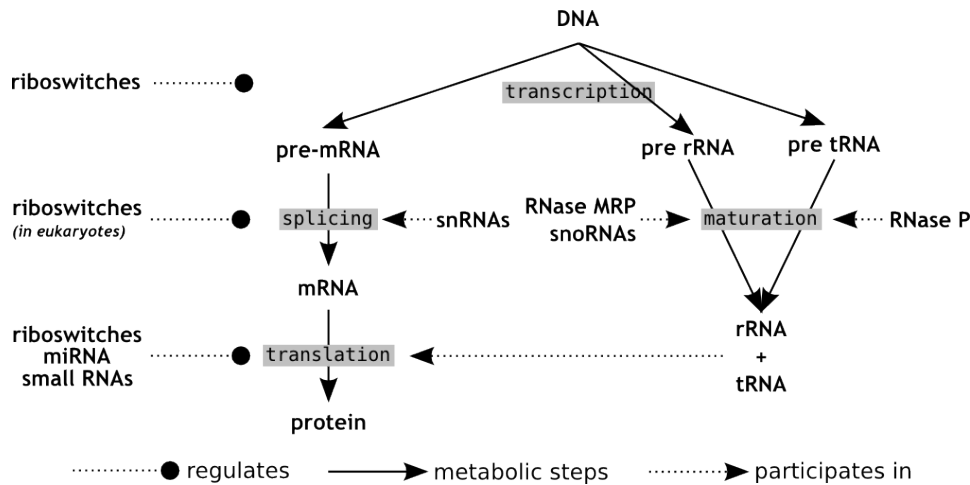


Figure 1.2: The major roles of ncRNAs. Freely inspired from: [http://en.wikipedia.org/wiki/Non-coding\\_RNA](http://en.wikipedia.org/wiki/Non-coding_RNA)

- Strong sequence conservation in short regions that play the role of guide sequences (e.g. snoRNAs). These regions tend to be conserved as they need to recognize the target sequence by sequence complementarity.
- Strong conservation in very short regions important for RNA-protein contacts.
- Ability to sustain large insertions in peripheral regions with small structural impact on the catalytic core of the molecule.

Figure 1.3 displays some examples of the evolutionary restrictions that constrain ncRNAs observed in some highly conserved ncRNA molecules in hemiascomycetous yeasts.

The fact that many sequences can assume conformations that are compatible with the structural and functional constraints of native ncRNA (Schultes and Bartel, 2000) suggests a possible “sequence neutrality” that allows a faster evolution in sequence space. The telomerase ncRNA is an extreme example of it. Figure 1.4 presents the histogram of pairwise sequence identities between 24 telomerase RNAs from closely related ciliates genes. It is striking that an essential gene occurring in all eukaryotes with, it is believed, the same function, presents a mean pair-wise sequence identity of only 54%, i.e., way beyond the capabilities of any sequence alignment tool.

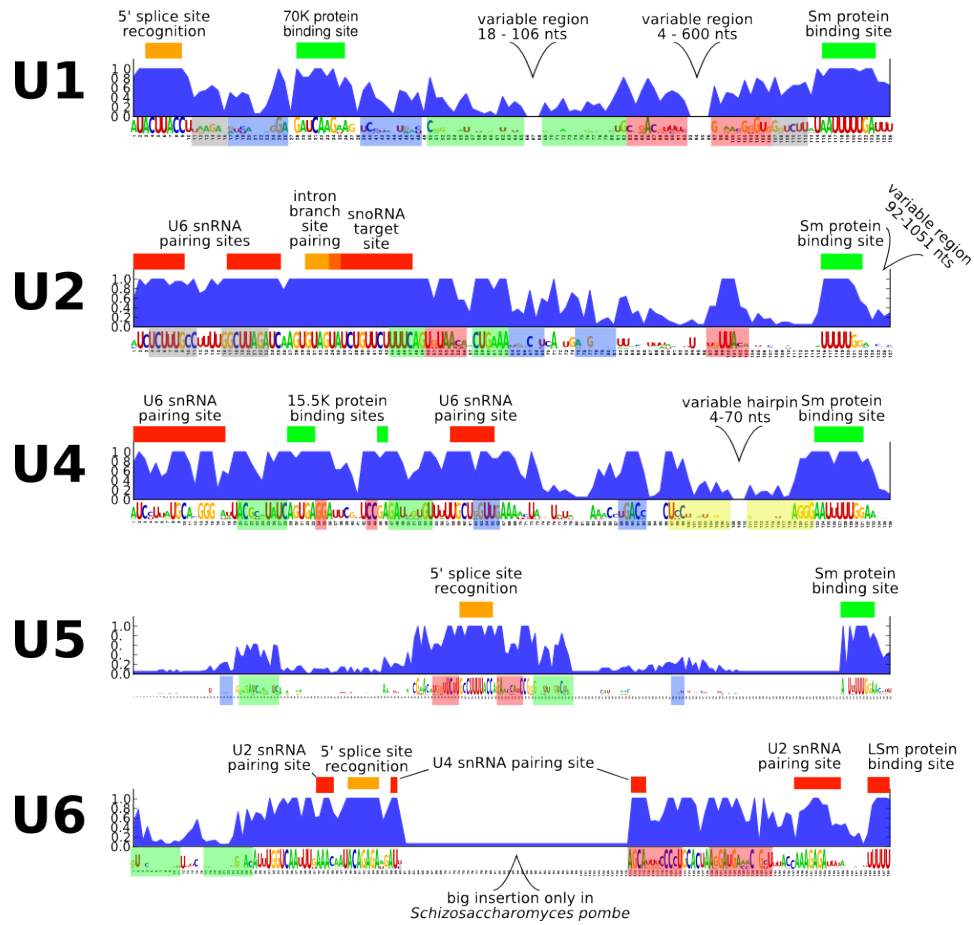


Figure 1.3: ncRNAs evolutionary constraints. Conservation profile of five snRNAs from 20 yeast species. Blue curves indicate sequence conservation measured as the entropy of the alignment column. Colored horizontal bars indicate regions responsible for inter molecular interaction with other ncRNAs (red), proteins (green) or introns (orange). Higher conservation is observed in regions involved in intermolecular interaction, with significant sequence variation outside these regions. Large indels are found mainly in *Saccharomyces*. Some of these indels, like those present in the U1 variable regions, are not essential for yeast survival. This conservation pattern is evidence for a noticeable sequence flexibility of ncRNAs: the positions not involved in intermolecular interactions are allowed to change drastically as long as structural features (mainly helical domains) are preserved.

Figure adapted from:

Cruz JA, Westhof E. *Evolution of RNA Structure and Sequence in Hemiascomycetes*. Darwin09 – Trends in Complex Systems, November 23th-27th, 2009, Palma de Mallorca, Spain

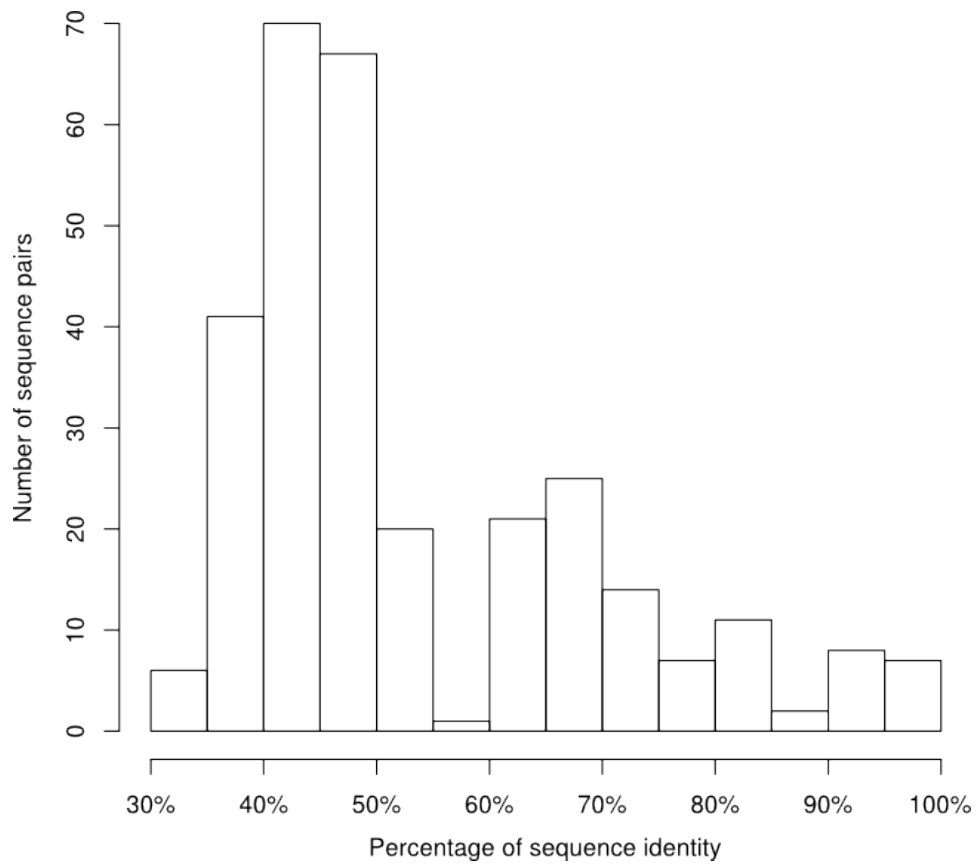


Figure 1.4: Telomerase ncRNA conservation. Histogram of the sequence identities of 276 pairs of 24 ciliate Telomerase ncRNA genes.

## 1.5 Article – The Dynamic Landscapes of RNA Architecture

The review referred in section 1.2 was published in the following article:

Cruz, J. A. and Westhof, E. (2009). *The Dynamic Landscapes of RNA Architecture*. Cell 136, 604-609.

# The Dynamic Landscapes of RNA Architecture

José Almeida Cruz<sup>1</sup> and Eric Westhof<sup>1,\*</sup>

<sup>1</sup>Architecture et Réactivité de l'ARN, Université de Strasbourg, Institut de Biologie Moléculaire et Cellulaire du CNRS, 67084 Strasbourg, France

\*Correspondence: e.westhof@ibmc.u-strasbg.fr  
DOI 10.1016/j.cell.2009.02.003

**A wealth of information on RNA folding and ribonucleoprotein assembly has emerged from analyses of structures and from the use of innovative biophysical tools. Although integrating data obtained from static structures with dynamic measurements presents major challenges, such efforts are opening new vistas on the RNA folding landscape.**

Although only 2% of the human genome codes for proteins, it is now realized that almost all of the genome is transcribed and new biological functions for RNA transcripts are frequently being discovered (Amaral et al., 2008). The biological roles performed by noncoding RNAs (see Review by C.P. Ponting, P.L. Oliver, and W. Reik in this issue of *Cell*) depend on the native three-dimensional structures that they form, both by themselves and in complexes with ligands and proteins. Most of our current knowledge about RNA structures comes from X-ray crystallography or nuclear magnetic resonance (NMR). Although these techniques reveal snapshots of a dynamic reality, they convey little information about the steps taken by a linear chain of nucleotides to fold into complex and intricate three-dimensional arrangements. Beginning to fill this gap are increasingly sophisticated biophysical tools, such as single-molecule optical traps, time-resolved fluorescent resonance energy transfer (FRET), and hydroxyl radical footprinting. With these tools, it is possible to monitor conformational changes undergone by RNA molecules during their folding and during assembly of ribonucleoproteins (RNPs). Here, we examine our present understanding of RNA architecture and RNP assembly, which is built upon a foundation of static structures, in the light of the insights gained through analysis with biophysical tools.

## RNA Architecture Is Modular and Hierarchical

Three-dimensional structures reveal that RNAs have a hierarchical organization in which secondary structural elements,

such as double-stranded helices, hairpins, and single-stranded loops, are connected by tertiary interactions. Although double-stranded helices are maintained by Watson-Crick base pairs and require only monovalent ions, the tertiary contacts are dominated by non-Watson-Crick base pairs and generally require the presence of divalent ions, especially magnesium ions (Tinoco and Bustamante, 1999).

There are also many recurrent structural assemblies that can best be described as modules given that they have only a minor dependence on the surrounding sequences or contacts. Such modules formed by non-Watson-Crick interactions organize internal loops or helical junctions and are found embedded within or between regular helices. Some of the most frequent modules are the sarcin-ricin loop, the K-turn, or the C-loop. Although they are often associated with a similar structural role (a kink in a helical domain or variations in helical twist), they generally bind to a great variety of ligands or proteins (Lescoute and Westhof, 2006).

To begin, we describe some of the static features of RNA assembly as deduced from folded architectures. They reveal complex networks of interactions, most of which are weak, that cooperatively stabilize the fold.

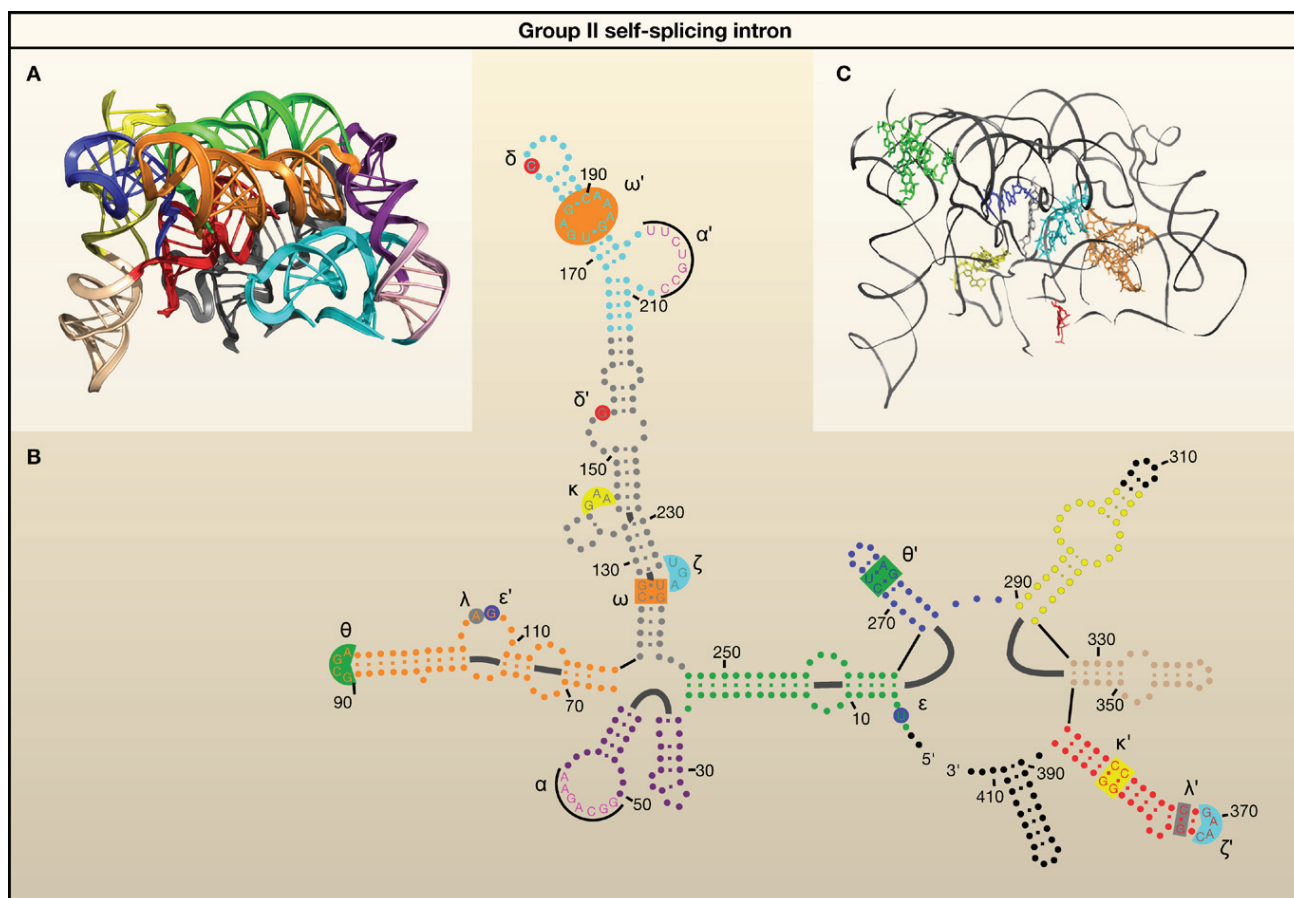
## RNA-RNA Self-Assembly Motifs

RNA architecture is dominated by continuous base stacking leading to co-axial stacks of helical domains packed parallel or orthogonal to one another as beautifully displayed by the recent structure

of a group II self-splicing intron (Toor et al., 2008) (Figure 1). The pack of stacks is maintained by an intricate network of contacts, which are either further Watson-Crick base pairs (as in kissing loops or pseudoknots) or non-Watson-Crick pairs (Lescoute and Westhof, 2006). The latter belong overwhelmingly to the A-minor interactions in which two consecutive adenine nucleotides interact via their sugar edges with the sugar edges of Watson-Crick paired nucleotides. The sequence specificities of these contacts vary from complete absence (as in ribose zippers) to exquisitely precise contacts (such as those between a GAAA tetraloop and an 11 nt motif). The lack of a strong link between specific sequences and many of the forms of A-minor interactions (meaning that sequence variations are neutral for RNA-RNA interactions) imply that further constraints must exist to guarantee specific and native folding of structured RNAs. This begs the question—how is specificity of folding achieved?

## The Central Role of Junctions in RNA Architecture

Helical junctions are the point of connection between a group of helical segments. They are particularly important for RNA folding given their role in promoting the correct co-axial stacking of helical domains and thus the correct positioning in space of the RNA-RNA assembly motifs (Lescoute and Westhof, 2006). Junctions are often organized by sets of non-Watson-Crick pairs (for instance a sarcin-ricin module) in many structured RNAs including ribosomal



**Figure 1. Bridging the 2D and 3D Worlds**

Architecture of the self-spliced product of a group II intron ribozyme (Toor et al., 2008; PDB code 3BWP). (A) RNA compaction occurs through helical packing. (B) A conventional representation of the secondary structure indicating the co-axial stacks of helices and the long-range tertiary contacts through either Watson-Crick base pairs, as in the  $\alpha$ - $\alpha'$  loop-loop interaction, or non-Watson-Crick base pairs, as in the  $\theta$ - $\theta'$  GNRA tetraloop/helix contact. In (A) and (B), the catalytically active helix is shown in red. (C) The tertiary contacts are represented on a simplified 3D representation in the same orientation as in (A).

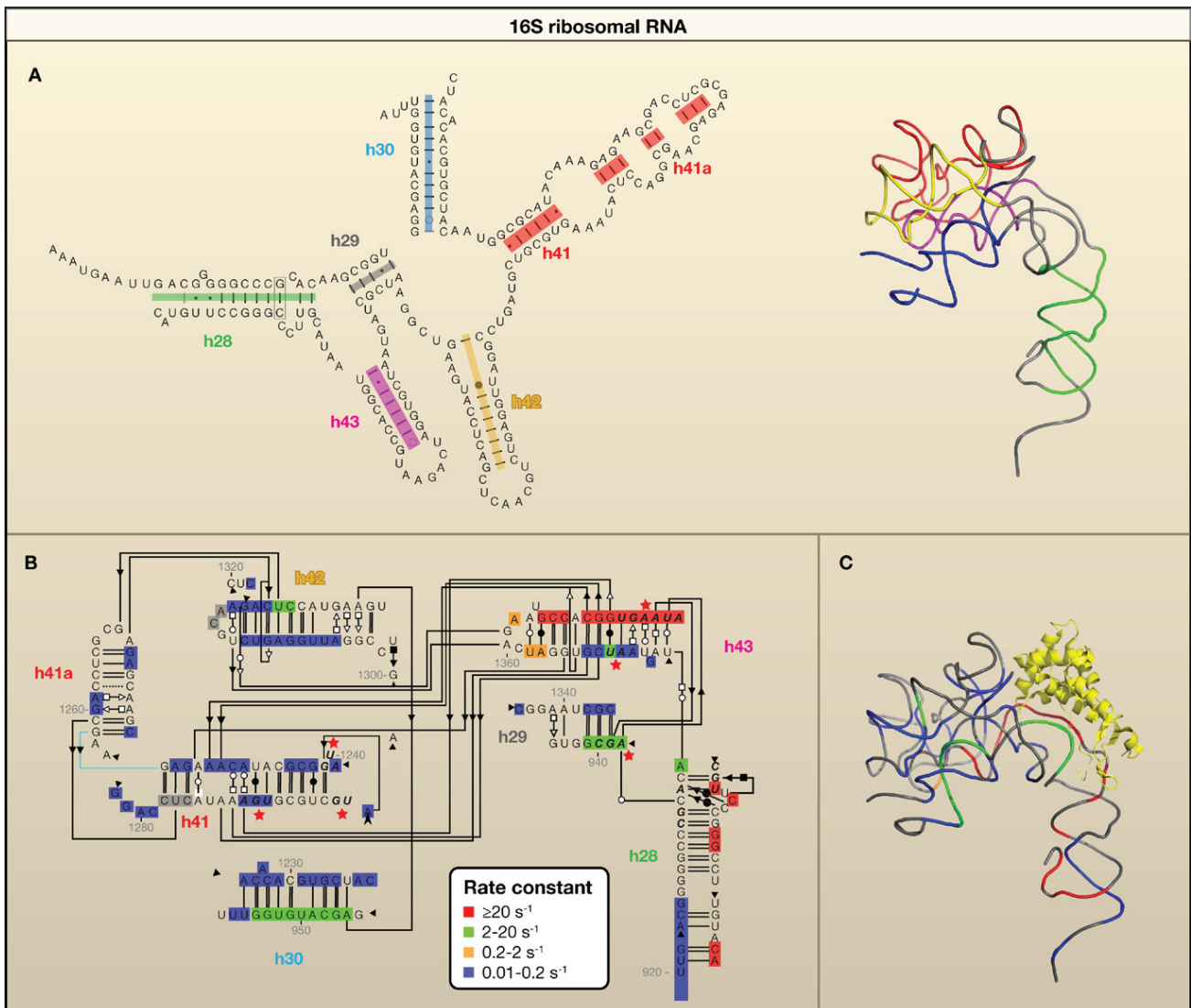
RNAs. Junctions allow for conformational diversity that can be modulated by either RNA-RNA or RNA-ligand interactions. The hammerhead ribozyme provides a clear illustration of this trait. Its active site consists of a three-way helical junction containing a central core with 15 highly conserved nucleotides, which are essential for catalytic activity (see Essay by T. Cech in this issue of *Cell*). Its activity, however, is strongly dependent on loop-bulge or loop-loop tertiary interactions in nonconserved regions far away from the active site (Martick and Scott, 2006). These long-range tertiary interactions stabilize the active conformation of the junction positioning the relevant nucleotides in the exact positions for catalysis. Likewise, in riboswitches, helical junctions often form the binding site for the ligand that regulates its activity

(Montange and Batey, 2008). In the 30S ribosome, the primary assembly proteins (S4, S15, and S7) bind to key junctions of the 16S ribosomal RNA (16S rRNA) (Brodersen et al., 2002), which are preorganized by sets of non-Watson-Crick base pairs.

#### Preorganization and Quasi-hierarchical Folding

Compared to tertiary interactions, base stacking interactions and Watson-Crick base pairs have a greater contribution to the energetic stability of RNA structures. This difference in relative contribution underlies the hierarchical model for RNA architecture in which preorganized secondary structural domains fold independently and simultaneously in the initial stage of folding, followed by the formation of tertiary interactions (Tinoco and

Bustamante, 1999). However, it is now appreciated that the hierarchical view of folding is a first-order approximation in that tertiary structure formation can also lead to secondary structure rearrangements or precede formation of a helical domain as recently shown for the adenine riboswitch (Greenleaf et al., 2008; Noeske et al., 2007). Furthermore, data from time-resolved hydroxyl radical footprinting have shown that the overall folding of group I introns relies strongly on specific tertiary interactions that assist in the formation of native-like intermediate structures that rapidly fold into the native conformation. Single mutations affecting the interaction between a GAAA tetraloop and an 11 nt motif alter folding speed and accuracy as they produce non-native intermediate structures prone to becoming trapped



**Figure 2. Tracking the Dynamics of Folding**

(A) (Left) A standard representation of a part of the secondary structure of *E. coli* 16S ribosomal RNA showing the co-axial stacking. Only helices h28 to h30 and h41 to h43 are represented. (Right) A three-dimensional view of the same elements with colors corresponding to the co-axial stacks (PDB code 1FJG).

(B) The local 16S tertiary interaction network, which incorporates co-axial stacking and non-Watson-Crick contacts (Lescoute and Westhof, 2006). The nucleotide protection data of *E. coli* 16S rRNA, obtained by time-resolved X-ray hydroxyl radical footprinting (Adilakshmi et al., 2008), are superimposed. The protection rate of bases is indicated by the color code: bases with higher protection rate form tertiary RNA-RNA or protein-RNA interactions first. Stars and boldface nucleotides represent the binding sites of the protein S7. The code for the non-Watson-Crick pairs is as follows: Watson-Crick edge, circle; Hoogsteen edge, square; sugar edge, triangle. The symbols are darkened when the two nucleotides approach in the *trans* orientation.

(C) The three-dimensional representation of the 16S rRNA backbone with S7 protein (yellow) (PDB code 1FJG).

in metastable conformations (Chauhan and Woodson, 2008). Thus, when tertiary interactions form cooperatively with helices, kinetic traps, which can lead to metastable conformations, are more easily avoided. In this way, quasi-hierarchical folding guides an RNA through the conformational space into the native conformation. However, in addition to the internal RNA interaction networks,

ligand or protein binding leads to external interaction networks that contribute to the final RNA structure.

#### Protein Binding

A recent examination of assembly of the 30S ribosome in vitro using time-resolved X-ray hydroxyl radical footprinting (Adilakshmi et al., 2008) led to the following conclusions (Figure 2): (1) the three main domains fold simultaneously

and autonomously; (2) in agreement with pulse-chase experiments using quantitative mass spectrometry (Talkington et al., 2005), the primary binding proteins bind very early in the folding process; (3) assembly is nucleated at various locations of the 16S rRNA secondary structure at the same time, leading to parallel routes of folding to the native fold; (4) and finally, the binding sites for protein on the

RNA are not protected at the same rates. Those protected fastest belong to sets of non-Watson-Crick base pairs that organize helical junctions, whereas those protected more slowly are the result of slow reorganization and induced fit of the RNA-protein complexes (Figure 2). The assembly of the *Tetrahymena* telomerase RNP, which is induced by p65 binding to an evolutionarily conserved GA bulge, is also hierarchical, as shown by a single-molecule FRET approach (Stone et al., 2007).

### Ligand Binding

Riboswitches, regions of mRNAs responsible for gene regulation in bacteria, plants, and fungi, are able to change conformation in the presence of specific small metabolites (see Review by L.S. Waters and G. Storz in this issue of *Cell*). Metabolite binding can repress gene expression either by folding the riboswitch with a transcription termination structure or by sequestering the Shine-Dalgarno translation initiation sequence (Breaker, 2008). Alternatively, the presence of the metabolite can lead to activation of transcription (by formation of an antiterminator structure) or initiate translation (by releasing the Shine-Dalgarno region). Riboswitches prefold into a ligand recognition domain that typically forms around a multihelical junction. In one class of riboswitches (which includes purine, glmS, and SAM-II), ligand binding to the pocket stabilizes the fold. Ligand binding induces mainly local adjustments to the prefolded conformation. NMR spectroscopy and X-ray experiments show that both free and ligand-bound riboswitches share helical domains and even tertiary loop-loop interactions. The ligand-free riboswitch, however, has a dynamic and unstructured binding site (Noeske et al., 2007). In a second class of riboswitches (which includes TPP, SAM-I, and M-Box), the ligand brings together two domains of the binding pocket that are far apart in the preorganized structure (Montange and Batey, 2008). The above examples show that hierarchical folding is far from being a straightforward and general model applicable to any length RNA and at all timescales. Several recent papers convey the diversity in folding processes noting that hierarchical folding can differ in folding speed, specificity of the initial

tertiary assembly, existence of misfolded intermediates, and occurrence of local rearrangements of the native structure (Russell et al., 2006; Pereira et al., 2008; Waldsich and Pyle, 2008).

### Induced Fit and Catalysis

Most structures of RNP complexes reveal that recognition for the RNA component involves an induced fit. RNP complexes display high cooperativity in contacts and mutual induced fits between the components, much of which is not seen in either the isolated components or in partially assembled complexes. Similarly, in catalytic RNAs, loop-loop interactions between peripheral domains produce local conformational rearrangements in the active site, thereby exerting a massive influence on catalytic activities and on the requirement of Mg<sup>2+</sup> ions for activity. This is exemplified by the recent crystal structures of hammerhead ribozymes (Martick and Scott, 2006). The recent crystal structure of a group II intron fragment (Toor et al., 2008) displays a distortion in the helical domain constituting the active site that is induced by the close packing of neighboring interconnected helices (Figure 1). Such structural distortion and stabilization of local networks propagated at a distance by multiple tertiary contacts between RNA domains are far beyond the reach of computational simulations currently available.

### Coupling between Electrostatic and Architectural Hierarchies

RNA molecules are very negatively charged and thus their assembly is strongly coupled to the electrostatic environment. The majority of folding data have been obtained by *in vitro* experiments, which are affected by ionic and temperature conditions. Cations promote folding by creating an ionic atmosphere around the RNA molecule that counteracts the repulsive force of the negatively charged backbone and allows helix packing. In addition to this nonspecific effect on folding, cations also have specific roles, such as binding directly after partial dehydration to particular pockets of the structure. Differences in ion types and in ionic concentrations *in vitro* can drastically affect folding speed and rates of misfolding, suggesting that the energy landscape

of folding is rugged, with many possible pathways (and kinetic traps) on the way to the final fold (Chauhan and Woodson, 2008). As discussed above, *in vitro* folding is very susceptible to misfolding due to the high energy content of the helical elements as observed recently by single-molecule unfolding experiments (Woodside et al., 2008). Starting from folded RNAs, pulling leads to the breakdown of tertiary structures followed by unfolding of the helices (Greenleaf et al., 2008). In contrast, with high Mg<sup>2+</sup> concentration and high temperature, a group II intron folds after a slow step into on-pathway intermediate states leading rapidly to the native conformation (Waldsich and Pyle, 2008; Steiner et al., 2008). During *in vivo* folding, the coupling between electrostatics and architecture is monitored by the polymerization process itself and by binding of protein factors. *In vitro*, higher ionic concentrations can compensate for the lack of *in vivo* folding factors.

### Cotranscriptional Folding

Because the folding of RNA helices is 2–3 orders of magnitude faster than the rate of transcription, base-base recognition takes place as soon as the emerging strand of RNA reaches sufficient length to allow folding. Transcription speed, modulated by the elongation speed of RNA polymerase and sequence-specific pauses, can thus influence the RNA folding dynamics in diverse ways. The early transcribed regions can start to fold, potentially privileging locally stable structures and competing with more stable global structures that would form with a longer transcript. Thus, transcription speed can drastically affect the propensity of group I introns to fold properly or misfold (Jackson et al., 2006). Polymerase pausing is important for efficient folding of some noncoding RNAs in *E. coli* (such as RNase P, signal recognition particle [SRP], and transfer-messenger RNA [tmRNA]) by allowing for temporary sequestration of non-native helices that late in the transcription process will form the native structure more efficiently (Wong et al., 2007). Force-dependent kinetic measurements using single-molecule techniques beautifully demonstrate the multiple pathways present in transcription termination by bacterial RNA polymerase. They show how tran-



scription termination is fine tuned by energetic competition between anti-termination hairpin formation and closure, alternative pairing with upstream sequence, and the stability of the DNA-RNA hybrid (Larson, et al., 2008). In the case of the adenine riboswitch, FRET experiments show that the ligand binds the riboswitch after the transcription of the binding domain but before the complete transcription of the expression platform, highlighting the dependence on the order of transcription (Lemay et al., 2006). In the flavin mononucleotide (FMN) riboswitch the ligand concentration necessary to switch off transcription is higher than the apparent dissociation constant. If the FMN concentration is not sufficiently high, transcription will be completed before ligand binding reaches thermodynamic equilibrium (Wickiser et al., 2005). This last observation suggests ways by which evolution can fine tune responses to a given concentration of a metabolite by changing the binding affinity (through sequence variations) or by changing transcription speed (Breaker, 2008).

### RNA Chaperones

Beyond the evidence concerning cotranscriptional folding, many other observations can only be explained by posttranscriptional effects on RNA conformation. Hairpin ribozymes consist of four helices H1 to H4. Mahen et al. (2005) inserted complementary sequences that prevented catalysis by impeding the formation of H1 in the 5' and 3' ends of the ribozyme, obtaining two variants of the molecule. In vitro experiments have shown that catalytic activity of the 3'-end variant is less impaired than that of the 5'-end variant. This was expected given that helix H1, which is transcribed first, has time to fold in the case of the 3'-end variant but not in the 5'-end variant. Surprisingly, in vivo, both variants lose the ability to self-cleave. This result strongly suggests that cotranscriptional folding is not sufficient for guaranteeing native folding and, thus, that cellular factors affect in vivo folding. One mechanism could be the ability to recruit proteins participating in folding during or after transcription (Mahen et al., 2005). RNA chaperone activity is generally understood as resulting from nonspecific unfolding mecha-

nisms acting on the folded/misfolded equilibria, thereby promoting the native structures. Recently, Bhaskaran and Russell (2007) showed that the CYT-19 protein, a DExD/H-box helicase, unfolds both native and misfolded helices of the group I intron ribozyme of *Tetrahymena thermophila* but has a preference for less stable structures lacking tertiary interactions.

### Folding Robustness and Evolvability

The explicit relations between sequence, structure, and function make the study of RNA a source of insights into understanding mutational robustness in biological macromolecules, defined as the ability to maintain structure or function upon mutation (Wagner, 2008). For structured RNAs, the most obvious way mutations can affect function is by provoking alterations in the native structure or in its stability. Due to compensatory mutations and base-pair isostericity (structural similarity), RNA molecules can sustain a fair number of mutations without dramatic variation to the structure or loss of function (Leontis et al., 2002). Another, more subtle way by which mutations can affect RNA function is by changing the folding pathways and the local helical stabilities (Larson, et al., 2008; Woodside et al., 2008). Multiple sequence alignments of homologous RNAs yield a rich display of the sequence space or the range of variation accessible to a family of structured RNAs. It is, however, striking to observe in structured RNAs that invariant residues are infrequent. This observation, together with the frequent structural neutrality of RNA-RNA interactions, offers a solid basis for understanding the robustness of RNA architecture. Further, the observed robustness of RNA molecules is compatible with their ability to evolve in form and function. Recent theoretical work on RNA secondary structures explains this apparent contradiction by exploring the concept of neutral networks in sequence space. A neutral network connects sequences with similar structure and purportedly similar function. A molecule that is structurally robust will have a larger neutral network, increasing the number of reachable neighbor structures from other networks (Wagner, 2008). This model is in accordance with experimental results

obtained by Schultes and Bartel (2000) who artificially evolved two ribozymes with distinct sequences and functions, one single mutation at a time, until converging into a unique common sequence at the frontier of both neutral networks. This experiment shows, at least in vitro, that two functionally distinct molecules can be separated by only a handful of mutations such that a small number of evolutionary events are enough to produce distinct functions (Schultes and Bartel, 2000).

### Predictions of Folding Models and Pathways

The many factors involved in RNA folding confound our ability to make predictions for how the linear nucleotide chain of an RNA molecule achieves its native structure and which (and how many) intermediate conformations exist. The early secondary structure prediction tools were based on maximizing the number of stacked Watson-Crick base pairs and the assumption that the energy of tertiary interactions could be treated as perturbations (Tinoco and Bustamante, 1999). However, those algorithms cannot easily take into account cotranscriptional constraints to the folding process but, rather, produce a set of candidate native structures without providing any insight regarding the folding pathways. Kinetic folding is an alternative approach that simulates the folding process, either directly or indirectly, reproducing the conformation pathway of RNA molecules. Unfortunately, the number of possible alternative conformations at each step of the folding process grows exponentially with sequence length, rendering any exhaustive or systematic exploration of the folding space computationally infeasible for medium or large RNAs. One way to circumvent the combinatorial explosion of alternatives is to use known data about folding dynamics to restrict the search space. A recent algorithm (Geis et al., 2008) attempts to combine both approaches, relying on the observation that locally optimal substructures or combinations of such structures are important folding intermediates. It progresses stepwise through sets of subintervals of the full sequence. The most stable structures inside each subinterval are generated using a

dynamic programming algorithm, and the folding path is retrieved if the generated substructure was selected as part of the global structure. This algorithm is able to make predictions for RNAs of up to 1500 nt, showing that the integration of qualitative knowledge about the folding process is a promising approach to tackling this computationally demanding task. Yet, this approach has its limitations given that three-dimensional information is not considered. For example, it would not be possible using this approach to predict the stabilization of a riboswitch in the presence of a ligand.

### Computational Challenges

Biophysical measurements show that conformational rearrangements, which are stabilized or induced during the hierarchical assembly process, drive RNA architecture or RNP complex formation cooperatively and through multiple pathways. Integrating data from such dynamic views with the static folded architectures presents new computational challenges for modeling and simulations. Induced fit changes in conformation or distortions propagated at a distance either by tertiary interactions or as a result of protein binding cannot yet be adequately simulated. Similarly, the simulation of folding kinetics with various concentrations of ligands is currently out of reach.

Could the new biophysical insights contribute to bioinformatics? In other words, how can we integrate genomic data and sequence alignments with the biophysical data? Conversely, how can we exploit the emerging wealth of structural knowledge to search genomes and identify new functional RNAs?

### Perspectives

The study of the kinetics of the folding processes leading to native RNA architectures has made huge progress in recent years following the recognition of the regulatory roles of RNAs and the application of single-molecule and new fluorescence techniques. Since the early work of Yanofsky on transcriptional attenuation (Merino and Yanofsky, 2005), it has been accepted that alternative pairings between RNA segments play key roles in biological regulation.

The deeply rooted biological functions of riboswitches (Breaker, 2008) forced upon us the realization that static structures, though central and key for our molecular understanding, do not give the complete picture or framework. Furthermore, the diversity in interactions and functions of sense/antisense RNA complexes is now appreciated, not only in bacteria but also in eukaryotic cells. Thus, in the years to come, the molecular biophysics of intermolecular complex formation between RNA strands as well as between RNAs and proteins will remain a frontier for research. Single-molecule studies, together with molecular simulations, have brought RNA folding into the realm of statistical mechanics by revealing the intrinsic molecular dynamics. This comes amidst a shifting view of biological systems that puts a greater emphasis on the uniqueness of single cells in space and time, a trend that is coming about from a combination of deep sequencing, the view of stochastic gene expression, and a growing appreciation of cell-to-cell variability. Making measurements of RNA at biologically relevant time and space granularity is more urgent than ever.

### ACKNOWLEDGMENTS

J.A.C. is supported by the Ph.D. Program in Computational Biology of the Instituto Gulbenkian de Ciência, Portugal (sponsored by Fundacao Calouste Gulbenkian, Siemens SA, and Fundacao para a Ciencia e Tecnologia; SFRH/BI/33194/2007).

### REFERENCES

- Adilakshmi, T., Bellur, D.L., and Woodson, S.A. (2008). *Nature* 455, 1268–1272.
- Amaral, P.P., Dinger, M.E., Mercer, T.R., and Matlack, J.S. (2008). *Science* 319, 1787–1789.
- Bhaskaran, H., and Russell, R. (2007). *Nature* 449, 1014–1018.
- Breaker, R.R. (2008). *Science* 319, 1795–1797.
- Brodersen, D.E., Clemons, W.M., Carter, A.P., Wimberly, B.T., and Ramakrishnan, V. (2002). *J. Mol. Biol.* 316, 725–768.
- Chauhan, S., and Woodson, S.A. (2008). *J. Am. Chem. Soc.* 130, 1296–1303.
- Geis, M., Flamm, C., Wolfinger, M.T., Tanzer, A., Hofacker, I.L., Middendorf, M., Mandl, C., and Stadler, P.F. (2008). *J. Mol. Biol.* 379, 160–173.

Greenleaf, W.J., Frieda, K.L., Foster, D.A.N., Woodside, M.T., and Block, S.M. (2008). *Science* 319, 630–633.

Jackson, S.A., Koduvayur, S., and Woodson, S.A. (2006). *RNA* 12, 2149–2159.

Larson, M.H., Greenleaf, W.J., Landick, R., and Block, S.M. (2008). *Cell* 132, 971–982.

Lemay, J.F., Penedo, J.C., Tremblay, R., Lilley, D.M.J., and Lafontaine, D.A. (2006). *Chem. Biol.* 13, 857–868.

Leontis, N.B., Stombaugh, J., and Westhof, E. (2002). *Nucleic Acids Res.* 30, 3497–3531.

Lescoute, A., and Westhof, E. (2006). *Nucleic Acids Res.* 34, 6587–6604.

Mahen, E.M., Harger, J.W., Calderon, E.M., and Fedor, M.J. (2005). *Mol. Cell* 19, 27–37.

Martick, M., and Scott, W.G. (2006). *Cell* 126, 309–320.

Merino, E. and Yanofsky, C. (2005) *Trends Genet.* 21, 260–264.

Montange, R.K., and Batey, R.T. (2008). *Annu. Rev. Biophys.* 37, 117–133.

Noeske, J., Buck, J., Fürtig, B., Nasiri, H.R., Schwalbe, H., and Wöhnert, J. (2007). *Nucleic Acids Res.* 35, 572–583.

Pereira, M.J.B., Nikolova, E.N., Hiley, S.L., Jai-karan, D., Collins, R.A., and Walter, N.G. (2008). *J. Mol. Biol.* 382, 496–509.

Russell, R., Das, R., Suh, H., Travers, K.J., Laderach, A., Engelhardt, M.A., and Herschlag, D. (2006). *J. Mol. Biol.* 363, 531–544.

Schultes, E.A., and Bartel, D.P. (2000). *Science* 289, 448–452.

Steiner, M., Karunatilaka, K.S., Sigel, R.K.O., and Rueda, D. (2008). *Proc. Natl. Acad. Sci. USA* 105, 13853–13858.

Stone, M.D., Mihalusova, M., O'Connor, C.M., Prathapam, R., Collins, K., and Zhuang, X. (2007). *Nature* 446, 458–461.

Talkington, M.W.T., Siuzdak, G., and Williamson, J.R. (2005). *Nature* 438, 628–632.

Tinoco, I., and Bustamante, C. (1999). *J. Mol. Biol.* 293, 271–281.

Toor, N., Keating, K.S., Taylor, S.D., and Pyle, A.M. (2008). *Science* 320, 77–82.

Wagner, A. (2008). *Proc. Biol. Sci.* 275, 91–100.

Waldsich, C., and Pyle, A.M. (2008). *J. Mol. Biol.* 375, 572–580.

Wickiser, J.K., Winkler, W.C., Breaker, R.R., and Crothers, D.M. (2005). *Mol. Cell* 18, 49–60.

Wong, T.N., Sosnick, T.R., and Pan, T. (2007). *Proc. Natl. Acad. Sci. USA* 104, 17995–18000.

Woodside, M.T., Anthony, P.C., Behnke-Parks, W.M., Larizadeh, K., Herschlag, D., and Block, S.M. (2008). *Science* 314, 1001–1004.

## Chapter 2

# RNA Structure Comparison

Comparison between the atomic structure of biological macromolecules is a recurrent operation with particular importance for several biological domains such as the analysis of evolutionary and functional relationship between homologous molecules, the automatic search in structural databases, the discovery of structural modules in experimental structures, the evaluation of structure prediction tools, among others. In its simplest terms, structure comparison is about determining how similar (or different) two molecular structures are at the atomic scale. This simple definition hides a complex problem: How to quantify the structural divergence between two models of the same molecule? Which domains of the comparing structures most contribute to the quantified divergence? How to interpret the divergence value in a biologically meaningful way?

In the present work we are interested in a particular structure comparison problem: Finding appropriate metrics to evaluate predicted RNA models of a molecule of known structure.

Three-dimensional RNA structure prediction is an expanding field of research. A number of recently published tools aim to produce complete and biologically significant three-dimensional RNA models from sequence information in a more or less automatic fashion (Ding et al., 2008; Martinez et al., 2008; Parisien et al., 2009; Sharma et al., 2008; Das et al., 2010; Jossinet et al., 2010; Cao and Chen, 2011; Rother et al., 2011). Having appropriate methods and tools to evaluate predicted models and to compare prediction tools is a pressing need of the field (Parisien et al., 2009; Hajdin et al., 2010).

What is a biologically significant structure prediction? How to evaluate a given predicted model? How to know if a given prediction tool or approach is consistently effective? How do those tools sample the space of solutions? In which particular scenarios (molecule type, size, complexity, ...) they perform better or worse? A practical way to address these questions is to systematically compare the predicted models against known, atomic reso-

lution, native structures, obtained by X-Ray crystallography and study the observed similarities and differences.

An additional motivation to search for good comparison metrics is that automatic structure prediction tools tend to produce hundreds or even thousands of models which are impossible to manually analyze one-by-one, thus requiring some sort of automatic comparison methodologies.

The most commonly used comparison metric, both for proteins as for RNAs, is the Root Mean Square Deviation (*RMSD*). Although *RMSD* is simple to formulate and to compute, it lacks a clear interpretation. To provide a meaningful measure of structure similarity, a comparison metric should take into account the nature of the molecules being compared and their most relevant structural features. It should also provide indications on the features that contribute or penalize the similarity value. Many structure comparison metrics have been proposed for protein structures (Holm and Sander, 1993; Falicov and Cohen, 1996; Gerstein and Levitt, 1998; Eidhammer et al., 2000; Siew et al., 2000; Yang and Honig, 2000; Lancia and Istrail, 2003; Betancourt and Skolnick, 2001; Carugo, 2003; Zhang and Skolnick, 2004; Mamidipally et al., 2009). Unfortunately, tools specifically developed for protein structure comparison do not adapt to RNA comparison and the amount of work developed by the RNA structure community is significantly smaller. In the present chapter we propose two RNA specific comparison metrics that address some of the enumerated issues.

## 2.1 Structure Comparison

The pairwise comparison of RNA structures<sup>1</sup> requires (i) an alignment, i.e. some method to establish a correspondence between the comparing elements of each structure and (ii) a metric, i.e. a measure of similarity between the aligned structures.

We start with a formal definition of pairwise comparison between two RNA structures defined by the coordinates of their atoms in  $\mathbb{R}^3$ . Let  $R$  be the reference structure,  $S$  a predicted model of  $R$  and  $r_i, s_j$  their respective elements<sup>2</sup>. Let  $A$  represent an alignment between  $R$  and  $S$ :

$$A(R, S) = [(r_{i_1}, s_{j_1}), (r_{i_2}, s_{j_2}), \dots, (r_{i_n}, s_{j_n})],$$

with

$$i_x \neq i_y \text{ and } j_x \neq j_y, \forall x \neq y$$

---

<sup>1</sup>Although we discuss specifically the RNA structure comparison problem, it is worth to bear in mind that most of the generic concepts also apply to proteins with minimal adaptations.

<sup>2</sup>When comparing RNA sequences the compared elements are the sequence nucleotides. When compared structures the compared elements can be the corresponding atoms of each comparing structure or some representation of them (e.g. the center of mass of the atoms of a nucleotide, a specific atom of each base, ...).

and  $M$  a similarity measure between  $R$  and  $S$  given the alignment  $A$  such as:

$$M_A(R, S) = \begin{cases} 0 & \text{if } R \text{ is structurally identical to } S. \\ > 0 & \text{otherwise.} \end{cases}$$

By “structurally identical” we mean that there is a rigid body transformation  $T$ , i.e. a sequence of translations and rotations applied to the atomic coordinates of  $S$  such as:

$$TS = R.$$

The problem of defining the alignment  $A$  is a classic bioinformatics problem of sequence alignment. When the nucleotide sequences of  $R$  and  $S$  differ by insertions or deletions finding the optimal alignment  $A$  can be far from trivial. In our case, however, we are interested in comparing predicted RNA models with native structures, therefore the alignment is reduced to a one-to-one correspondence between each  $r_i$  and  $s_j$ <sup>3</sup>. Thus, we will concentrate on the search for meaningful measures of  $M$ .

### 2.1.1 Root Mean Square Deviation

The *RMSD* is the most commonly used molecular structure comparison metric. More generally, the *RMSD* can be used to compare any set of predicted values with the actual set of observed values. It is a measure of prediction precision. A formal generic definition of RMSD can be given as follows: Let  $X$  be a set of  $N$  observed values generated by a given phenomena and  $Y$  a list of  $N$  predicted values produced by some model of the same phenomena. The *RMSD* between  $X$  and  $Y$  is:

$$RMSD(X, Y) = \sqrt{\frac{\sum_{i=1}^N (x_i - y_i)^2}{N}},$$

where  $x_i$  and  $y_i$  are the individual values of  $X$  and  $Y$  respectively.

To compute the *RMSD* between two molecular structures  $R$  and  $S$  the definition is similar<sup>4</sup>:

$$RMSD(R, S) = \sqrt{\frac{\sum_{i=1}^N d(r_i, ts_i)^2}{N}},$$

---

<sup>3</sup>Sometimes, for technical reasons, the sequences from  $R$  and  $S$  can differ (e.g. when the predicted model corresponds to part of the native structure or when prediction programs are unable to deal with RNA dimers and a small sequence must be added to the model to transform it into a single stranded). In all such cases, however, the alignments are fairly easy to obtain.

<sup>4</sup>In the discussion that follows all comparison between  $R$  and  $S$  implicitly assumes an alignment  $A$ . For simplicity sake we will use  $RMSD(X, Y)$  instead of  $RMSD_A(X, Y)$

In which  $r_i$  represents the atom coordinates<sup>5</sup> of  $R$ ;  $ts_i$  the atom coordinates of  $TS$ , a rigid body transformation of  $S$  that minimizes the RMSD between  $R$  and  $S$  (Kabsch, 1976); and  $d$  the euclidean distance on  $\mathbb{R}^3$ .

Another definition of *RMSD*, also called *RMSD<sub>D</sub>* (for “distance” RMSD), does not require the rigid body transformation:

$$RMSD_D(R, S) = \sqrt{\frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^N (d(r_i, r_j) - d(s_i, s_j))^2}{\frac{N(N-1)}{2}}}.$$

Although more simpler to compute than the *RMSD*, the *RMSD<sub>D</sub>* has a time complexity of  $O(N^2)$  which is harder to compute for large molecules than the linear *RMSD* (Kabsch, 1978). Additionally, the *RMSD<sub>D</sub>* fails to distinguish between mirror images of the molecule, which is not the case of the *RMSD* (Maiorov and Crippen, 1994).

Most of the studies on *RMSD* approach the problem by randomly sampling the conformation space and analyzing the obtained *RMSD* distribution. The major difficulty of this sampling approach (other than the computational challenge posed by the huge size of the structure space) is to find a random model of the molecular structure that would allow a representative sampling of the huge conformation space. While protein structures can be roughly approximated by a self avoiding chain on a three-dimensional cubic lattice (Shakhnovich and Gutin, 1990), there is no such simple model for RNA structures.

An interesting, recently published, work applied a similar approach to the study of RNA (Hajdin et al., 2010). The authors used discrete molecular dynamic simulation to sample the conformation space with coarse-grained RNA models. The *RMSD* distributions are then related with the sequence length of the models. From the *RMSD* distributions for the several studied models the authors established a P-value for successful predictions of  $P < 0.01$ . A model is considered a successful prediction if its *RMSD* is lower than:

$$RMSD_{P<0.01} < 5.1 \times N^{0.41} - 19.8,$$

for model predictions based on imposed secondary structure and:

$$RMSD_{P<0.01} < 6.4 \times N^{0.41} - 16.9,$$

for model predictions with no imposed secondary structure.

Although these values could be useful references, they are still short of a structural interpretation of *RMSD*. If one can usually agree that for a

---

<sup>5</sup>The choice of the atoms to use depends on the type of molecules being compared. In proteins is frequent to use the  $C^\alpha$  atoms. In RNA the use of all heavy atoms is an usual practice but also the center of mass of nucleotides or the phosphates from the backbone can be used.

$RMSD(R, S) < RMSD_{P < 0.01}$ ,  $S$  will be structurally close to  $R$  and, at the other extreme, for an  $RMSD(R, S) \geq R_g(R)$ <sup>6</sup>,  $S$  will be too distinct from  $R$  to be significant (Maiorov and Crippen, 1994). What can one say about the intermediate values?

## 2.2 Deformation Index

A metric specific for RNA should take into account the particular structural features of RNAs, such as secondary structure and base-base interactions. As seen before (see 1.1), RNA molecules present a hierarchical architecture formed by secondary structure elements – helices – organized by long-range tertiary interactions – mainly non Watson-Crick pairs (Leontis et al., 2002). Through the analysis of the available X-Ray structures one can easily observe that, on average, 88%, 63% and 17% of all bases of a structured RNA participate in stacking interactions, Watson-Crick pairs (WC) and non Watson-Crick pairs (non WC) respectively (see Figure 2.1).

From these numbers it is clear that the correct prediction of the WC base pairs would contribute to the similarity of the predicted model against the reference structure. A subtler effect, but no less important, is achieved by the non Watson Crick interactions (non WC). Even if, on average, only 17% of all bases are involved in non WC interactions, they occur in key regions for the structural organization: helical junctions; long range loop-loop and loop-helix interactions; and in recurrent structural modules. Therefore, the correct prediction of a few non WC bases would play a major role in the predicted model quality. Similarly, the correct prediction of stacking interactions, in particular those that occur outside the helical stacks, would be important to model the single stranded regions of the molecule which are, as we will see in the next chapter, the most challenging regions in RNA structure prediction.

In summary, the importance of WC base pairs for a correct secondary structure prediction follows from the significant proportion of bases involved, the importance of non WC pairs, key for establishing the correct interactions and orientations between the secondary structure elements and base stacking, can be important for the correct prediction of certain single stranded regions. The complete set of base-base interactions establishes an “interaction network” and it is not possible to produce a biologically meaningful model of an RNA molecule without correctly predicting it.

Thus, an alternative way of measuring the quality of a modeled structure is to evaluate how well the interaction network of the native structure is predicted. This evaluation can be achieved by computing the Matthews Correlation Coefficient (MCC) (Matthews, 1975) of the predicted base-base interactions against the interactions present in the native structure.

---

<sup>6</sup> $R_g(R)$  represents the radius of gyration of  $R$  (see Appendix A).

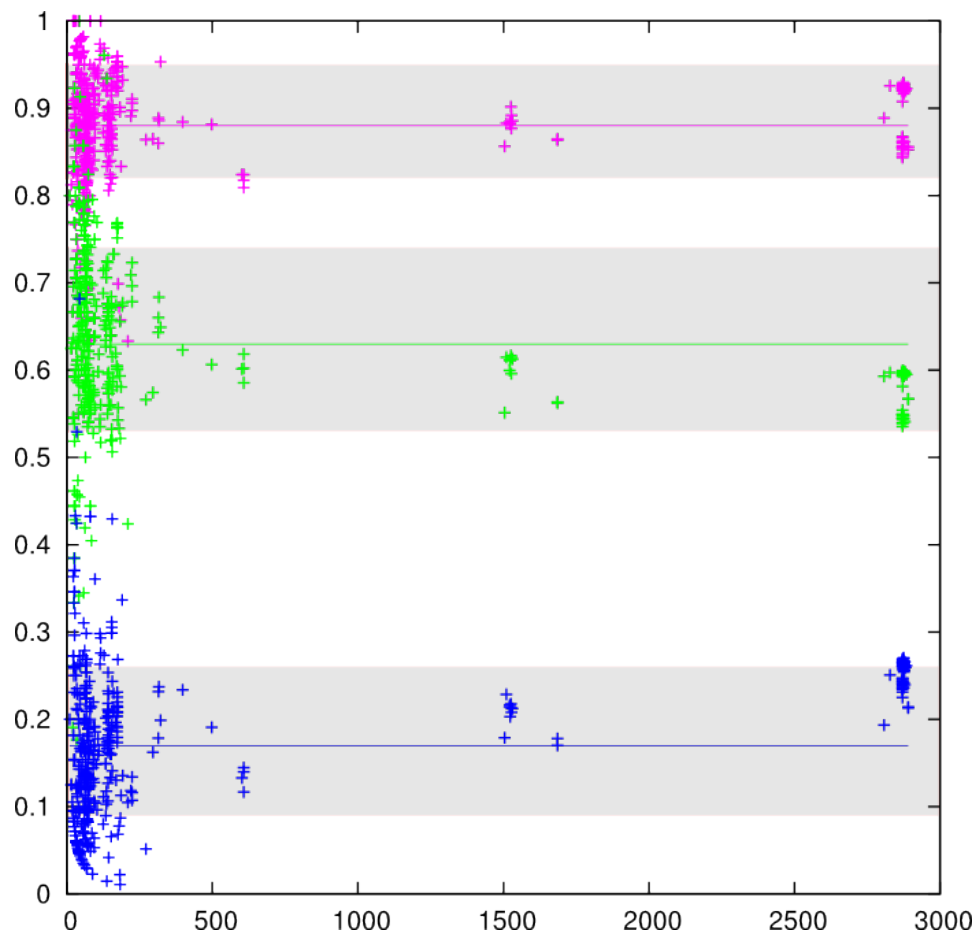


Figure 2.1: Proportion of bases participating in base stacking interactions:  $88\% \pm 7\%$  (magenta); Watson-Crick base pairs:  $64\% \pm 10\%$  (green); and non Watson-Crick base pairs:  $17\% \pm 8\%$  (blue). Solid horizontal lines represent the mean for each type of interactions and gray shaded regions the respective standard deviation. Results for 481 non redundant structured RNA crystal structures.



In general, the *MCC* between a set *A* of predicted and a set *B* native values can be estimated by:

$$MCC(A, B) = \sqrt{\left(\frac{TP}{TP + FP}\right) \times \left(\frac{TP}{TP + FN}\right)},$$

where *TP* (True Positives) is the number of correctly predicted interactions, *FP* (False Positives) is the number of incorrect predictions and *FN* (False Negative) is the number of interactions present in the native structure but not predicted in the model. The *MCC* varies from 0, for a model in which no interaction was predicted, to 1, for a model in which all interactions were correctly predicted.

Replacing *A* and *B* by the interactions in the predicted model *I<sub>S</sub>* and the native structure *I<sub>R</sub>* we can establish the Interaction Network Fidelity (INF) measure:

$$INF(I_S, I_R) = MCC(I_S, I_R).$$

Comparing the INF and RMSD values obtained for different predicted models of the same native molecule – in this case 9847 models of the rat 28S rRNA loop E produced with the MC-Sym 3D RNA prediction tool (Parisien and Major, 2008) – we observe that, even if some correlation between INF and RMSD exists (Pearson correlation coefficient  $P = 0.6$ ), for a given high value of INF many RMSD values are observed. On the other hand for small values of RMSD a wide range of INF is observed (see Figure 2.2).

To obtain the RMSD weighted by the INF value of the predicted model we propose the “Deformation Index” (*DI*) metric defined as:

$$DI(S, R) = \frac{RMSD(S, R)}{INF(I_S, I_R)}.$$

This way the RMSD assumes its own value for predicted models in which all interactions were correctly predicted ( $INF = 1$ ) and is infinite if no interaction was correctly predicted ( $INF = 0$ ).

To investigate in more detail the effect of *DI* in the structure comparison we randomly selected three of the predicted models with the *RMSD* and *INF* values shown in Table 2.1. These modules are depicted in Figure 2.3. We observe that: model *A* – with low *RMSD* and high *INF* – closely reproduces the key structural features of the native model; model *B* – with high *RMSD* and *INF* – is largely penalized by the badly predicted nucleotides A14-G15 in the loop and the shifted backbone even though the interaction network is as well predicted as in model *A*; and, finally, model *C* – with low *RMSD* and lower *INF* – fails to predict more than the double of interactions than models *A* and *B*.

Notice that the *RMSDs* for all three models are much higher than what would be considered a successful prediction based on the criteria of (Hajdin

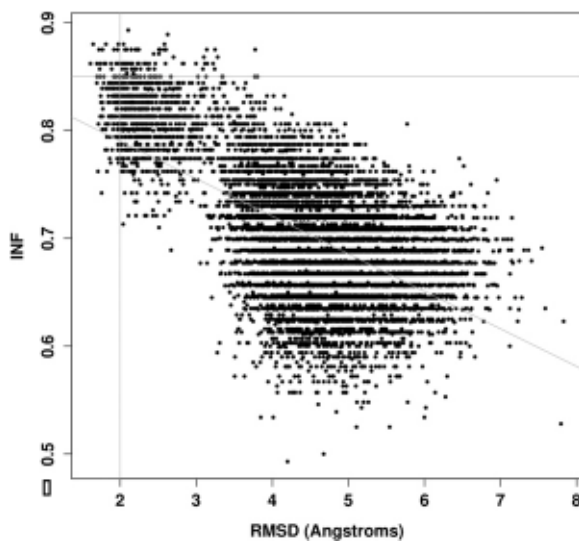


Figure 2.2: Distribution of  $RMSD$  vs.  $INF$  values. Each point corresponds to an individual structure generated with MC-Sym (Parisien and Major, 2008).  $RMSD$  and  $INF$  values were computed by comparison with the crystallographic structure. The oblique line is the linear regression ( $P = 0.6$ ), the horizontal line corresponds to an  $INF$  of 0.85, and the vertical line to an  $RMSD$  of 2.0 ÅRMSD. Adapted from (Parisien et al., 2009)

Model	$RMSD$	$INF$	$DI$
A	1.64Å	0.88	1.86
B	3.76Å	0.88	4.30
C	2.03Å	0.71	2.85

Table 2.1:  $RMSD$ ,  $INF$  and  $DI$  values for three predicted models of rat 28S E-loop

et al., 2010) for molecules of this size with imposed secondary structure, which should be zero:

$$RMSD_{P<0.01} < 5.1 \times 27^{0.41} - 19.8 = -0.1,$$

This example shows the need for considering the specific RNA features (provided by  $INF$  and  $DI$ ) when evaluating predicted models.

## 2.3 Deformation Profile

A comparison metric should also provide meaningful indications about the domains of the predicted model that most contribute to the discrepancy in

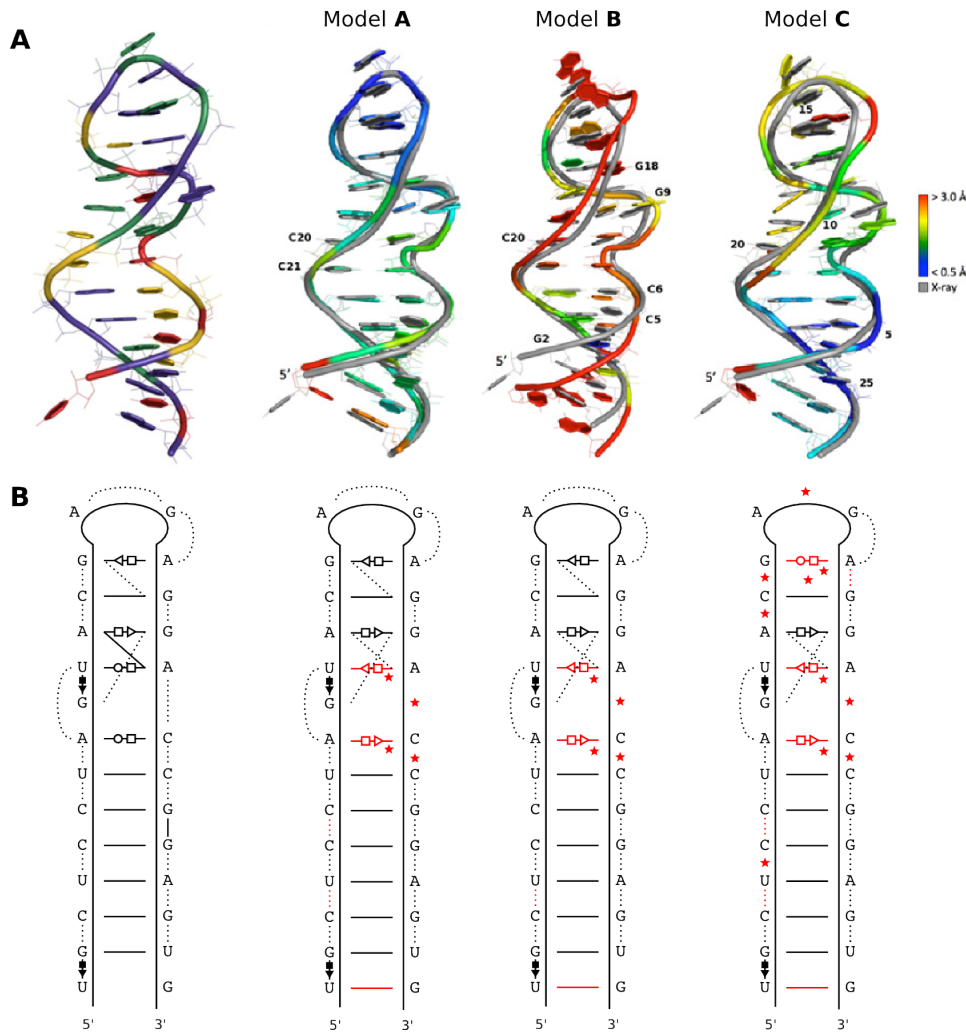


Figure 2.3: Native and predicted models of rat 28S E-loop. (A) The leftmost structure is the crystal structure. The predicted models are shown in colors and the crystal structure in gray (PDB: 1Q9A). Well modeled regions are in blue ( $RMSD < 0.5 \text{ \AA}$ ) and badly modeled regions in red ( $RMSD > 3.0 \text{ \AA}$ ). Model A presents a good  $INF$  (0.88;  $TP=29$ ;  $FP=6$ ;  $FN=2$ ), a good  $RMSD$  (1.64  $\text{\AA}$ ) and a  $DI$  of 1.86. Model B has a good  $INF$  (0.88;  $TP=28$ ;  $FP=5$ ;  $FN=3$ ), but a bad  $RMSD$  (3.76  $\text{\AA}$ ) and a  $DI$  of 4.3. Although the geometry of the base pairs is well modeled, the thread through the phosphate atoms is shifted. Model C has a bad  $INF$  (0.71;  $TP=21$ ;  $FP=7$ ;  $FN=10$ ), but a good  $RMSD$  (2.03  $\text{\AA}$ ) and a  $DI$  of 2.85. The thread through the phosphate atoms is well superimposed, but the base-pairing geometry is wrong. (B) Interaction networks for each of the modules. Interactions in black are True Positives (TP) and in red are False Positives (FP). Red stars correspond to False Negatives (FN). Adapted from Figures 2 and 4 of (Parisien et al., 2009).

respect to the native structure. It is clear that the full complexity of such a comparison cannot be reduced to a “single value” metric. To convey this information we devised the “Deformation Profile” (*DP*) matrix, which is a 2D grid that provides, for each base of the molecule, how well it is predicted in respect to the overall structure.

The formal definition of *DP* is:

$$DP(S, R)_{i,j} = AVG\_DIST(T_{(S_i, R_i)} S_j, R_j),$$

in which  $S, R$  are, respectively, the predicted model and the reference structure,  $S_i, R_i$  are the  $i^{th}$  base of the respective structure,  $T_{(S_i, R_i)}$  is the solid body transformation on  $S$  that superimposes the bases  $S_i$  and  $R_i$  minimizing their *RMSD* and *AVG\_DIST* is given by:

$$AVG\_DIST(A, B) = \frac{\sum_{i=1}^N d(A_i, B_i)}{N},$$

in which  $A$  and  $B$  are two bases,  $A_i$  and  $B_i$  are the  $i^{th}$  atom of each base<sup>7</sup>,  $d$  is the euclidean distance on  $\mathbb{R}^3$  and  $N$  is the number of common atoms between  $A$  and  $B$ . The steps to compute *DP* are illustrated in figure 2.4.

An example of the applicability of the *DP* is given by comparing two predicted models of the hammerhead ribozyme (Dunham et al., 2003)  $S1$  and  $S2$  with the reference crystal structure  $R$  of the ribozyme (PDB: 1NYI) (see Figure 2.5). The models  $S1$  and  $S2$  were produced by the 3D prediction tool MC-Sym (Parisien and Major, 2008) and the two structures present very distinct values of *RMSD*:  $RMSD_{S1} = 3.4$  and  $RMSD_{S2} = 12.2$ .

Applying the *DP* to both molecules we obtain the matrices from Figures 2.6 and 2.7 and the average values for the main domains of  $R$  and  $S2$  in table 2.2.

Surprisingly, the large *RMSD* difference between  $S1$  and  $S2$  given by:

$$RMSD(S2, R)/RMSD(S1, R) = 3.58$$

does not verify for any of the individual domains with a maximum ratio of 1.32 for helices H1 and H2 and even the case of Loop L1 that is better predicted in model  $S2$  than in model  $S1$  (0.87). The structural discrepancy between  $S2$  and  $R$  that justifies such an important *RMSD* difference comes from the interdomain scale as we notice that it is the geometrical relationship between helices H1xH2 and H1xH3 that presents the biggest ratios (2.46 and 3.36 respectively), in the same order of the *RMSD* ratio. The difference

<sup>7</sup>In practice we assume that both bases are of the same nature and compare the atoms of the same name of both bases.

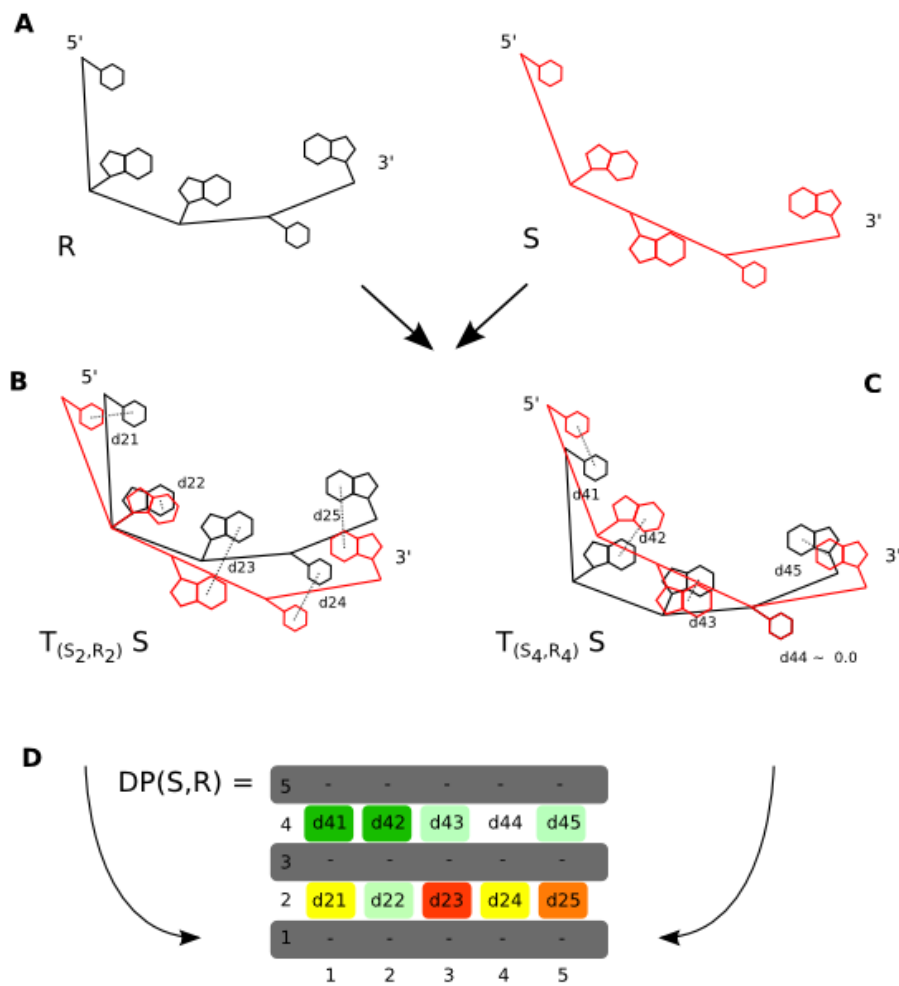


Figure 2.4: Building steps of the deformation profile. (A) The predicted model  $S$  is compared with the native model  $R$ . After superimposing  $S$  on  $R$  minimizing the  $RMSD$  between bases 2 (B) and 4 (C), the average distances between all atoms of the corresponding bases is recorded in the  $DP$  matrix (D). Figure adapted from (Parisien et al., 2009).

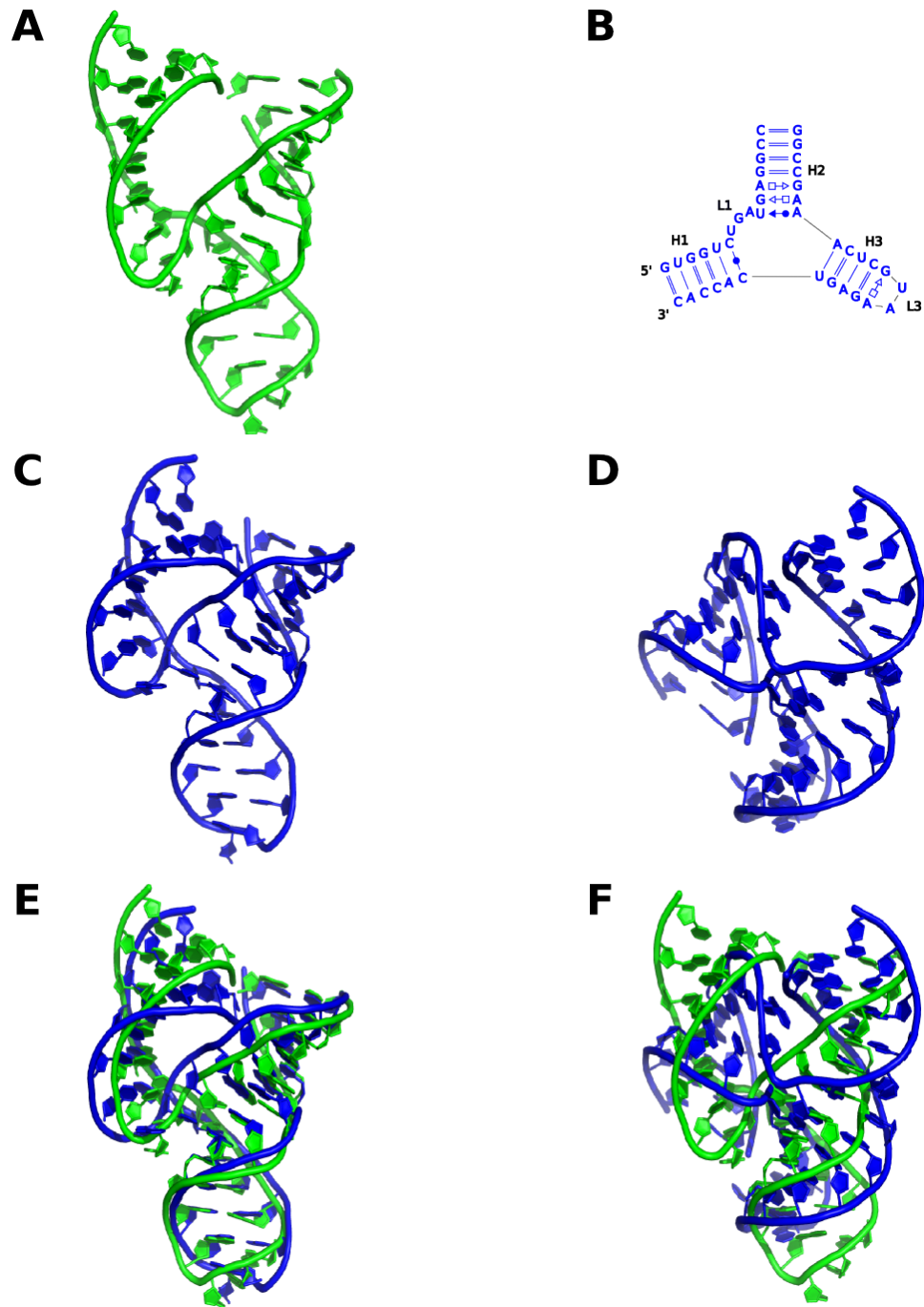


Figure 2.5: (A) Native hammerhead ribozyme (PDB: 1NYI) and its (B) interaction network. (C) Predicted model *S1*. (D) Predicted model *S2*. Superposition of (E) *S1* and (F) *S2* over *R*. Figure adapted from (Parisien et al., 2009)

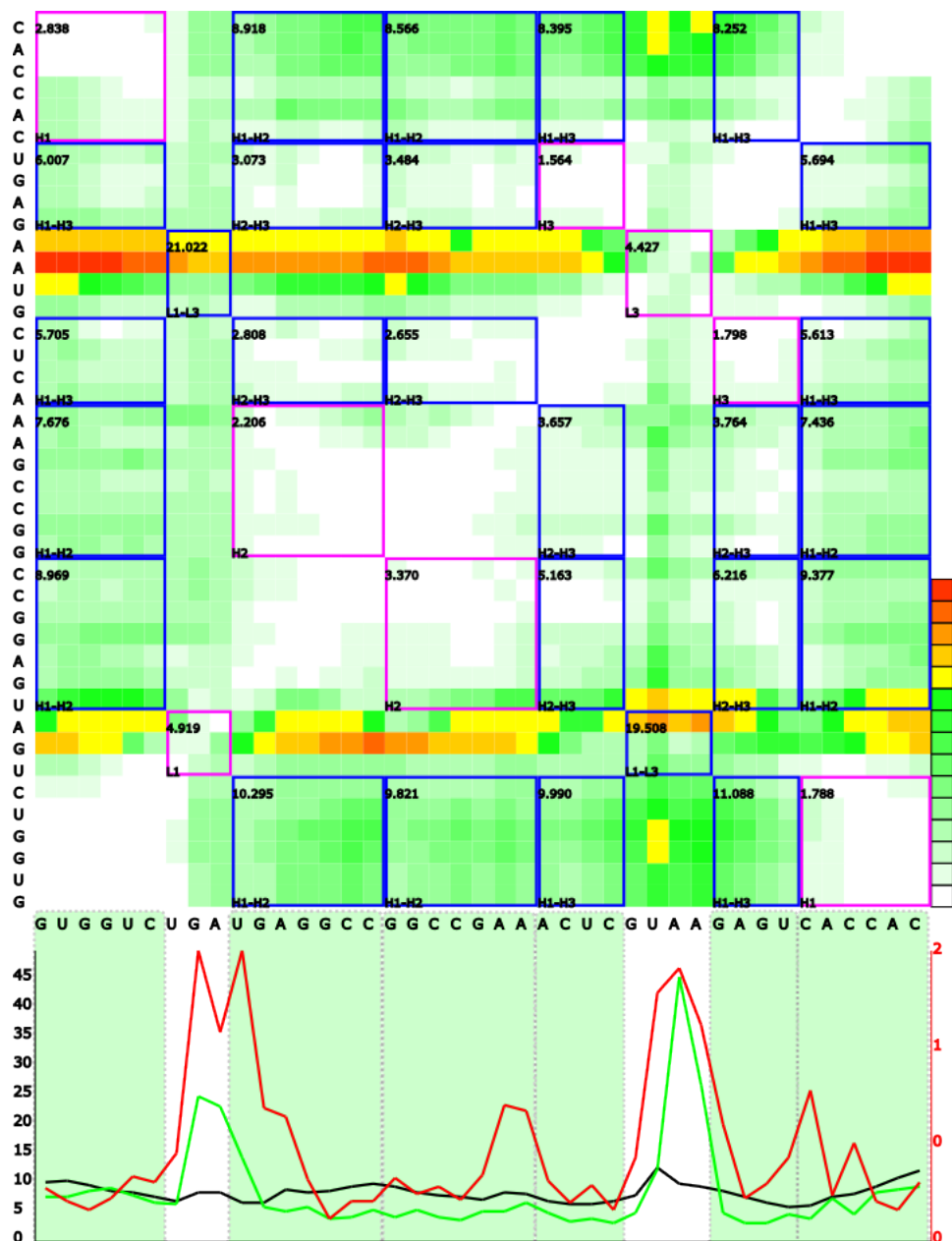
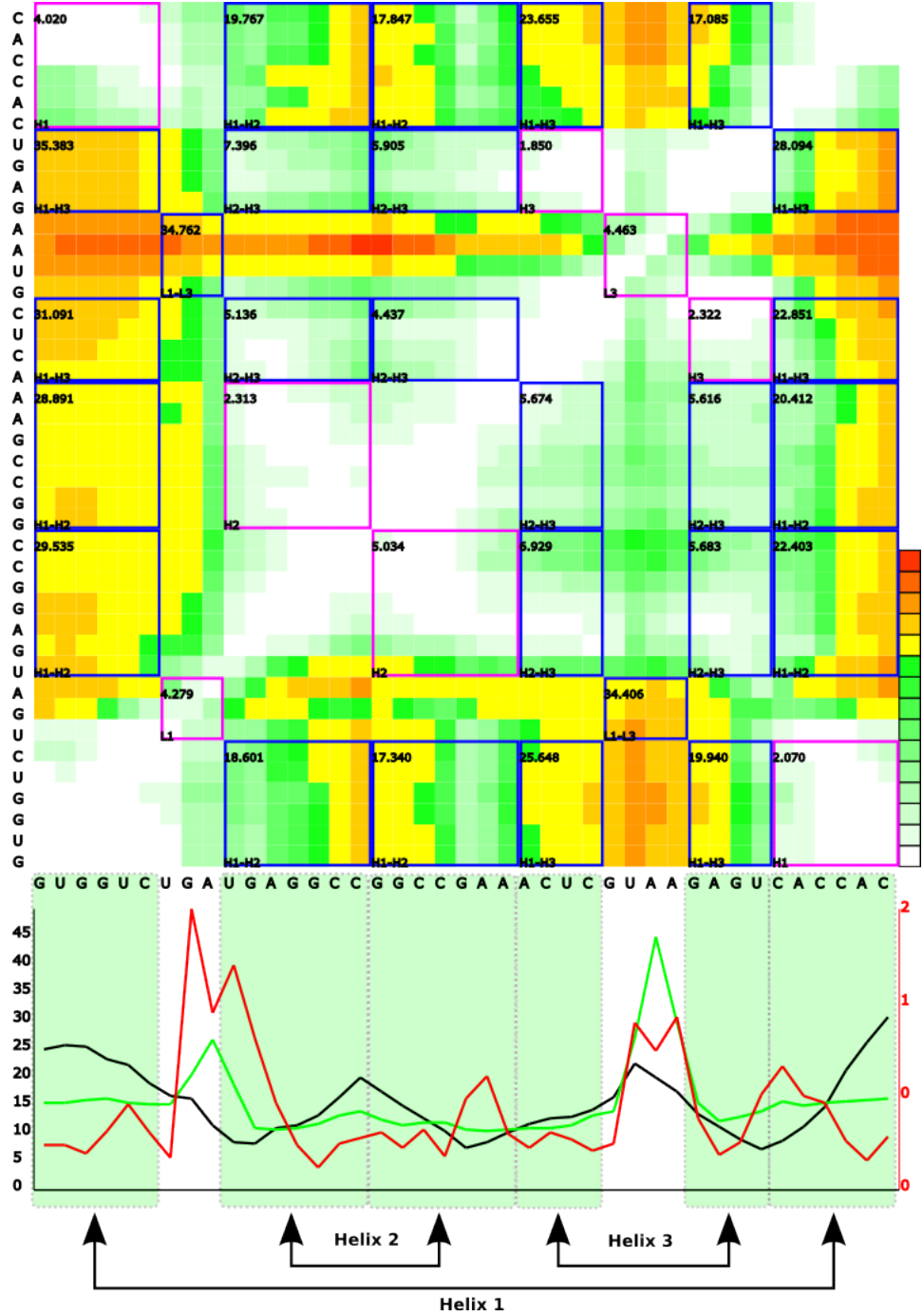


Figure 2.6: Deformation Profile matrix for model S1.

Figure 2.7: Deformation Profile matrix for model *S2*.



<b>Intradomain</b>	<i>S1</i>	<i>S2</i>	<b>ratio</b>
Helix H1	2.31	3.04	1.32
Helix H2	2.79	3.67	1.32
Helix H3	1.68	2.08	1.24
Loop L1	4.92	4.28	0.87
Loop L2	4.43	4.46	1.01
<b>Interdomain</b>	<i>S1</i>	<i>S2</i>	<b>ratio</b>
H1 x H2	8.88	21.85	2.46
H1 x H3	7.59	25.47	3.36
H2 x H3	3.85	5.85	1.52
L1 x L3	20.26	30.54	1.51

Table 2.2: Deformation Profile (*DP*) values and ratios for several intra and interdomain regions of *S1* and *S2* models in comparison with the native structure *R*. The intradomain value of a given domain *D* is the average of all *DP* values of all bases belonging to *D*. The interdomain value between two domains *D1* and *D2* is the average of all *DP* values of all bases from *D1* with respect to the bases of *D2*.

between both models is due to a rotation of nearly 180° degrees of the axis of helix H1. All other domains and interdomain regions are reasonably well predicted in both models – helices H2 and H3 are coaxial (DP ratio 1.52) – or equally badly predicted – L1 x L3 (DP ratio 1.51). This interpretation would be clear from the direct analysis of the structure for a trained eye (see figure 2.5). However, recalling what was already said, it is not possible to analyze by hand the thousands of models generated by the automatic structural prediction.

We believe that both *DI* and *DP*, provide reliable structural comparison for RNA molecules at all scales – nucleotide level, single domain and inter domain – and, as we will try to show in the next chapter, can play a useful role in the validation and study of prediction tools and techniques.

## 2.4 Article – New metrics for comparing and assessing discrepancies between RNA 3D structures and models

This chapter is an extended summary of the following article:

Parisien\*, M., Cruz\*, J. A., Westhof, E., and Major, F. (2009). *New metrics for comparing and assessing discrepancies between RNA 3D structures and models*. RNA 15(10):1875-1885. (\* equal contribution)

# New metrics for comparing and assessing discrepancies between RNA 3D structures and models

MARC PARIEN, <sup>1,3</sup> JOSÉ ALMEIDA CRUZ, <sup>2,3</sup> ÉRIC WESTHOF, <sup>2</sup> and FRANÇOIS MAJOR <sup>1</sup>

<sup>1</sup>Institute for Research in Immunology and Cancer, Department of Computer Science and Operations Research, Université de Montréal, Montréal, Québec H3C 3J7, Canada

<sup>2</sup>Architecture et Réactivité de l'ARN, Université de Strasbourg, Institut de Biologie Moléculaire et Cellulaire du CNRS, 67084 Strasbourg Cedex, France

## ABSTRACT

To benchmark progress made in RNA three-dimensional modeling and assess newly developed techniques, reliable and meaningful comparison metrics and associated tools are necessary. Generally, the average root-mean-square deviations (RMSDs) are quoted. However, RMSD can be misleading since errors are spread over the whole molecule and do not account for the specificity of RNA base interactions. Here, we introduce two new metrics that are particularly suitable to RNAs: the deformation index and deformation profile. The deformation index is calibrated by the interaction network fidelity, which considers base–base–stacking and base–base–pairing interactions within the target structure. The deformation profile highlights dissimilarities between structures at the nucleotide scale for both intradomain and interdomain interactions. Our results show that there is little correlation between RMSD and interaction network fidelity. The deformation profile is a tool that allows for rapid assessment of the origins of discrepancies.

**Keywords:** RNA; structure; comparative analysis; three-dimensional modeling; RMSD

## INTRODUCTION

Determining RNA three-dimensional (3D) structures is key in studying RNA function (Gesteland et al. 2006). Physical methods such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy are the most common ways for determining RNA 3D structures at high resolution. However, these methods cannot be applied to all RNAs and RNA systems. Alternative methods include interactive modeling (Michel and Westhof 1990; Massire and Westhof 1998; Martinez et al. 2008) and conformational space searching (Das and Baker 2007; Ding et al. 2008; Parisien and Major 2008; Jonikas et al. 2009).

The development and improvement of alternative methods are highly dependent on what we learn from experi-

mentally resolved structures. In particular, close inspection of rRNA structures revealed the presence of structural motifs that we can recognize from sequence (Lescoute et al. 2005). To assist the production of new knowledge, systematic methods to annotate RNA 3D structures (Gendron et al. 2001; Lemieux and Major 2002; Yang et al. 2003; Djelloul and Denise 2008), discover and analyze structural motifs (Huang et al. 2005; Lemieux and Major 2006; Lisi and Major 2007; Abraham et al. 2008; Xin et al. 2008), and formally represent RNA structures (Dowell and Eddy 2004; St-Onge et al. 2007) have been developed. This systematization of knowledge generation and integration in ever-improving predictive methods is typical of the post-ribosomal X-ray crystallographic era. A problem that has been largely neglected, however, is how one can measure quantitatively the improvements brought by new approaches or methods.

The classical index for comparing predictive methods is to benchmark with the average root-mean-square deviations (RMSDs) after optimal superimposition between the modeled RNA 3D structures they produce and their corresponding experimental structures. RMSDs are extremely useful, and obtaining models close to experimental structures is a noble exercise. RMSDs capture the general 3D

<sup>3</sup>These authors contributed equally to this work.

**Reprint requests to:** François Major, Institute for Research in Immunology and Cancer, Department of Computer Science and Operations Research, Université de Montréal, Montréal, Québec H3C 3J7, Canada; e-mail: [major@iro.umontreal.ca](mailto:major@iro.umontreal.ca); fax: (514) 343-5839; or Eric Westhof, Architecture et Réactivité de l'ARN, Université de Strasbourg, Institut de Biologie Moléculaire et Cellulaire du CNRS, 15 Rue René Descartes, 67084 Strasbourg Cedex, France; e-mail: [E.Westhof@ibmc.u-strasbg.fr](mailto:E.Westhof@ibmc.u-strasbg.fr); fax: +33 (0)-3-88-41-70-46.

Article published online ahead of print. Article and publication date are at <http://www.majournal.org/cgi/doi/10.1261/rna.1700409>.

shape of an RNA, but give little information about its base-pairing and base-stacking patterns, local deviations of the structure, intradomain deformation, or interdomain deviations. Most importantly, RMSDs spread errors over the whole molecule to obtain the best global superimposition so that it is very difficult to localize the origins of the modeling defects and thus to improve the modeling process (Yang and Honig 2000; Gendron et al. 2001; Shatsky et al. 2002). RNA molecules have specific structural features, such as a modular and hierarchical architecture of structural elements like helices, hairpins, and single-stranded loops connected by tertiary interactions. In addition, RNA bases associate in well-defined patterns of pairings that usually stack on each other. As modeling and predictive methods are getting increasingly accurate, it is now desirable that their results could be compared based on the reproducibility of these important and specific RNA structural features rather than on global average measurements.

Here, we introduce two new RNA 3D structure comparison tools: (1) an RNA 3D structure comparison index, the deformation index (DI), which evaluates and indicates the deviations between two RNA 3D structures with both RMSDs and base interactions; and (2) a deformation profile (DP), which depicts the conformation differences between two models at local, interdomain, and intradomain scales. These new tools provide quantitative measures to compare the accuracy in reproducing the base–base interaction networks of different 3D models, as well as the ability to evaluate the local and global prediction precision and quality of RNA molecules.

## RESULTS

### Deformation index

We define the DI as the RMSD between two optimally aligned 3D structures (general shape) divided by the base interaction network fidelity (INF). The INF is computed from the base-stacking and base-pairing annotations of both structures. For practical reasons, we use two automated annotation procedures that have been proposed recently: MC-Annotate (Gendron et al. 2001; Lemieux and Major 2002) and RNAview (Yang et al. 2003). Note that the index uses, but is not related to, the annotation programs, which are obviously prone to the quality of the reference structures.

### Base-stacking and base-pairing interactions

MC-Annotate detects that two bases stack using the Gabb et al. (1996) method. The base-stacking annotation results are described using the Major and Thibault (2007) nomenclature, which indicates the relative orientation of the two bases. The relative orientation is determined by comparing the direction of the normal vectors of each base, i.e., the

rotational vector obtained by a right-handed axis system defined by atoms N1 to N6 around the pyrimidine ring (Fig. 1A).

Two possible relative orientations in each base result in four base-stacking types: upward ( $\gg$ ), downward ( $\ll$ ), outward ( $\langle \rangle$ ), and inward ( $\rangle \langle$ ) (see Fig. 1B). Two vectors pointing in the same direction (upward and downward) corresponds to the base-stacking type in canonical A-RNA double helices. Upward or downward is chosen depending on which base is referred to first (i.e.,  $A \gg B$  means B is stacked upward of A, or A is stacked downward of B). The two other types are, respectively, inward ( $A \rangle \langle B$ ; A or B is stacked inward of, respectively, B or A) and outward ( $A \langle \rangle B$ ; A or B is stacked outward of, respectively, B or A).

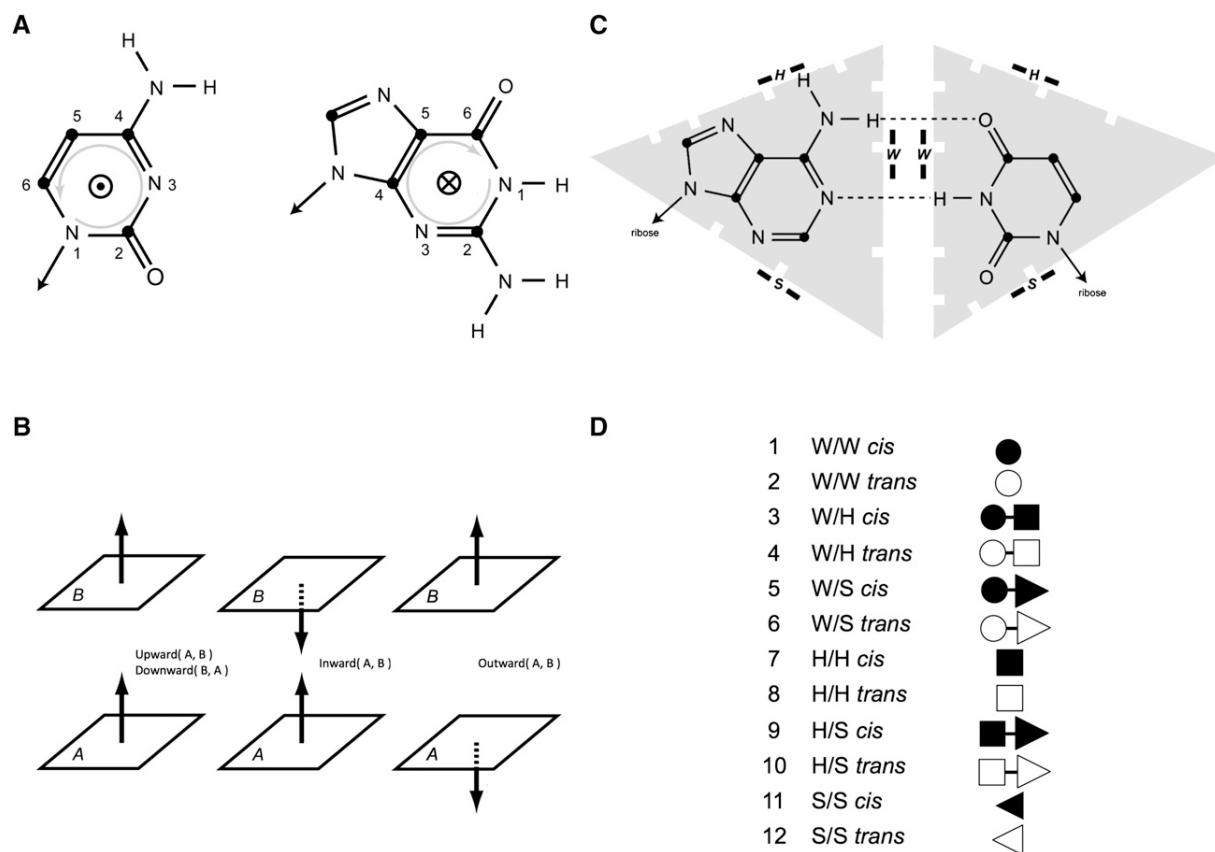
MC-Annotate uses an unsupervised machine-learning approach to detect H-bonds and H-bonding patterns (Lemieux and Major 2002), and RNAview uses geometrical constraints (Yang et al. 2003). Both programs describe their base-pairing annotations using the Leontis and Westhof nomenclature. Each type describes the interacting edge of the two bases. Three interacting edges are defined: the Watson–Crick edge: ● (*cis*), ○ (*trans*); the Hoogsteen edge: ■ (*cis*), □ (*trans*); and the sugar edge: ◀ (*cis*), < (*trans*) (Fig. 1C; Leontis and Westhof 2001). The *cis/trans* notation reflects the relative orientation of the backbone according to the median of the plane formed by the two bases. In Figure 1C, the base pair is *cis* since the riboses are positioned on the same side of the base-pair plane. When two bases interact by the same edge, only one symbol is used. For instance, a *trans* X–Y Hoogsteen base pair is either written “H/H *trans*” or  $X \square Y$ . Figure 1D lists all possible base-pairing types that are described by this nomenclature.

The DI considers the full set of interactions, i.e., base-stacking and base-pairing interactions defined by the classical two-dimensional (2D) structure (A–U and G–C Watson–Crick and G–U Wobble base pairs that form in the stems); extended 2D structures (the noncanonical base pairs, but that can be represented in the dot–bracket notation); and tertiary structure interactions, such as non-helical stacking and long-range base pairs. Note that ~40% of the interactions in crystallized ribosomal RNAs enter the latter category (Stombaugh et al. 2009).

### Interaction network fidelity

A stacking or pairing interaction,  $I$ , involves two distinct nucleotides,  $N_i$  and  $N_j$ ,  $i < j$ , to form an interaction ( $(N_i, N_j, I)$ ), where  $I$  is one of the above base-pairing or base-stacking types. The annotation of a 3D structure produces a set,  $S$ , of such interactions. Given the two sets of interactions in two distinct RNA structures, we can then compare them using simple set theory operations.

Let  $S_r$  be the set of interactions in a reference structure (usually an experimentally resolved structure) and  $S_m$  the set of interactions of a modeled structure. The interactions



**FIGURE 1.** Base-stacking and base-pairing nomenclature. (A) Normal vectors in pyrimidines and purines. Using a right-handed axis system, the normal vector in the pyrimidine (*left*) comes out of the paper plane (atom numbers counterclockwise), whereas it is reversed in the pyrimidine ring of the purine (atom numbers clockwise). (B) The four base-stacking types. Using the normal vectors (represented by arrows), we can distinguish three types of base stacking. If base A is below base B, the normal vector of A points to B, and both normal vectors point in the same direction (*left*), then base B is stacked upward of A (or symmetrically base A is stacked downward of B). If the normal vectors of A and B point toward each other (*middle*), then bases A and B stack inward. If the normal vectors flee each other (*right*), then bases A and B stack outward. (C) Base edges. Each base is divided into three edges: the Watson–Crick (W) edge is at the tip of the base and where the chemical groups involved in Watson–Crick base pairs interact; the Hoogsteen (H) edge is on the opposite side of the ribose; and the sugar (S) edge is on the side of the ribose. Here is a *cis* A–U Watson–Crick base pair, and we write W/W *cis* and represent it using the black dot. The fact that any edge in any base can interact with any other edge in a partner results in six different base–base interactions: W/W, W/H, W/S, H/H, H/S, and S/S. Since there are two possible relative orientations of the ribose according to the place formed by the two bases of a base pair, then this nomenclature describes 12 different base-pairing patterns. (D) The 12 base-pairing patterns, or types, and their associated symbols.

found in the intersection of both sets are true positives,  $TP = S_r \cap S_m$ . The interactions in  $S_m$  that are not present in  $S_r$  are false positives,  $FP = S_m \setminus S_r$ . The interactions absent in  $S_m$  but present in  $S_r$  are false negatives,  $FN = S_r \setminus S_m$ .

The Matthews correlation coefficient (MCC) is estimated by:

$$MCC = \sqrt{PPV \times STY},$$

$$\text{where } PPV(\text{specificity}) = \frac{|TP|}{|TP| + |FP|},$$

$$\text{and } STY(\text{sensitivity}) = \frac{|TP|}{|TP| + |FN|},$$

(Gorodkin et al. 2001). When the model reproduces exactly the base interactions of the reference, then  $|FP| = |FN| = 0$ ,  $|TP| > 0$ , and thus  $MCC = 1$ . When the model does not reproduce any of the interactions of the reference structure, then  $MCC = 0$ , since  $|TP| = 0$ .

We define the interaction network fidelity (INF) between structures A and B as the MCC,  $INF(A,B) = MCC(A,B)$ . We propose a new measure of the resemblance between two structures A and B (for example, a model and its corresponding experimental structure), which is quantified by a deviation index,

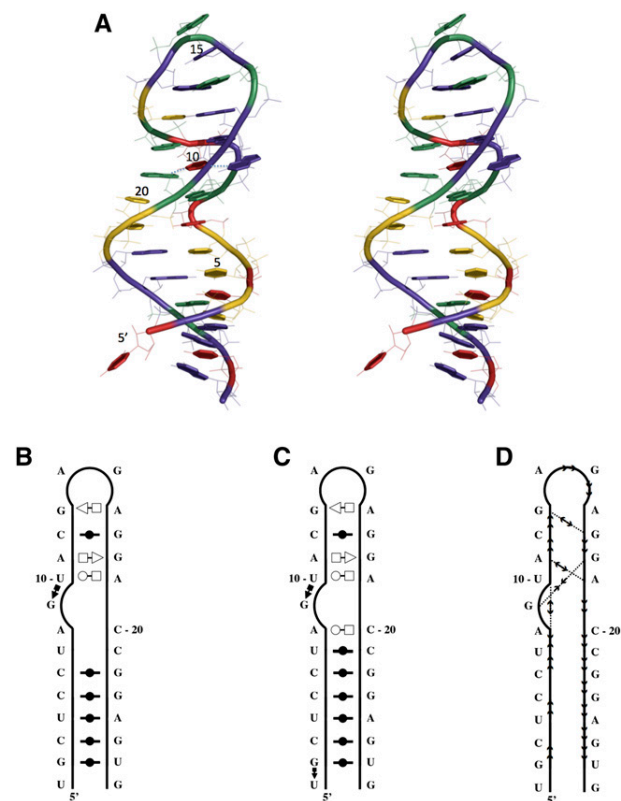
$$DI(A,B) = RMSD(A,B)/INF(A,B).$$

Not having an INF, the DI would simply be the RMSD. However, given an INF from 0 to 1, then the RMSD

between A and B could either have a large (and even infinite) DI if the two structures share no common interactions ( $INF = 0$ ), or meaningful RMSD as  $INF$  approaches 1 (i.e., the majority of the interactions in A are reproduced in B).

### Example: Modeling the rat 28S rRNA loop E 3D structure

Consider the crystal structure of the rat 28S rRNA loop E (PDB code 1Q9A; resolution 1.04 Å; Correll et al. 2003) shown in Figure 2A. MC-Annotate (Fig. 2B) and RNAview (Fig. 2C) were used to compute the base-pairing network of this structure. Since RNA structure annotation is subject to interpretation and small geometrical variations—for instance, MC-Annotate is stricter than RNAview—we therefore take the intersection of both programs. MC-Annotate also computes the base-stacking network (see Fig. 2D).



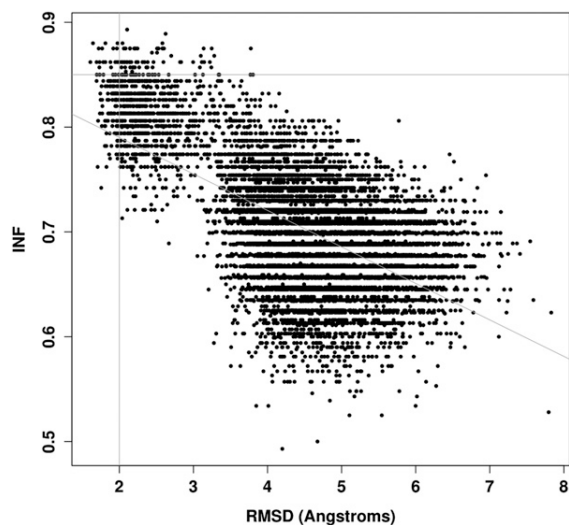
**FIGURE 2.** The rat 28S rRNA loop E structure. (A) Stereoview of the crystal structure (PDB code 1Q9A). (Green) Adenosines, (yellow) cytosines, (violet) guanosines, (red) uracils. The thread through the phosphate atoms is shown using a cylinder. Each base ring is filled and highlighted by thick covalent bonds. The H-bonded bases of the characteristic loop E structure, here the G9-U10-A19 base triple, are linked with dotted lines. Note that U1 in this crystal structure is not paired with G27. The image was generated using Pymol. (B) Secondary structure annotated by MC-Annotate. (C) Secondary structure annotated by RNAview. (D) Stacking annotation.

To illustrate the benchmarking of RNA 3D structure modeling results, we generated loop E 3D structures using MC-Sym (Parisien and Major 2008; see Materials and Methods). We generated a decoy of 9847 3D structures, where each structure is at least at 1 Å RMSD from each other. The RMSDs (all atoms but H) between these structures and the crystal structure range from 1.6 Å to 7.8 Å, whereas the  $INF$  values range from 0.49 to 0.89 (Fig. 3). We note that for a given RMSD threshold, we have a wide range of  $INF$  values, and for a given  $INF$  threshold, we have a wide range of RMSDs. However, as RMSDs worsen, the  $INF$  values also worsen. We note an absence of population in the upper right corner (i.e., high RMSD and high  $INF$  values). The Pearson correlation coefficient between RMSD and  $INF$  values is  $P = 0.60$  for this particular decoy.

For further analyses, we randomly selected three of the MC-Sym-generated structures. Structure A is located in the upper-left corner of Figure 3 and is shown in Figure 4A. This structure has good RMSDs (1.64 Å) with the crystal structure, and good  $INF$  and DI values, 0.88 and 1.86, respectively. Since RMSDs are averaged values, they do not inform about the maximum modeling error. Therefore, we also report the max RMSD( $i, j$ ) ( $j > i$ ), i.e., the maximum RMSD over any sequence fragment defined by  $i$  and  $j$ ;  $j > i$ . If we exclude from the analysis the dangling nucleotide U1 in the crystal structure, the fragment that has maximum RMSD with the crystal structure is C20–C21 with 1.7 Å. This is shown by the fact that C20 and C21 are base paired in the generated structures, as annotated by RNAview, but they have problematic geometries in the crystal structure, as indicated by the absence of annotation by MC-Annotate (Fig. 2B).

Structure A contains 29 TP, i.e., 29 of the 30 base interactions (10 base pairs and 20 base stacks) in the crystal structure. Six FP are made: (1, 2) two upward stacking between C3–U4 and C5–C6. Note that in principle these base-stacking interactions make sense since they are located in a stem. They were not detected in the crystal structure by MC-Annotate; (3) a flip of the C20 base around the glycosidic bond creates an inward stacking A19–C20; (4) as assumed in the modeling, A8–C20 now form a base pair (H/W *trans*); (5) the dangling nucleotide U1 in the models is base paired to G27 as a canonical W/W *cis* type; and (6) as assumed in the modeling, U7–C21 now form a base pair (S/H *trans*). Due to the C20 base flip, the upward stacking A19–C20 and C20–C21 are not reproduced, making two FN.

Structure B was selected in the upper right section of Figure 3, i.e., it has a good  $INF$  (0.88), but a bad RMSD with the crystal structure (3.76 Å). It is shown in Figure 4B. If we remove U1, the worst fragment is G2–C20 (19-nucleotides [nt] long) with 3.66 Å. This is shown in Figure 4B by a shifted backbone in almost all nucleotide positions. Structure B contains 28 of the 30 base interactions (10 base pairs and 20 base stacks) in the crystal structure. Five FP are made. They are the same as in structure A, but the upward stacking between C5–C6 is absent as in the crystal structure.



**FIGURE 3.** Distribution of (RMSD, INF) values. For each MC-Sym generated structure, the RMSD and INF values when compared with the crystal structure are plotted. The oblique line is the linear regression ( $P = 0.6$ ). The horizontal line is at an INF of 0.85, and the vertical line at 2.0 Å RMSD.

The two FN due to the C20 base flip are also present in structure B. In addition, no inward base stacking is detected between G9 and G18.

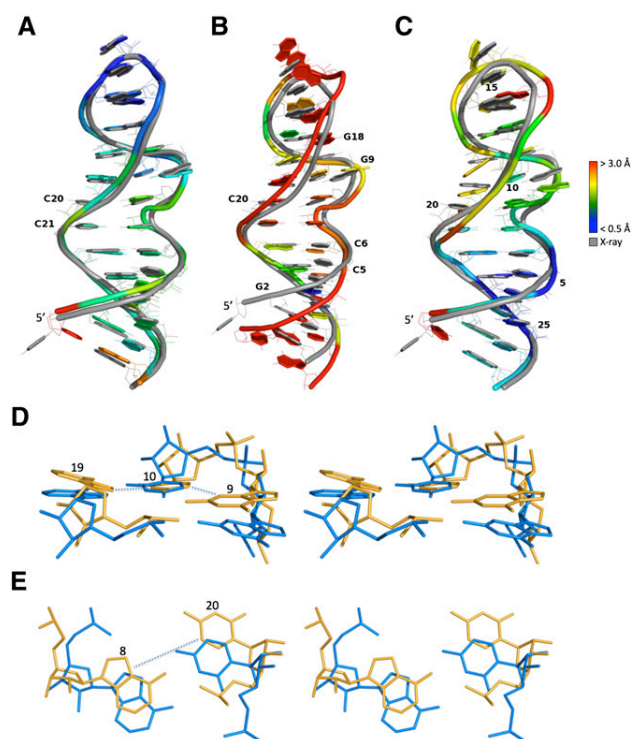
Finally, structure C (Fig. 4C) was selected in the lower left region of Figure 3, i.e., bad INF (0.71), but relatively good RMSD (2.03 Å). Again, the worst fragment is G2–C20, but its RMSD is now 2 Å. What hurts the RMSD of this model is related to difficulties to reproduce the base triple and the A8–C20 base pair of the crystal structure; typical errors in RNA modeling. In our particular case, it is noteworthy that the bases in the generated base triple have a more planar geometry than those observed in the crystal structure (Fig. 4D). As for the A8–C20 base pair, its H/W *trans* type now makes a consensus between MC-Annotate and RNAview. Structure C contains 21 of the 30 base interactions in the crystal structure. Seven FP are made: (1–5) are the same as in structure A; but, in addition, (6) an upward stacking between A16–G17 is detected that was not detected in the crystal structure; (7) the flanking base pair of the GAGA tetraloop, which is changed to a W/H *trans* (S/H *trans* in the crystal). The three FN of structure B are also made in structure C (two are due to the C20 base flip) (Fig. 4E). In addition, four upward stackings are not detected between A11–C12, C12–G13, A14–G15, and U4–C5. The outward stacking between G13–G17 and the G9–U10 S/H *cis* base pair are also not detected. The tenth FN is the absence of the S/H *trans* G13–A16 base pair.

### Deformation profile

The DP is a distance matrix representing the average distance between a predicted model (PM) and reference model

(RM). The DP matrix is obtained by (1) computing all 1-nt superimposition of PM over RM and then (2) for each superimposition, computing the average distance between each base in RM and the corresponding base in PM. Let  $RM_i$  and  $PM_i$  represent the  $i$ th nucleotides of RM and PM respectively, let  $SUP(A_i, B_i)$  be the model that results from the superposition of model B over the reference model A, minimizing the RMSDs between all the atoms of the nucleotides  $A_i$  and  $B_i$ , and let  $AVG\_DIST(A_i, B_i)$  be the average distance between all atoms of the nucleotides  $A_i$  and  $B_i$ . Thus, the *deformation profile* of PM regarding RM is defined as:

$$DP_{ij} = AVG\_DIST[SUP(RM_i, PM_i)_j, RM_j]$$

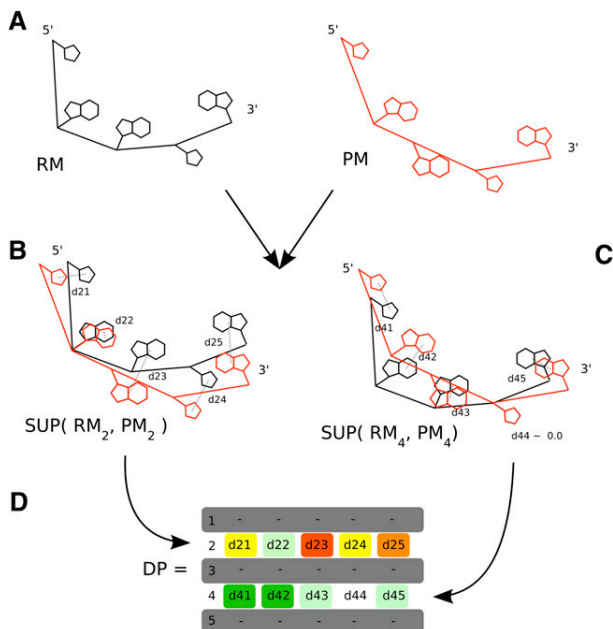


**FIGURE 4.** Three models of the rat 28S rRNA loop E. The models are shown colored and the crystal structure in gray (PDB code 1Q9A). (Blue) Well modeled regions (RMSD < 0.5 Å), (red) badly modeled regions (RMSD > 3.0 Å). The models were optimally aligned (all atoms but H) with the crystal structure. (A) Model with a good INF (0.88; TP 29; FP 6; FN 2) and a good RMSD (1.64 Å); DI = 1.86. (B) Model with a good INF (0.88; TP 28; FP 5; FN 3), but a bad RMSD (3.76 Å); DI = 4.30. Although the geometry of the base pairs is well conserved, the thread through the phosphate atoms is shifted. (C) Model with a bad INF (0.71; TP 21; FP 7; FN 10), but a good RMSD (2.03 Å); DI = 2.85. The thread through the phosphate atoms is well superimposed, but the base-pairing geometry is wrong. Structural features that lead to a bad INF include: (D) base-stacking parameters that differ between the crystal (yellow) and model (blue) structures, such as G9, which shows a high rise in the crystal structure when compared with the model, and A19, for which a tilt can be observed between the crystal and model structures; and (E) base-pairing parameters that differ between the crystal and model structures, such as C20, which flips (propeller twist of 180°) between the crystal and model structures.

Figure 5 illustrates the process of computing a DP matrix.

Once a pair of nucleotides ( $PM_i$ ,  $RM_i$ ) is superimposed, every other pair of nucleotides will be closer or farther depending on how well  $PM_i$  predicts  $RM_i$ . Those average distances are represented in the  $i$ th row of the matrix. Thus, the row average provides information about local similarity regarding the  $i$ th nucleotide. For example, an individual row with higher values than the rest of the matrix (Figs. 6, 7, represented as yellow/red rows in the DP matrices) usually means a particularly poorly predicted nucleotide. The  $j$ th column of the matrix contains the average atomic distances between the  $j$ th nucleotides of PM and RM, for each superimposition. Thus, the column average indicates how the distance between  $PM_j$  and  $RM_j$  depends on the overall prediction of all nucleotides. Finally, the main diagonal contains the average atomic distance of each nucleotide, allowing a perspective of individual nucleotide conformation similarity.

An interesting property of DP is the ability to reveal similarity information at several structural scales. The rectangles corresponding to the intersection of two strands indicate the relative similarity between those strands. This way, one can easily apprehend the structural similarity at intradomain (such as between both strands of a helix or the nucleotides of a loop) and interdomain scales (such as between two helices or two loops).



**FIGURE 5.** Building steps of the deformation profile. (A) A predicted model (PM) will be compared with the reference model (RM). After superimposing PM over RM, minimizing the RMSD between nucleotides 2 (B) and 4 (C), the average distances between all atoms of corresponding nucleotides is calculated and recorded in DP matrix (D).

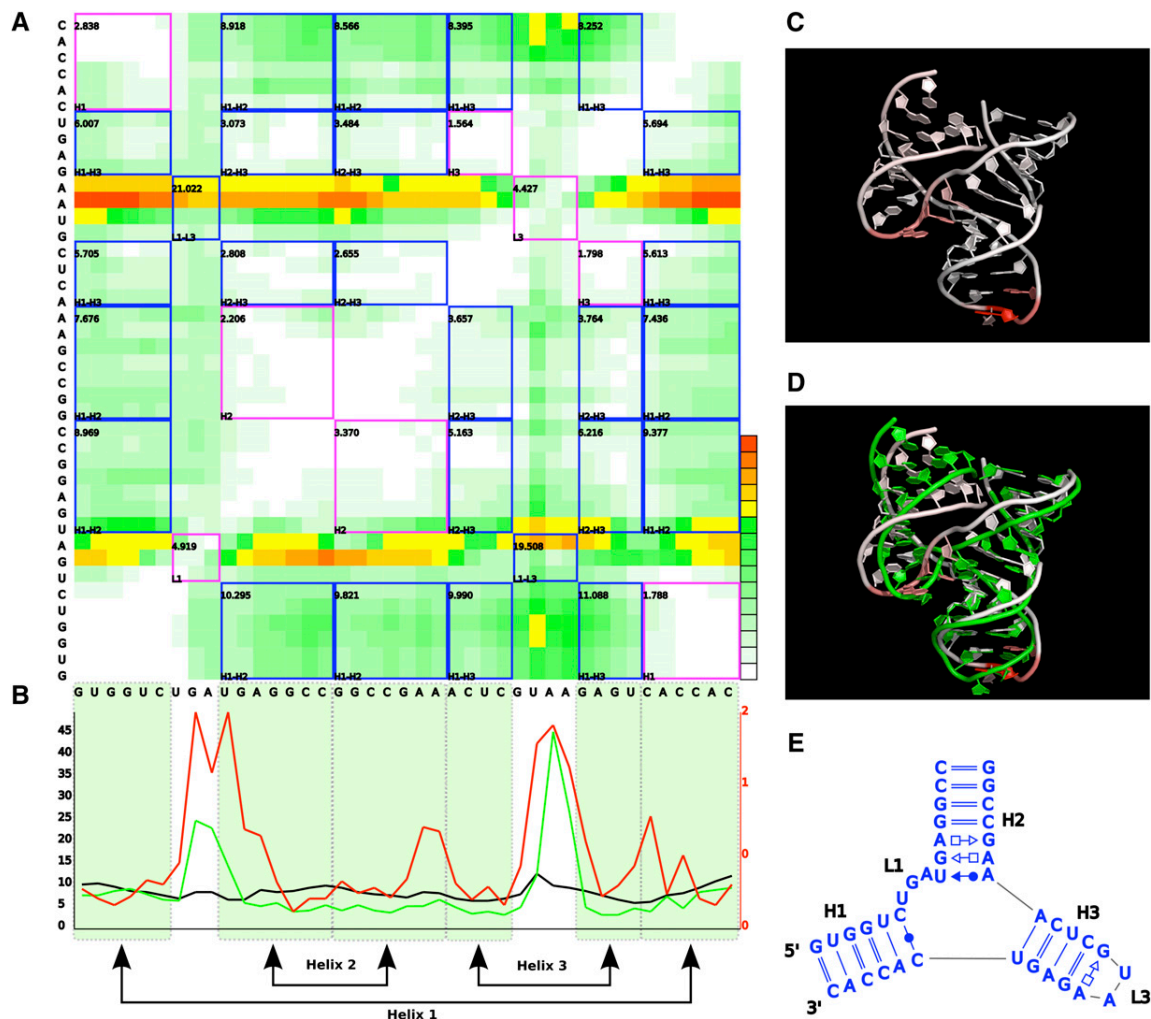
It is worth noticing that values in a DP are not normalized across the whole matrix. Values close to the main diagonal tend to be smaller than values farther away. This is because nucleotide pairs closer from the superimposing pair tend to have smaller average atomic distances than those farther away. Consequently, one should only compare DP values from regions at similar distances to the main diagonal or, obviously, values from DPs of distinct models.

### Example: The hammerhead ribozyme

To exemplify the deformation profile, we compared three predicted models of a hammerhead ribozyme with the reference crystal structure (PDB: 1NYI) (Dunham et al. 2003). We generated a decoy of 9999 3D structures, where each structure is at least at 1 Å RMSD from each other. The RMSDs (all atoms but H) between these structures and the crystal structure range from 2.5 Å to 15.8 Å. Selecting models from decoys is a thorny question. Here, we limited our analysis to a series of structural properties offered by the MC-Pipeline website (see Materials and Methods). We reduced the decoy by performing a five-clustering of the 10,000 models, and selecting one model per cluster that has a small volume (<25,000), a good P-Score (<-15), and to either be bipolar or coplanar (at the >0.7 level) (Laederach et al. 2007). The “thresholds” were established by comparing each structural property with RMSDs to the crystal structure (Supplemental Fig. S1). The selected models and their properties are shown in Table 1.

From the modeling results, we further analyzed models 553, 633, and 2698, the resulting DPs of which are pictured in Figures 6 and 7, and Supplemental Figure S2, respectively. The models share 3.4, 12.2, and 4.9 Å RMSDs with the crystal structure, respectively. The helical regions of the models score fairly well and much better than interhelical and interloop regions (Table 2). Not surprisingly, nucleotides involved in canonical WC base pairing are better predicted than nucleotides involved in noncanonical base pairs or in loops. The 3- and 2-nt-long single-stranded regions (L1 and L3) present the worst deformation score of all short (<5-nt) contiguous regions (Supplemental Fig. S3), except for L3 in model 2698, which was particularly well predicted. The difficulty in predicting L1 and L3 also reflects in the poor prediction of the relative positions of L1 and L3. The main difference between prediction quality among the three models is due to the relative position of helix H1 with respect to the other two helices. Noticeably, the coaxial stacking of helices H2 and H3 was reasonably well predicted in all three models. While model 553 scored well in all helix–helix relative positions, models 633 and 2698 present a displacement of helix H1 regarding H2 and H3. In model 2698, helix H1 is slightly twisted, which significantly penalizes H1×H2 and, to a lesser extent,





**FIGURE 6.** Deformation Profile between predicted model 553 and the hammerhead ribozyme crystal structure. (A) DP matrix. Blue and pink squares inside the matrix correspond to intra- and interdomain similarity relationships, respectively. Numbers in the left top corner of each square are the average value of all positions inside the square. Color scale goes from 0 Å (white) to (but not including) 20 Å (dark green) in 10 equal steps and from 20 Å (yellow) to 80 Å (red) in five equal steps. (B) Average values of rows (green), columns (black), and main diagonal (red) of the matrix. (Shaded green regions) Helical strands. (C) 3D structure of the model. Each nucleotide is colored according to the respective row average value, from minimum (white) to maximum deformation (red) value. (D) Superimposition of the model and reference 3D structures. (E) Interaction network of the original molecule.

H1×H3. In model 633, helix H1 has its double-helical axis rotated by half a turn, pointing in the opposite direction of H1 in the reference molecule, which is reflected in the high values of H1×H2 and H1×H3.

## DISCUSSION

So far, the field of 3D structural modeling has been driven by RMSD comparisons. In particular, GDT-TS (global distance test) is a measure that accounts for the number of atoms that are within 1, 2, 4, and 8 Å of the RMSD from a reference structure (Zemla et al. 1999; Ginalski et al. 2005). A perfect model scores 1.0. Recently, optimal GDT-TS scores of ~0.35 for a tRNA (~75 nt) and 0.20 for the

P4–P6 domain of a group I intron (~150 nt) have been reported (Jonikas et al. 2009). In our study, the optimal score for the hammerhead ribozyme (~40 nt) is 0.68. However, when objectively selecting models from decoys by applying *K*-clustering, GDT-TS scores of 0.20, 0.06, and 0.60 are obtained, respectively. In comparison, protein structure predictions now reach GDT-TS scores as high as 0.75 on average (Zhang 2008). These results highlight the fact that there is a need for improved RNA model selection and generation methods.

RMSD-based measures might be a sufficient criterion for modeling protein structures since the backbone trace is indicative of the structure and correct positioning of the side chains (Dunbrack and Cohen 1997). However, RNA

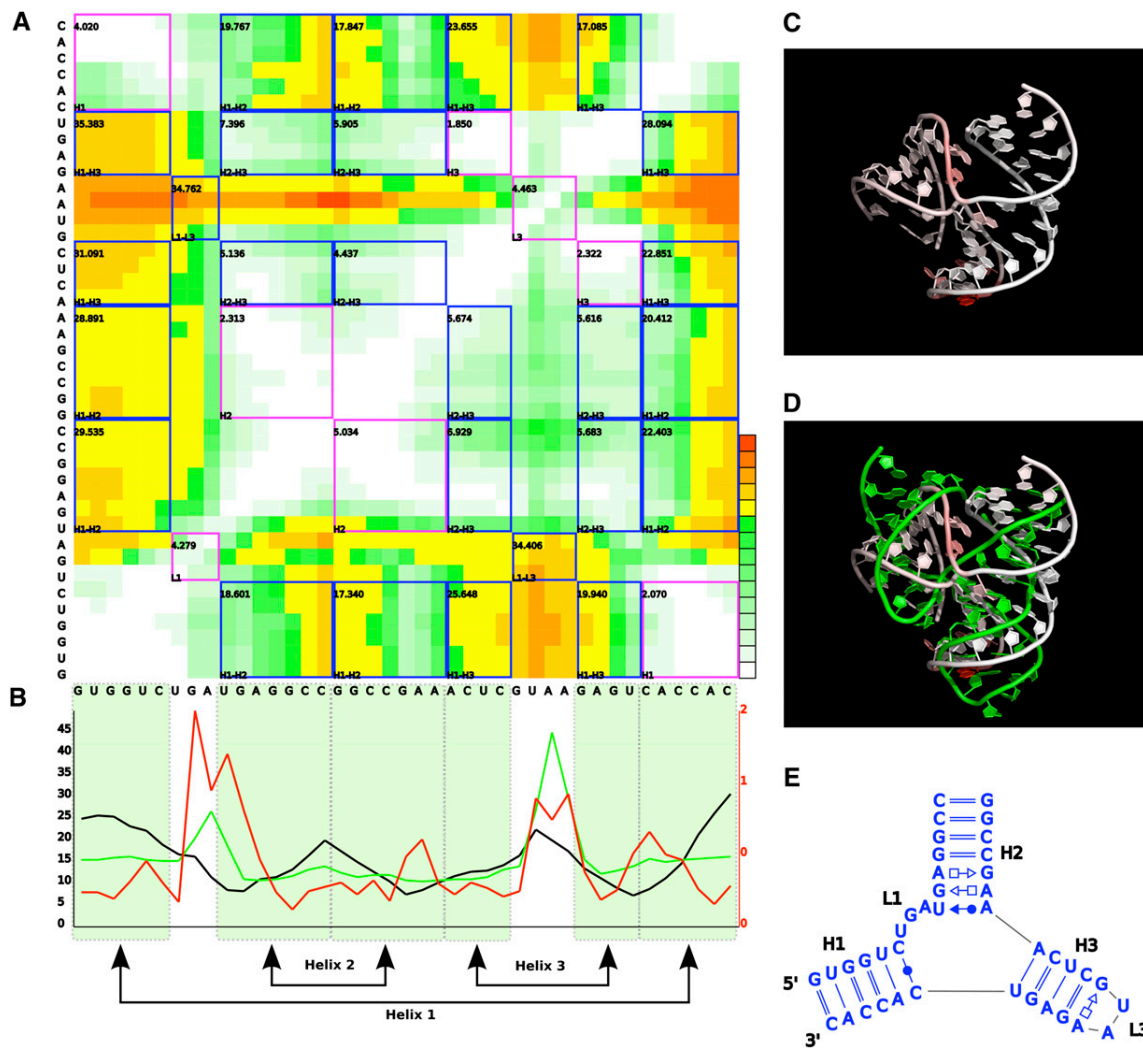


FIGURE 7. Same as Figure 6 but for model 633.

structures contain specific patterns of interacting side chains that are characteristic of folded modules and typical to each overall architecture (Lescoute and Westhof 2006). To evaluate adequately the accuracy of a predicted model, it is key to assess how well such tertiary modules and the non-Watson-Crick base pairs have been reproduced. We show that, in the context of the modeling example we used, the Pearson coefficient between RMSD and INF values ( $P = 0.6$ ) presents little correlation between the two indexes. Our results further show that RMSDs do not provide information about the quality and fidelity of the base interaction network. Besides, the Pearson coefficient for structures with  $\text{RMSD} \geq 3.0 \text{ \AA}$  ( $P = 0.2$ ) is even weaker. These results point to the potential risk of using averaged values such as RMSD in evaluating the quality of RNA 3D models and, thus, the structure prediction methods that generate them. Besides, if the correlation on a small hairpin RNA example is already low, then it is expected to be even lower on larger RNAs.

Besides, the INF is less subject to variations than RMSD for an RNA under thermal motion (Grishaev *et al.* 2008). Intrahelical distortions include: collective atomic motion resulting in slight helix twisting that rarely affect base-base interactions (Fig. 4B), and relative atomic motion that is handled by discretizing the base-base interactions using symbolic annotation (Gendron *et al.* 2001; Leontis and Westhof 2001; Lemieux and Major 2002; Yang *et al.* 2003). Interhelical disposition from thermal motion affects the angle between helices, which greatly affects atomic distances and thus RMSD. However, such changes in general concern only a small fraction of the base-base interactions, and thus do not affect much the INF (Table 1).

In the structure prediction field, models  $< 3 \text{ \AA}$  of RMSD from an experimental structure are considered accurate. Our results suggest extreme prudence at this particular value, since in our test case the INF value of such models can be as low as 0.7. In our example, structure C has an INF

**TABLE 1.** Structural parameter values for five models of the hammerhead ribozyme

Model <sup>a</sup>	Bipol <sup>b</sup>	Copl <sup>b</sup>	Rand <sup>b</sup>	RMSD	P-Sc <sup>c</sup>	Vol <sup>d</sup>	INF <sup>all e</sup>	INF <sup>bp e</sup>	GDT-TS <sup>f</sup>	Cluster
553	0.83	0.05	0.11	3.4	-23.6	23,635	0.82	0.90	0.60	2
633	0.80	0.12	0.08	12.2	-26.0	23,861	0.87	0.94	0.15	1
2698	0.81	0.12	0.07	4.9	-21.0	24,900	0.84	0.89	0.38	4
3778	0.84	0.05	0.12	12.2	-20.6	24,338	0.86	0.92	0.15	3
6870	0.76	0.08	0.16	13.9	-16.5	23,599	0.79	0.89	0.09	5

<sup>a</sup>“Model” represents one model per cluster (Cluster) selected from the results of a “five-clustering.”

<sup>b</sup>Bipolar (Bipol), coplanar (Copl), and random (Rand) are measurements against the RMSD. These parameters describe the field of nucleobase normal vectors, which have been shown to be highly organized in solved RNA structures (Laederach et al. 2007). A threshold at 0.7 for the bipolar scores corresponds to a low RMSD (see Supplemental Fig. S1).

<sup>c</sup>The P-Score (P-Sc) against the RMSD measures the A-RNA likelihood of the phosphate chain—measured using the probabilities of valence angles of three consecutive atoms and the torsion angles of four consecutive atoms. The probabilities, P, are converted in pseudo-energies, E, using the Boltzmann relation:  $E = -RT \log(P)$ .

<sup>d</sup>Approximated ellipsoidal volume (Vol) against the RMSD. The volume is computed as described by Hao et al. (1992). A threshold at 25,000 corresponds to a low RMSD (see Supplemental Fig. S1).

<sup>e</sup>The INF values over base pairing and base stacking (INF<sup>all</sup>) and base-pairing interactions alone (INF<sup>bp</sup>).

<sup>f</sup>Global distance test (GDT-TS) values measure the average percentage of atoms within 1, 2, 4, and 8 Å from the target structure (Zemla et al. 1999; Ginalski et al. 2005). The higher the value, the better the model compared with the target structure.

of 0.71. This structure, despite 21 TP, also had seven FP and 10 FN. If we look between 3 and 5 Å of RMSD, then INF values can be as low as 0.5; with a wider range of INF values (0.5–0.9) located at or near 4 Å of RMSD. Clearly, assessing the quality and accuracy of any given RNA 3D model needs both the RMSD and INF values.

Capturing the dissimilarity between two structures in a single value, as does RMSD, is a practical way of assessing the accuracy of predicted models. However, a single value cannot provide enough information about the shape of the actual structure and the local dissimilarities. Understanding the contribution of individual domain—nucleotides, helices, single-stranded regions—to an overall dissimilarity score demands the intervention of a human expert, which is not compatible with the analysis of dozens or hundreds of candidate models produced by automatic prediction tools. The proposed deformation profile provides a compact representation of RNA model dissimilarities from nucleotide length to intradomain scales and can be used in complement to the DI to fully assess the quality of predicted models.

Consequently, a full quantification of the comparison between two RNA 3D structures should include the overall RMSD, max RMSD( $i, j$ ), INF, as well as the DI. If only one value is to be used, then the DI is the most significant one since it reflects the overall features encoded by the RMSD calibrated by the quality of the reproduced interaction network, which is encoded by the INF value. As the size of modeled RNAs increases, the importance of using both quantifiers increases as well since the correlation between RMSD and INF values is expected to decrease. Finally, phosphate or backbone atom-only, as well as canonical base-paired region-only RMSD, should be avoided since they are not indicative of the quality of the produced models, and the field has now made sufficient progress in RNA 3D modeling and prediction methods so that all-atom models are now the gold standard.

## MATERIALS AND METHODS

### Generating MC-Sym decoys

To generate a decoy for the Loop E, we produced an MC-Sym script from the dot-bracket notation supported by the RNAview annotated secondary structure, “(((((((.....((((.....)))))))))))).” The Dot2Sym program is an MC-Tool to generate MC-Sym input scripts from dot-bracket notations (see Supplemental Information). Note that no base-pairing type information is used, and MC-Sym in such a case attempts all consistent base-pairing types.

For the hammerhead ribozyme, we also obtained a first script from Dot2Sym using the following dot-bracket input: “(((((((.....((((.....))))))))((((.....)))))))).” The script was manually edited and can be found in the provided Supplemental Information. We reduced the 10,000 structure decoys to a list of five models using the five-clustering and the following SQL query:

**TABLE 2.** Intradomain and interdomain scores for all helices, loops, helix–helix, and loop–loop combinations

Intradomain	Model 553	Model 633	Model 2698
Helix H1	2.31	3.04	3.04
Helix H2	2.79	3.67	3.89
Helix H3	1.68	2.08	2.03
Loop L1	4.92	4.28	4.72
Loop L3	4.43	4.46	1.18
Interdomain	Model 553	Model 633	Model 2698
H1 × H2	8.88	21.85	13.25
H1 × H3	7.59	25.47	9.10
H2 × H3	3.85	5.85	6.49
L1 × L3	20.26	34.54	13.13

The intradomain score of domain D is the average of all positions ( $i, j$ ) of the Deformation Profile where both nucleotides  $i$  and  $j$  belong to D. The interdomain score of domains D1 × D2 is the average of all positions ( $i, j$ ) and ( $k, l$ ) of the deformation profile where nucleotides  $i$  and  $k$  belong to D1 and  $j$  and  $l$  belong to D2.

```
SELECT * FROM BKOiY0dM2m T1 INNER JOIN (SELECT
MIN(PScore) AS minP, Cluster FROM BKOiY0dM2m WHERE
((Bipolar >= 0.7) OR (Coplanar >= 0.7)) and Volume <= 25000
and PScore <= -15 GROUP BY Cluster) T2 ON T1.PScore =
T2.minP and T1.Cluster = T2.Cluster WHERE T1.Volume <=
25000
```

See the MC-Sym FAQ (<http://www.major.irc.ca/MC-Sym/faq.html>), commands.html page generated by MC-Sym, and the MC-Pipeline website for details (<http://www.major.irc.ca/MC-Pipeline>). The 3D structures were visualized and rendered using Pymol (DeLano 2002).

## RMSD

RMSD values were for all-atom but H, as computed using the MC-RMSD program. MC-RMSD is part of the MC-Tools, which are available from the authors.

## Deformation profile

All the data processing, PDB file manipulation, and superimposition used to compute the Deformation Profile were done in Python using Bio.PDB (<http://biopython.org>) and NumPy (<http://numpy.scipy.org/>). The script to produce DP matrices is available from the authors.

## SUPPLEMENTAL MATERIAL

Supplemental material can be found at <http://www.rnajournal.org>.

## ACKNOWLEDGMENTS

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) to F.M. F.M. is a member of the Centre Robert-Cedergren of the Université de Montréal. M.P. holds a Ph.D. scholarship from the NSERC. J.A.C. is supported by the Ph.D. Program in Computational Biology of the Instituto Gulbenkian de Ciência, Portugal (sponsored by Fundacao Calouste Gulbenkian, Siemens SA, and Fundacao para a Ciencia e Tecnologia; SFRH/BD/33528/2008).

Received April 23, 2009; accepted July 10, 2009.

## REFERENCES

- Abraham M, Dror O, Nussinov R, Wolfson HJ. 2008. Analysis and classification of RNA tertiary structures. *RNA* **14**: 2274–2289.
- Correll CC, Beneken J, Plantinga MJ, Lubbers M, Chan YL. 2003. The common and the distinctive features of the bulged-G motif based on a 1.04 Å resolution RNA structure. *Nucleic Acids Res* **31**: 6806–6818.
- Das R, Baker D. 2007. Automated de novo prediction of native-like RNA tertiary structures. *Proc Natl Acad Sci* **104**: 14664–14669.
- DeLano WL. 2002. *The PyMOL molecular graphics system*. DeLano Scientific, San Carlos, CA.
- Ding F, Sharma S, Chalasani P, Demidov VV, Broude NE, Dokholyan NV. 2008. Ab initio RNA folding by discrete molecular dynamics: From structure prediction to folding mechanisms. *RNA* **14**: 1164–1173.
- Djelloul M, Denise A. 2008. Automated motif extraction and classification in RNA tertiary structures. *RNA* **14**: 2489–2497.
- Dowell RD, Eddy SR. 2004. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics* **5**: 71–85. doi: 10.1186/1471-2105-5-71.
- Dunbrack RL Jr, Cohen FE. 1997. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci* **6**: 1661–1681.
- Dunham CM, Murray JB, Scott WG. 2003. A helical twist-induced conformational switch activates cleavage in the hammerhead ribozyme. *J Mol Biol* **332**: 327–336.
- Gabb HA, Sanghani SR, Robert CH, Prevost C. 1996. Finding and visualizing nucleic acid base stacking. *J Mol Graph* **14**: 6–11, 23–24.
- Gendron P, Lemieux S, Major F. 2001. Quantitative analysis of nucleic acid three-dimensional structures. *J Mol Biol* **308**: 919–936.
- Gesteland RF, Cech TR, Atkins JF, eds. 2006. *The RNA world*. CSHL Press, Cold Spring Harbor, NY.
- Ginalski K, Grishin NV, Godzik A, Rychlewski L. 2005. Practical lessons from protein structure prediction. *Nucleic Acids Res* **33**: 1874–1891.
- Gorodkin J, Stricklin S, Stormo G. 2001. Discovering common stem-loop motifs in unaligned RNA sequences. *Nucleic Acids Res* **29**: 2135–2144.
- Grishaev A, Ying J, Canny MD, Pardi A, Bax A. 2008. Solution structure of tRNA<sup>Val</sup> from refinement of homology model against residual dipolar coupling and SAXS data. *J Biol NMR* **42**: 99–109.
- Hao MH, Rackovsky S, Liwo A, Pincus MR, Scheraga HA. 1992. Effects of compact volume and chain stiffness on the conformations of native proteins. *Proc Natl Acad Sci* **89**: 6614–6618.
- Huang H-C, Nagaswamy UMA, Fox GE. 2005. The application of cluster analysis in the intercomparison of loop structures in RNA. *RNA* **11**: 412–423.
- Jonikas MA, Radmer RJ, Laederach A, Das R, Pearlman S, Herschlag D, Altman RB. 2009. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA* **15**: 189–199.
- Laederach A, Chan JM, Schwartzman A, Willgohe E, Altman RB. 2007. Coplanar and coaxial orientations of RNA bases and helices. *RNA* **13**: 643–650.
- Lemieux S, Major F. 2002. RNA canonical and noncanonical base pairing types: A recognition method and complete repertoire. *Nucleic Acids Res* **30**: 4250–4263.
- Lemieux S, Major F. 2006. Automated extraction and classification of RNA tertiary structure cyclic motifs. *Nucleic Acids Res* **34**: 2340–2346.
- Leontis NB, Westhof E. 2001. Geometric nomenclature and classification of RNA base pairs. *RNA* **7**: 499–512.
- Lescoute A, Westhof E. 2006. The interaction networks of structured RNAs. *Nucleic Acids Res* **34**: 6587–6604.
- Lescoute A, Leontis NB, Massire C, Westhof E. 2005. Recurrent structural RNA motifs, isostericity matrices, and sequence alignments. *Nucleic Acids Res* **33**: 2395–2409.
- Lisi V, Major F. 2007. A comparative analysis of the triloops in all high-resolution RNA structures reveals sequence structure relationships. *RNA* **13**: 1537–1545.
- Major F, Thibault P. 2007. RNA tertiary structure prediction. In *Bioinformatics: From genomes to therapies* (ed. T Lengauer), Vol. I, pp. 491–539. Wiley-VCH, Weinheim, Germany.
- Martinez HM, Maizel JV Jr, Shapiro BA. 2008. RNA2D3D: A program for generating, viewing, and comparing three-dimensional models of RNA. *J Biomol Struct Dyn* **25**: 669–683.
- Massire C, Westhof E. 1998. MANIP: An interactive tool for modeling RNA. *J Mol Graphics Modell* **16**(4-6): 197–205, 255–257.
- Michel F, Westhof E. 1990. Modeling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J Mol Biol* **216**: 585–610.
- Parisien M, Major F. 2008. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **452**: 51–55.
- Shatsky M, Nussinov R, Wolfson HJ. 2002. Flexible protein alignment and hinge detection. *Proteins* **48**: 242–256.

- St-Onge K, Thibault P, Hamel S, Major F. 2007. Modeling RNA tertiary structure motifs by graph-grammars. *Nucleic Acids Res* **35**: 1726–1736.
- Stombaugh J, Zirbel CL, Westhof E, Leontis NB. 2009. Frequency and isostericity of RNA base pairs. *Nucleic Acids Res* **37**: 2294–2312.
- Xin Y, Laing C, Leontis NB, Schlick T. 2008. Annotation of tertiary interactions in RNA structures reveals variations and correlations. *RNA* **14**: 2465–2477.
- Yang AS, Honig B. 2000. An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J Mol Biol* **301**: 665–678.
- Yang H, Jossinet F, Leontis N, Chen L, Westbrook J, Berman H, Westhof E. 2003. Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res* **31**: 3450–3460.
- Zemla A, Venclovas C, Moulton J, Fidelis K. 1999. Processing and analysis of CASP3 protein structure predictions. *Proteins Suppl* **3**: 22–29.
- Zhang Y. 2008. Progress and challenges in protein structure prediction. *Curr Opin Struct Biol* **18**: 342–348.

## Chapter 3

# RNA Puzzles

Since the publication of the first complete atomic resolution of an RNA structure – the X-Ray Crystal structure of the yeast tRNA<sup>Phe</sup> at 2.7 Å (Sussman et al., 1978) – X-Ray crystallography remains today the single method to determine the structure of full RNA molecules at atomic resolution (less than 3 Å). The lengthy process required to obtain the structure of an RNA molecule by the crystallographic process and the importance of those structures to the understanding of the molecular function motivate the search for alternative ways of structure determination. Experimentally RNA folds can also be obtained using NMR (Nuclear Magnetic Resonance) and Cryo-electron microscopy, although in both cases the resolution has not the same meaning as the crystallographic resolution. The ability to predict the tertiary structure of an RNA molecule based only on its sequence information is a long time appealing goal. The illusory simplicity of this goal is eloquently expressed in the provocative statement of Tinoco and Bustamante:

“(...) If 10% of protein fold researcher[s] switched to RNA, the problem [of RNA folding prediction] could be solved in one or two years (...)” (Tinoco and Bustamante, 1999)

Even if fully automatic RNA structure prediction is still a hope, important theoretical advances in the last decades took us closest to it:

- The development of predictive models for RNA secondary structure (Tinoco et al., 1973) currently available in a number of tools (Hofacker et al., 1994; Zuker, 2003; Reuter and Mathews, 2010).
- The comparative sequence analysis modeling (Michel and Westhof, 1990).
- The systematization of the knowledge about RNA architecture and interactions (Leontis and Westhof, 2001).

- The rapid increase in the number of RNA structures (Berman et al., 2000). and sequence alignments (Gardner et al., 2009) available in public databases.

Today, the acknowledged importance of ncRNAs in molecular processes pushed a number of research groups to approach the RNA structure prediction problem. The current RNA prediction methodologies start to produce the first practical results for molecules larger than a few nucleotides and several recent results foreshadow exciting developments in this field.

As it was true for the protein structure field 17 years ago (Moult et al., 1995), the RNA structure prediction community needs now an independent and continuous comparison of the different methodologies and their obtained results. In order to fulfill this need we developed **RNAPuzzles**, a collective experiment of *de novo* RNA structure prediction as a first approach to compare RNA structure prediction algorithms in a blind and independent fashion.

### 3.1 RNA Tertiary Structure Prediction

Tertiary structure prediction of RNA molecules consists in determining the 3D atomic coordinates of an RNA molecule from its sequence. The native structure – most often unknown — is commonly referred to as the “target” structure and the predicted structure as the “model”. A number of prediction approaches have been proposed. They can be characterized by the following factors:

- **Degree of automatism:** Prediction tools automatism ranges from the computer assisted RNA modeling to the fully automatic prediction. In the lower end of automatism the model is built by the human expert assisted, in a more or less intelligent way, by the computational tool. In the opposite end, fully automatic tools produce an RNA model with no human intervention other than the input data and parameter settings.
- **Homology information:** Some prediction tools can take advantage of available homologous structures to predict new models. In these cases the tool will replace mutated nucleotides and propose new structures for the additional domains not present at the homologous structure. This approach can be particularly effective if sequence divergence is limited or if an important number of homologous structures is available (e.g. tRNAs,rRNAs) which is still, unfortunately, a rare case.
- **Additional input information:** Beyond the sequence information, some prediction tools can incorporate additional information to the prediction process such as: the known (or expected) secondary structure, chemical probing data, single molecule kinetic data, the expected

positions of structural models, ... Again, these data are not always available or fully reliable.

- **Resolution:** Although the produced models represent the atomic coordinates of the RNA molecule, these atomic coordinates are some times deduced from lower resolution models also known as coarse grained models. Coarse grain models are useful simplifications that avoid the full complexity of an atomic detail model presenting, however, a lower fidelity representation of the reality.

No current algorithm or methodology is able to single handedly predict with full accuracy an RNA structure. Because of that each prediction tool combines a set of different techniques hoping to take the most out of each one and to complement their shortcomings. The following techniques are some of the most commonly used:

- **Fragment assembly:** Combines known RNA fragments to build a full model. The choice of the fragments to use depends on sequence similarity and local structural compatibility such as known local conformation, recurrent structural models, or base-pair isosteric substitutions. The fragment database needs to be as large as possible in order to be representative of the conformational space. Fragment assembly works well for small molecules or localized structure prediction, but it has difficulties in predicting long-range interactions.
- **Conformation sampling:** Samples the RNA conformation space of the target sequence in order to find the “best” model (e.g. the most energetically favorable, the one with no steric clashes, the one that better fits some previous structural constrains, ...). This technique can be very useful to improve a pre-computed model by searching the nearest conformations for “better” models. The huge size of the RNA conformational space, however, renders impractical any systematic use of it without added information.
- **Molecular dynamics (MD):** Simulates the expected behavior of the nucleotide chain in its physicochemical context – solvation, ionic context, partner molecules, ligands, ... – It requires force field models derived from experimental work. MD can be an effective approach for *de novo* prediction (e.g. starting from a linear nucleotide chain and with no additional information), but it is computationally intensive, and it strongly depends on the accuracy of the force field and, as any other numerical simulation, is subject to numerical instability.
- **Reduced chain representation:** Uses a simplified model of the nucleotide chain, such as, one bead per nucleotide or a three beads model (each bead representing the backbone, sugar and residue atoms).



- **Discrete molecular dynamics (DMD):** Similar to MD but uses the reduced chain representation approach to represent the molecule.
- **Modular Fragments Assembly:** This is a computer assisted model building technique that consists in identifying the structural modules possibly present in the molecule and assembling them in a final model. The regions of the molecule for which no module could be predicted, usually single stranded regions, should be defined by the user. A modular fragments assembly software will help the user in producing the automatically predicted models, assembling them in a coherent 3D structure, refining the coordinates in order to avoid stereo chemical clashes.

Table 3.1 on page 55 presents a list of currently available 3D prediction tools with some of their characteristics.

## 3.2 RNA Puzzles

An independent comparison between existing prediction methods can be useful to potential users as it provide information regarding the most appropriate methods for each class of problems, but also to the prediction community as it helps to establish the state-of-the-art on prediction methods and motivates new developments by direct comparison of the strengths and shortcomings of each method.

The protein community has a long history on such type of comparison. Since 1994 and every two years the “**Critical Assessment of protein Structure Prediction**” (**CASP**) experiment (Moult et al., 1995) gathers the protein structure prediction community. In the last **CASP** experiment **CASP9** 2010, 129 targets were released and more than 61000 prediction models were received<sup>1</sup>.

Although the RNA structure prediction community is much smaller than the protein homologue, the current number of research groups, their enthusiasm and the significant advances in their obtained results convinced us of the opportunity and importance of organizing a **CASP**-like experiment for RNA prediction: the **RNAPuzzles**.

Another very important aspect is the sharp increase on the number of solved crystal structures in the few last years (Figure 3.1). Without the efforts and kind collaboration of the crystallography groups this experiment would not be possible.

The mechanics of **RNAPuzzles** is pretty straightforward:

---

<sup>1</sup>Data collected from <http://www.predictioncenter.org/casp9/numbers.cgi>.

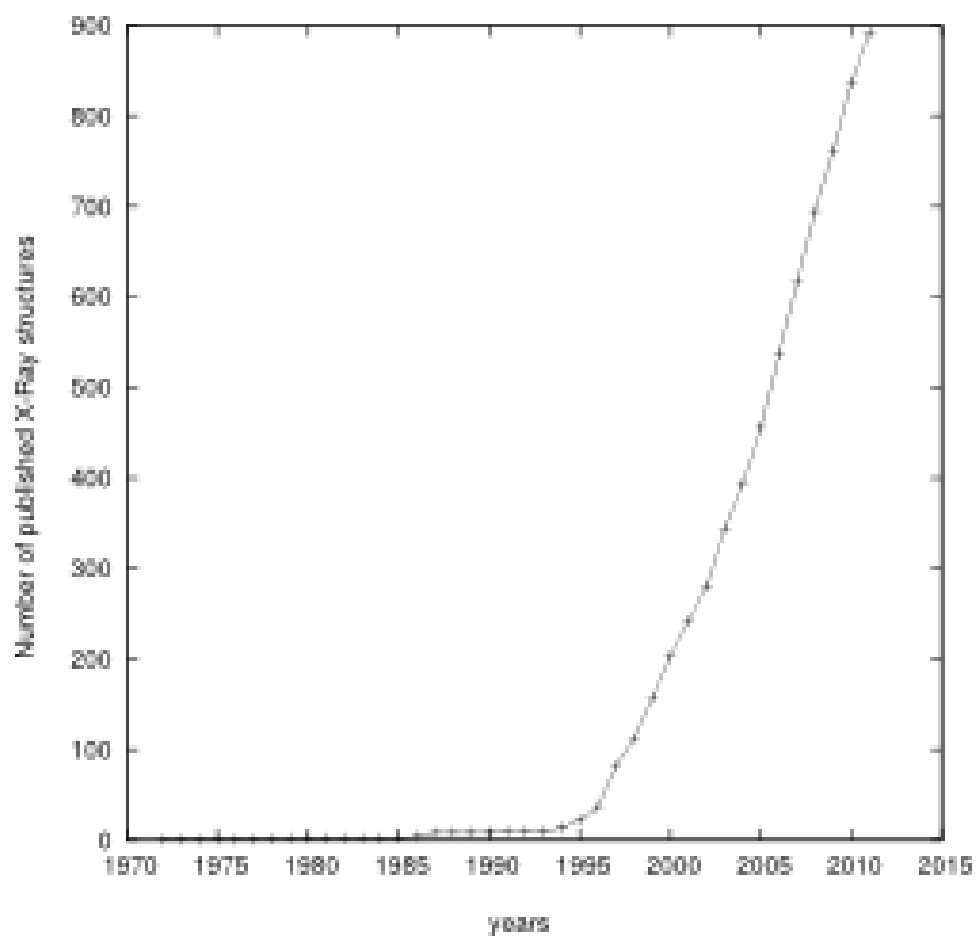


Figure 3.1: The picture shows the sharp increase on the number of newly published X-Ray structures (source: [www.pdb.org](http://www.pdb.org)).

1. **Target release:** The complete nucleotide sequences of unpublished X-Ray structures from a RNA molecule (kindly provided by an experimental group) is released to all of the interested groups;
2. **Prediction:** The participants submit their predicted models until a predefined date;
3. **Evaluation:** The predicted models are evaluated and compared with the target structures;
4. **Publication:** After the publication of the original X-ray structures the predicted models and comparison data are published.

The model evaluation is done in terms of stereochemical correctness and geometrical similarity with the experimental structure. The stereochemical correctness evaluation indicates if the atomic distances implied by the predicted model are compatible with the known H-bonds and van der Waals contact distances. It is measured as the `clash-score` given by the stereochemical evaluation tool `MolProbity` (Davis et al., 2007). To evaluate the geometrical similarity we computed the usual RNA structure comparison metrics `RMSD`, `Deformation Index`, `Deformation Profile` and `P-value` as described in the previous chapter (see Chapter 2).

### 3.2.1 Structure Comparison Pipeline

Despite of the conceptual simplicity of the `RNAPuzzles` mechanics, a number of technical issues had to be solved to allow for the automatic treatment of the predicted models and the publication of the results. In the first place all processes must be automatized as it would be impractical to manually treat all of the submitted models, the number of which we expect to grow in future rounds. As a result of this automatic process, the submitted data must be normalized in order to be processed by the different evaluation and visualization tools.

To solve this problem we developed a validation pipeline that performs all normalization, evaluation and data visualization steps with a minimal human intervention. The figure 3.2 depicts the developed pipeline.

### 3.2.2 RNA Puzzles Web Site

Finally, all the effort invested in this project only makes sense if the results are published and available to the public. We prepared a public web site (<http://paradise-ibmc.u-strasbg.fr/rnapuzzles/>) where all the information regarding the challenges, solutions, predicted models, respective evaluations and participants can be freely accessed. Figure 3.3 summarizes the published web site.

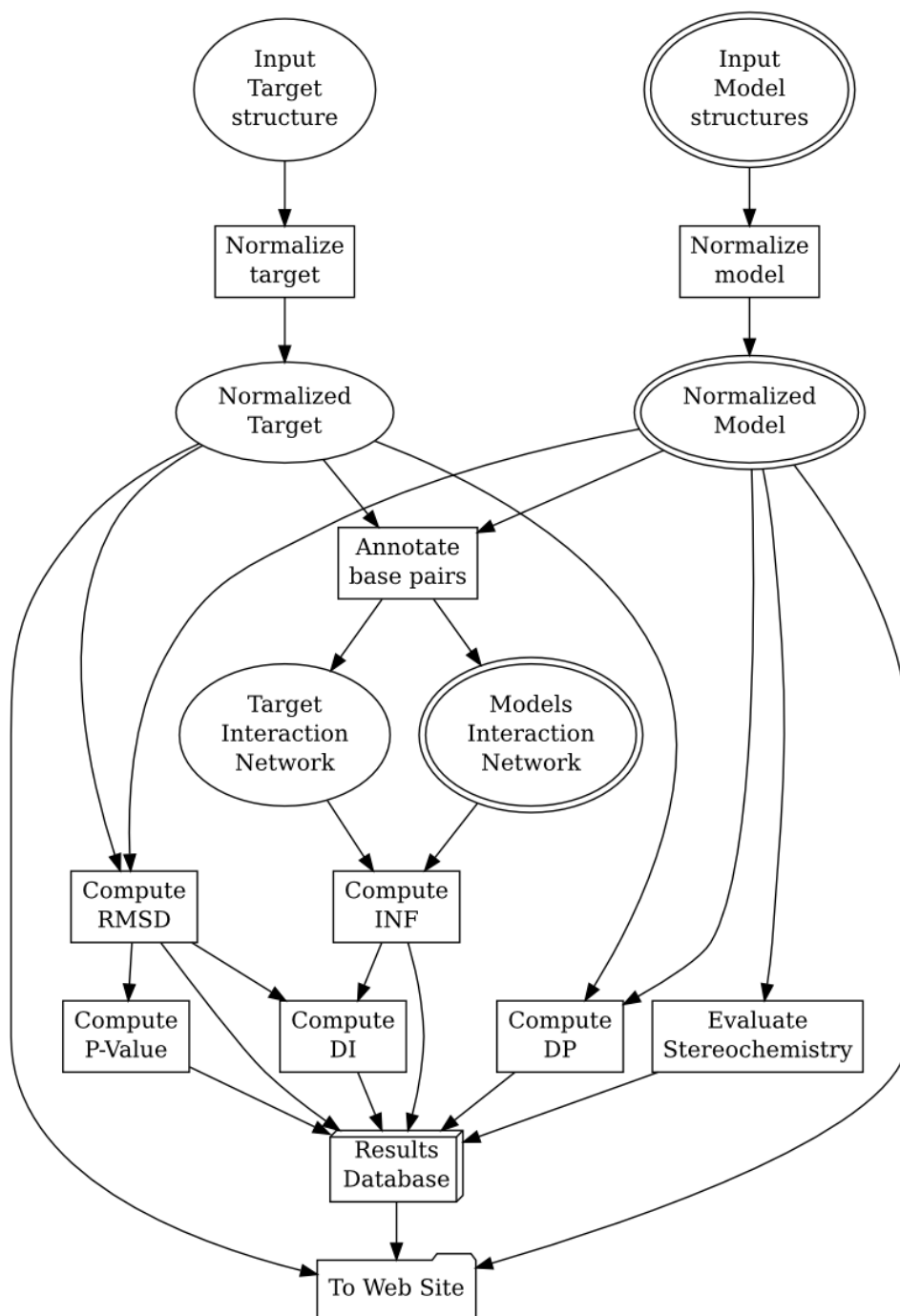
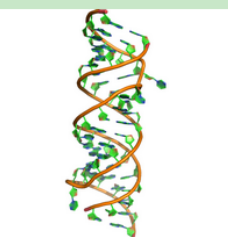


Figure 3.2: RNAPuzzles processing pipeline.

# RNA Puzzles

[Home](#) | [Open Challenges](#) | [Past Challenges](#) | [The Participants](#)



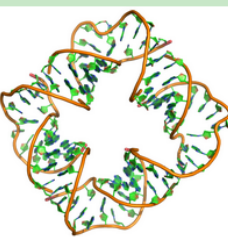
### Challenge 1 (November 2011)

What is the structure of the following sequence: 5' - CCGCCGCGCCAUGCCUGGGCGG - 3' knowing that the crystal structure shows a homodimer that contains two strands of the sequence. The strands hybridize with blunt ends (C-G closing base pairs).

Crystal structure kindly provided by Thomas Hermann:

Dibrov SM, McLean J, Hermann T. 2011. Structure of an RNA dimer of a regulatory element from human thymidylate synthase mRNA., *Acta Cryst. D, Biological Crystallography* 67, 97-104.

[results >](#)



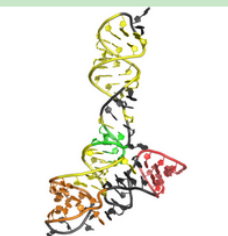
### Challenge 2 (November 2011)

The crystal structure shows a 100 nt square that assembles from four inner and four outer strands. The secondary structure shown was used for the design of the square. Actual base pairing in the crystal may deviate. 3D coordinates of the nucleotides in the inner strands (B,D,F,H) were provided. What are the structures of the outer strands (A,C,E,G)?

Crystal structure kindly provided by Thomas Hermann:

Dibrov SM, McLean J, Parsons J, Hermann T. 2011. Self-assembling RNA square. *PNAS* 108, 6405-6408.

[results >](#)



### Challenge 3 (November 2011)

A domain of a riboswitch was crystallized. The sequence is the following: 5' - CUCUGGAGAGAACC G U U U A A U C G G U C G C G A A G G A G C A A G C U C U G C G C A U A U G C A G A G U G A A C U C U C A G G C A A A G G A C A G A G - 3'

The crystallized sequence was slightly different (an apical loop was replaced by a GAAA loop) but it was not mentioned to protect the crystallographers.

Crystal structure kindly provided by Dinshaw Patel:

Huang L, Serganov A, Patel DJ. 2010. Structural Insights into Ligand Recognition by a Sensing Domain of the Cooperative Glycine Riboswitch. *Mol. Cell.* 40, 774-786.

[results >](#)

More informations: [e.westhof \[AT\] ibmc-cnrs.unistra.fr](mailto:e.westhof@ibmc-cnrs.unistra.fr)  
 Last update May 20th, 2011

Figure 3.3: RNAPuzzles web site. The web site is available at (<http://paradise-ibmc.u-strasbg.fr/rnapuzzles/>)

### 3.3 Conclusions

At the moment of writing the first round of experiments was concluded with 3 target RNA sequences released, 7 participating research groups and 39 submitted models.

We believe that this first round of the **RNAPuzzles** successfully showed the current state of the RNA structure prediction field: The quality of the best predicted structures was remarkably good; The ability to predict the correct local conformation of the most common structures was patent (e.g., for a fairly complex riboswitch target the correct 3 way junction architecture correctly predicted). Some of the still open challenges on RNA structure prediction were clearly pinpointed such as the prediction of (i) single stranded regions, (ii) long range interactions, (iii) global architecture of the molecule and (iv) non WC interactions.

The complete description of the results and its detailed analysis can be found in (Cruz et al., ).

Twelve years separate us from the motivational article from Tinoco and Bustamante. The RNA structure community is still probably less than 10% of the protein structure community, nevertheless, the understanding of the RNA folding and the structure prediction capabilities are steadily progressing. We hope **RNAPuzzles** could contribute somehow to this progress.

<b>Tool</b>	<b>Automatism</b>	<b>Input</b>	<b>resolution</b>	<b>methodology</b>
Paradise (Jossinet et al., 2010)	semi automatic	homology secondary structure	atomic	computer assisted
RNABuilder (Flores and Altman, 2010)	automatic	homology secondary structure	atomic	molecular dynamics
FARNA (Das and Baker, 2007)	automatic	-	atomic	fragment assembly conformation sampling coarse grained potentials
MC-Sym (Parisien and Major, 2008)	automatic	secondary structure	atomic	fragment assembly conformation sampling
iFoldRNA (Sharma et al., 2008)	automatic	secondary structure	coarse-grained	discrete molecular dynamics
ModeRNA (Rother et al., 2011)	semi automatic	secondary structure	atomic	computer assisted scripting automation
Vfold (Cao and Chen, 2005)	automatic	-	coarse-grained	reduced chain thermodynamic based MD
RNA2D3D (Martinez et al., 2008)	semi automatic	secondary structure	atomic	computer assisted molecular dynamics

Table 3.1: RNA structure prediction tools.

### 3.4 Article – A CASP-like Evaluation of *de novo* structure predictions

This chapter is an extended summary of the following article:

Cruz JA, Boniecki M, Bujnick JM, Chen SJ, Cao S, Das R, Ding F, Dokholyan NV, Flores SC, Lavender CA, Major F, Mikolajczak K, Philips A, Puton T, Santalucia J, Hermann T, Rother K, Rother M, Serganov A, Skorupski M, Soltysinski T, Sripakdeevong P, Tuszynska I, Weeks KM, Waldsich C, Wildauer M, Leontis NB, Westhof E (2011). *A CASP-like evaluation of RNA 3D structure predictions.* in preparation.



# A CASP-like evaluation of RNA 3D structure predictions

José Almeida Cruz (1), Michal Boniecki(2), Janusz M. Bujnick (2,3), Shi-Jie Chen (4), Song Cao (4), Rhiju Das (5), Feng Ding (6), Nikolay V. Dokholyan (6), Samuel Coulbourn Flores (7), Christopher A. Lavender (8), François Major (9), Katarzyna Mikolajczak(2), Anna Philips(2,3), Tomasz Puton(3), John Santalucia (10), Thomas Hermann (11), Kristian Rother(3), Magdalena Rother(3), Alexander Serganov (12), Marcin Skorupski(3), Tomasz Soltysinski(2), Parin Sripakdeevong (5), Irina Tuszynska(2), Kevin M. Weeks (8), Christina Waldsich (13), Michael Wildauer (13), Neocles B. Leontis (14), Eric Westhof (1)\*

(1) *Architecture et Réactivité de l'ARN, Université de Strasbourg, IBMC-CNRS, 15 r. R. Descartes, F-67084 Strasbourg, France*

(2) *Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology ul. Ks. Trojdena 4, 02-109 Warsaw, Poland*

(3) *Laboratory of Structural Bioinformatics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, ul. Umultowska 89, 61-614 Poznan, Poland*

(4) *Department of Physics and Department of Biochemistry, University of Missouri, Columbia, Missouri 65211, USA*

(5) *Department of Biochemistry and Department of Physics, Stanford University, Stanford CA 94305, California, USA*

(6) *Department of Biochemistry and Biophysics, University of North Carolina, School of Medicine, Chapel Hill, North Carolina, USA*

(7) *Computational & Systems Biology Program, Institute for Cell and Molecular Biology, Uppsala University, Uppsala, Sweden*

(8) *Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina 27599-3290.*

(9) *Institute for Research in Immunology and Cancer (IRIC), Department of Computer Science and Operations Research, Université de Montréal, PO Box 6128, Downtown Station, Montréal, Québec H3C 3J7, Canada*

(10) *Department of Chemistry, Wayne State University, Detroit, Michigan 48202, USA*

(11) *Department of Chemistry and Biochemistry, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA*

(12) *Structural Biology Program, Memorial Sloan-Kettering Cancer Center, New York, New York 10021, USA*

(13) *Max F. Perutz Laboratories, Department of Biochemistry, University of Vienna, Dr. Bohrgasse 9/5, Vienna 1030, Austria*

(14) *Department of Chemistry and Center for Biomolecular Sciences, Bowling Green State University, Bowling Green, OH 43402, USA*

\*Corresponding author.

## **Abstract**

We report the results of a first, collective, blind experiment in RNA three-dimensional structure prediction. The goal is to assess the leading edge of RNA structure prediction techniques, compare existing methods and tools, and evaluate their relative strengths, weaknesses, and limitations in terms of sequence length and structural complexity. The results should give potential users insight into the suitability of available methods for different applications and facilitate efforts in the RNA structure prediction community in their efforts to improve their tools. We also report the creation of an automated evaluation pipeline to facilitate the analysis of future RNA structure prediction exercises.

## Introduction

The determination of the atomic structure of any biological macromolecule, RNA molecules being no exception, contributes regularly towards the molecular understanding of the molecular basis of the underlying biological process. Each of the current experimental methods for determining the three dimensional (3D) structure of an RNA molecule, X-ray crystallography, NMR and cryo-electron microscopy, require great expertise and substantial technical resources. Therefore, the ability to reliably predict accurate structures of RNA molecules based solely on their sequences, on in concert with efficiently obtained biochemical information, is an important problem and constitutes a major intellectual challenge [Tinoco 1999]. Recent decades have seen a number of significant theoretical advances towards this goal and include: (i) The development of predictive models for RNA secondary structure, pioneered by the seminal work of Tinoco and co-workers [Tinoco 1973] and made commonly available in a number of tools that perform reasonably well for sequences of moderate size [Hofacker 1994, Zuker 2003, Reuter and Mathews 2010]. (ii) The ability to meaningfully deduce RNA structures through comparative sequence analysis [Michel and Westhof 1990]. (iii) The systematization of the knowledge about RNA architecture and interactions [Leontis and Westhof 2001] to gain a handle on the rapid increase in the number and size of RNA molecules with published structures available in public databases [Berman 2000]. (iv) The availability of comprehensive sequence alignments [Gardner 2009] permitting the study of the relationship between structure and sequence. (v) The development of improved molecular dynamics force fields and techniques [Ditzler 2010]. Finally, (vi) the increasing availability of inexpensive computing power and data storage. As a consequence, exciting developments in the field of *de novo* structure prediction have occurred in the last few years: Computer-assisted modeling tools [Martinez 2008, Jossinet 2010], conformational space search [Parisien and Major 2008], discrete molecular dynamics [Ding 2008], knowledge-based, coarse grained refinement [Jonikas 2009], template-based [Flores 2010, Rother 2011] and force field based approaches [Das 2010] inspired by proven protein folding techniques adapted to the RNA field (review: [Rother 2011b]). All these new approaches are pushing the limits of automatic RNA structure prediction from short sequences of a few nucleotides to medium sized molecules with several dozens. Assuming the continuation of a steady progress, one could expect, in the near-to-medium future, that *de novo* prediction of RNA 3D structures will become as common and useful as RNA secondary structure prediction is today.

These promising results and the increasing number of available tools raise the need for objective evaluation and comparison. Indeed, the establishment of a benchmark for RNA structure prediction has become essential in order to optimize and improve the current methods and tools for structural prediction. Here, we present the results of a blind exercise in RNA structure prediction. Sequences of RNA structures solved by crystallographers were submitted, before publication, to active research groups that develop new methods and perform RNA 3D structure prediction. Comparisons between predicted and experimental x-ray structures were undertaken once the structures were published. The resulting benchmarks function as a snapshot of the current status of this field. On the basis of this successful first round, we would like to extend the idea established from the protein structure prediction community [Moult 2006] to RNA and to propose a continuous, open and collective structure prediction experiment, with the essential active participation of experimentalists.

## **RNA Puzzles**

*RNAPuzzles* is a collective blind experiment for *de novo* RNA structure prediction evaluation. With this initiative, we hope to (i) assess the cutting edge of RNA structure prediction techniques; (ii) compare the different methods and tools, elucidate their relative strengths and weaknesses, make clear their limits in terms of sequence length and structure complexity; (iii) determine what has still to be done to achieve an ultimate solution to the structure prediction problem; (iv) promote the available methods and guide potential users in the choice of suitable tools for different problems; (v) and encourage the RNA structure prediction community in their efforts to improve the current tools.

The procedure that governs *RNAPuzzles* is straightforward. Based on the successful first round, we propose the following steps:

□ Complete nucleotide sequences will be periodically released to all those interested and who agree to keep sequence information confidential. These target sequences correspond to experimentally determined crystallographic structures, kindly provided by experimental groups, and not yet published in any form. Confidentiality of RNA sequence information is essential to protect the target selection done by the crystallographers and the molecular engineering construction strategies necessary for crystallogenesis.

- The interested groups will have a specified length of time (usually 4 to 6 weeks) to submit their predicted models on a web site in a standard pdb format that respect atom naming and accepted nomenclature.
- The predicted models will be evaluated with regard to stereochemical correctness, topology and geometrical similarity with the experimental structure.
- After the publication of the original X-ray structures, all the predicted models, experimental results and comparison data will be made publicly available.

To set up and automate these steps, the *RNAPuzzles* team has put together a public web site in which the announcements relative to the experiments and their results will be published, and a processing pipeline to carry out model evaluation. The web site is publicly accessible at <http://paradise-ibmc.u-strasbg.fr/rnapuzzles/>.

## **Structure Analysis and Comparison**

The evaluation of the biological value of a structural model raises many questions. How to determine if a given model is a meaningful prediction? What is a biologically meaningful prediction? Which questions should a structural model answer? Although a 5 Å resolution model is useless for the fine details of a catalytic process, such a model can help to understand the overall architecture or the interplay between partners in a complex. Thus, while some questions require very high precision (1 Å or below), others may be answered with residue-level or domain level precision. Very high precision models cannot be addressed by contemporary modeling methods. However, lower-level biochemical understanding or a general architectural level can be usefully addressed by present-day modeling tools.

To evaluate the predictive success of the proposed models we established two general criteria:

(i) The predicted model must be geometrically and topologically as close as possible to the experimentally determined structure, used as the reference. It is assumed that the crystal structure or the NMR structure is correct within the limitations of the experimental methods.

(ii) The predicted model must be stereo-chemically correct (bond distances and intermolecular contacts are close to the experimentally observed values);

To geometrically compare the predicted models against the experimental structures we used the Root Mean Square Deviation (RMSD) measure and the Deformation Index (DI) [Parisien 2009]. The RMSD is the usual measure of distance between two superimposed structures defined by the formula:

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

in which A and B are the structures and  $(a_i - b_i)$  represents the distance between the  $i^{\text{th}}$  atoms of both structures. The DI is given by:

$$DI(A,B) = \frac{RMSD(A,B)}{MCC(A,B)},$$

in which MCC is the Matthews Correlation Coefficient [Matthews 1975] computed on the individual base pair and base stacking predictions. The reason for this choice is that the RMSD, as a measure of similarity, does not account for specific RNA features such as the correctness of base pair and stacking interactions. The DI score complements the RMSD values by introducing those specific features in the metric. Using the DI value, the quality of two models with close RMSDs, can be discriminated according to the accuracy of their predictions of the base pairing and stacking interactions of the experimental structure. As we observed in the first experiments, the ranking of the models is sensitive to the chosen metric (see Tables 1-3). Such observations were also made during the CASP competitions (see [Marti-Renom 2002]). This can be expected since the various methods rely on different approximations with different consequences. The metrics can however help the design of improved treatments of the approximations and methods.

In a recent work, Weeks, Dokholyan and co-workers showed that when sampling the conformational space of an RNA molecule using discrete molecular dynamics, the RMSD values are distributed normally with a mean related to the length of the molecule by the power law:

$$\langle RMSD \rangle = a \times N^{0.41} - b$$

where N is the number of nucleotides and a and b are constants that depend on whether secondary structure information is provided as input to the molecular dynamic simulation. From this observation it is possible to compute the significance level (P-value) of a prediction with a given RMSD with respect to an accepted structure. This P-value corresponds to the probability that a given structure

prediction is better than that expected by chance [Hajdin 2010]. Structure models with *P*-values less than 0.01 represent, in general, successful predictions.

The stereo-chemical correctness of the predicted models was evaluated with MolProbity [Davis 2007], which provides quality validation for 3D structures of nucleic acids. MolProbity performs a number of automatic analyses, from checking the lengths of the H-bonds present in the model to validating the compliance with the rotameric nature of the RNA backbone [Murray 2003]. The reduce-build script of MolProbity was used for adding hydrogen atoms to the heavy atoms of the models. As a single measure of stereo-chemical correctness, we chose the *clash score*, i.e. the number of steric clashes per thousand residues [Word 1999].

All the computed values are showed in a comparison summary page, which ranks the submitted models according to each of the computed metrics. In addition to the comparison summary, we provide a report for each of the predicted models. The report presents the structural superposition between predicted model and experimental structure, the analysis of the predicted base pairs, i.e. correctly predicted (true positives), wrongly predicted (false positives) and missed (false negatives) and a complete Deformation Profile matrix (DP) which provides an evaluation of the predictive quality of a model at multiple scales as described in [Parisien 2009].

## **The problems**

Two crystallography laboratories sent coordinates for the prediction contest: the laboratory of Thomas Hermann at UC San Diego and that of Dinshaw Patel and Alexander Serganov at the memorial Sloan-Kettering Cancer Center. The three trial experiments were the following.

### ***Problem 1: Dimer.***

What is the structure of the following sequence:

5' CCG CCG CGC CAU GCC UGU GGC GG 3',

knowing that the crystal structure shows a homodimer that contains two strands of the sequence. The strands hybridize with blunt ends (C-G closing base pairs). The solution structure corresponds to the regulatory element from human thymidylate synthase mRNA [Dibrov 2010] which, in the crystal, forms a dimer with two asymmetrical internal loops despite perfect sequence symmetry (Figure 1a and

1b). The crystal structure was resolved to 1.97Å resolution. A total of 14 predicted models were submitted with a RMSD ranging from 3.4 Å to 6.9 Å (mean RMSD of 4.7 Å) (Table 1).

***Problem 2 : Square.***

The crystal structure, which was resolved to 2.2Å resolution, shows a 100 nt square of double-stranded RNA that self-assembles from four identical inner and four identical outer strands [Dibrov 2011]. The secondary structure shown was used for the design of the square. Actual base pairing in the crystal may deviate. 3D coordinates of the nucleotides in the inner strands (B, D, F, H) were provided. What are the structures of the outer strands (A, C, E, G)?

The square is formed by 4 helices connected by 4 single stranded loops. All the helices are identical identical at sequence level and so are all the loops (Figure 3).

***Problem 3 : A riboswitch domain.***

A domain of a riboswitch was crystallized. The sequence is the following:

5' CUC UGG AGA GAA CCG UUU AAU CGG UCG CCG AAG GAG CAA GCU  
CUG CGC AUA UGC AGA GUG AAA CUC UCA GGC AAA AGG ACA GAG 3'

The crystallized sequence was slightly different (an apical loop was replaced by a GAAA loop) but this detail of RNA crystal engineering was not disclosed to modelers to protect the crystallographers (Figures 6a, 6b) [Huang 2010].

**Results**

Eight research groups participated in the first three RNAPuzzles experiments. The Bujnicki group used a hybrid strategy previously developed for protein modeling in the course of the CASP experiment [Kosinski 2003]. The Chen lab used a multiscale free energy landscape-based RNA folding model (Vfold model) [Cao and Chen 2011, Chen 2008]. The Das group used the stepwise assembly (SWA) method for recursively constructing atomic-detail biomolecular structures in small building steps [Sripakdeevong and Das 2011]. The Dokholyan group adopted a multiscale molecular dynamics approach as described in [Ding 2011]. The Flores group used the RNABuilder program, a computer assisted RNA modeling tool [Flores 2010]. For a more detailed description of particular methods see Supplementary Information.

The amount of time required to produce the models and the degree of automation varied sensibly with



the different approaches. These descriptions can be found in the supplementary material. One point should be emphasized. Compared to CASP protein targets, with RNA molecules, a RNA puzzle typically involves multiple ‘mini-puzzles’ such as separate tertiary modules and non-Watson-Crick pairs. There are several examples of this from this first round, e.g., the four corners and the four helices of the nanosquare. Thus, even a single puzzle is powerful for testing modeling methods.

***Problem 1: Dimer.***

Fourteen predicted models were submitted. The RMSDs range from 3.4 Å to 6.9 Å (with a mean of 4.7 Å). The base pair interactions are correctly predicted in almost all models with more than 85% of WC base pairs correctly predicted in all but two models and more than 75% of stacking interaction predicted in all but one model. Contrary to the X-ray structure, most of the proposed models present a symmetric structure. The only exceptions are the models from the Das’s laboratory (see Table 1). From the analysis of the Deformation Profile values (Figure 2) it is clear the internal loops were the domains most difficult to predict (Figure 1c and 1d) and that helix H2, probably because of its location between the loops, presents a particularly large interval of DP values. Several models present high values for the Clash Score, which could reflect the need for up-dated dictionaries of distances and angles or stronger constraints towards the dictionary values.

***Problem 2 : Square.***

Thirteen predicted models were submitted with RMSDs ranging 2.3 Å from 3.7 Å (mean RMSD of 2.9 Å) (Table 2). The standard deviation is the least of all three problems. Such values of RMSDs constitute significant predictions for molecules of this size (P-value <  $5 \times 10^{-15}$  for all models). As expected the helical regions are better predicted than loops with mean DP values between 5 and 10 for all loops and less than 5 for 3 of the helices (Figure 3c, 3d and 4) with the exception of helix 1 in which the three base pairs close to loop 4 deviate slightly from the canonical Watson-Crick geometry (Figure 5). As for Problem 1, the base pairing and stacking were essentially well predicted but, again, there are a couple of very high values of Clash Scores with most models giving values below that of the X-ray structure.

### **Problem 3 : A riboswitch domain.**

This was the most complex case to model with the most intricate tertiary structure. Twelve predicted models were submitted with RMSDs ranging from 7.2 Å to 23.0 Å (mean RMSD of 14.4 Å) (Table 3). The P-values are accordingly very high (except maybe for the first model). Although any model with an RMSD higher than 7 Å is at the limit of any useful prediction the overall molecule architecture was reasonably predicted by at least the two models with the lowest RMSD values. The inter-domain DP values for the 10 pairwise helix-helix predictions (Figure 7) shows that the Chen's model presents the lowest DP for the three way junction (P1-P2, P1-P3 and P2-P3) and a consistently lower than average DP for the coaxial stacking of P2-P3-P3a-P3b. This coaxial stacking was also reasonably predicted (DP < 15) by five of the models (Table 4). Finally, the active site, present in a 13 nucleotide internal loop between domains P3 and P3a, was predicted with an RMSD < 6 Å in all except one model (Table 5, Figure 6c). The non-Watson-crick base pairs are not well predicted.

### **Discussion**

Here we presented RNAPuzzles, a collective blind experiment for *de novo* RNA structure prediction evaluation. We hope that this initiative will function as an open forum where the members of the RNA modeling community can compare their methods, tools and results and newcomers to the field can get a head start. The success of *RNAPuzzle* will greatly depend on the engagement of the prediction community as well as the generosity of the experimental community. Most importantly, this work will hopefully convince more structural biologists to offer problems to the modeling community in the future.

This first contest had clear limitations and several improvements can be already planned. (1) As in CASP, ask the modelers to predict the deviation of their own models to the native structure, in particular in terms of per-residue (or per-atom) deviations (in Å) from the unknown native structure. This could be encoded in the B-factor field. The number of submissions should be limited and the submissions should be ranked by the authors. (2) Further along those lines, it would be worthwhile to improve model scoring and ranking so that an absolute ranking of all models, taking into account local and global model quality, is produced. (3) Because the RNA structure database grows, template-based methods are becoming increasingly important and, consequently, future RNA puzzles should also include structures of homologs of existing folds (e.g. a riboswitch with an alternative ligand or a mutation). (4) The extension of the contest to include structures of RNA-protein complexes.

The assessment of model accuracy requires reliable and meaningful metrics for comparisons between the models and the experimentally determined structures used as a “gold standard”. In addition to the metrics currently used (generic to all macromolecules or specific for RNA), it may be worthwhile to include metrics that have been shown to perform very well at both global and local level for the very wide range of model qualities (from very inaccurate to very accurate), have been generally accepted in the protein structure prediction field and are used by assessors in the CASP experiment. In particular, the GDT\_TS score [Zemla 2003] is defined as the average coverage (fraction of superimposed residues) of one structure by another in superpositions carried out with four different distance thresholds (for proteins these are typically 1, 2, 4, and 8 Å). The exact per-residue deviation values are ignored (e.g. residues with deviations ranging from 4.1 to 8 Å from the native have identical contributions to the score). The GDT\_TS score is unfortunately dependent on the molecule size. The TM-score [Zhang 2004] attempts to eliminate the dependence on protein size by taking into account the radii of gyration of compared structures. The value of the TM-score always lies in range (0, 1], with better templates having higher TM-scores.

## **Materials and Methods**

A brief description of the methodology used by the modeling groups together with comments follow for each group.

### **Bujnicki Group**

The Bujnicki group used a hybrid strategy previously developed for protein modeling in the course of the CASP experiment [Kisinski 2003]. Briefly, initial models were constructed by template-based modeling and fragments assembly with a comparative RNA modeling tool ModeRNA [Rother 2011], using constraints on secondary structure. For RNA Puzzle Problem 2 the secondary structure was provided by the organizers, while for Problem 3 it was calculated as a consensus of >20 methods using RNA metaserver (<http://genesilico.pl/rnametaserver/>). The initial models were expected to possess approximately correct Watson-Crick base-pairing and stacking interactions within individual structural elements but their mutual orientation and tertiary contacts required optimization.

The initial models were subjected to global refinement using SimRNA, a de novo RNA folding method [Rother 2011], which has been inspired by the REFINER method for protein folding [Boniecki 2003]. SimRNA uses a coarse-grained representation, with only three centers of interaction per nucleotide residue. The backbone is represented by atoms P of the phosphate group and C4' of the ribose moiety, whereas the base is represented by just one nitrogen atom of the glycosidic bond (N9 for purines or N1 for pyrimidines). The remaining atoms are neglected. Such a simplistic representation allows to retain the main characteristics of the RNA molecule such as base pairing and stacking, and a spiral shape of the backbone in helices), while it significantly lowers the computational cost for conformational transitions and energy calculation. As an "energy" function, SimRNA employs a statistical potential derived from frequency distributions of geometrical parameters observed in experimentally determined RNA structures. Terms of the SimRNA energy function (for the virtual bond lengths, flat and torsion angles, pairwise interactions between the three atom types) were generated using reverse Boltzmann statistics. For searching the conformational space, SimRNA employs Monte Carlo dynamics controlled by an asymmetric Metropolis method [Metropolis 1949] that accepts or rejects new conformations depending on the energy change associated with the conformational change, with the probability of acceptance depending on the temperature of the system. Simulations can be run in the isothermal or energy minimization (simulated annealing) mode, or in the conformation space search mode (replica exchange). While SimRNA allows for simulations that employ only the sequence information, starting from an extended structure, it can use user-defined starting structures and restraints that specify distances or allowed distance ranges for user-defined atom pairs. For RNA Puzzles, the Bujnicki group used restraints on secondary structure that allowed the predicted base pairs to be maintained. Following a series of simulations, lowest-energy structures were selected for the final refinement.

The final models were built by first reconstructing the full-atom representation using RebuildRNA (Lukasz P, Boniecki M, Bujnicki JM, unpublished) and then optimizing atomic detail of selected residues with SCULPT [Surles 1994] and HyperChem 8.0 (Hypercube Inc.). For Problem 2, the known coordinates of four strands were used as provided by the organizers and "frozen" at the optimization stage.

The computer calculation time (on a single processor) was as follows: ModeRNA: <2 h; SimRNA and RebuildRNA: ~150h; SCULPT <1h; HyperChem: ~12h.

In the case of the Bujnicki group the proportion of human vs computer time was relatively large (approximately equal), as the RNA Puzzles experiment was regarded as an opportunity for training in the use of various modeling methods, in a spirit very similar to the collective work of that group during the CASP5 modeling season [Kosinski 2003]. Consequently, a large fraction of human time involved discussions and communication between the two parts of the team physically located in two different cities (Poznan and Warsaw). The human time devoted to interactions with software (preparation of input files, setting up simulations, analyses of output files, and manual refinement using the graphical user interfaces of SCULPT and HyperChem) summed up to about 30h, with the majority of time devoted to Problems 2 and 3.

### **Chen Group**

We used a multiscale approach to predict RNA 3D structure from the sequence (Cao and Chen 2011). For a given RNA sequence, we first predict the 2D structure from the free energy landscape using the Vfold model (Cao and Chen 2005, 2006a, 2006b, 2009; Chen 2008). The Vfold model allows us to compute the free energies for the different RNA secondary structures and pseudoknotted structures, from which we can predict the (low-free energy) folds. Distinguished from other existing models, the Vfold model is based on the virtual bond (coarse-grained) structural model and hence it enables direct evaluation of the entropy parameters for the different RNA motifs, especially for the pseudoknotted structure. Such a physics-based approach to the evaluation of the entropy and the free energy may lead to more reliable 2D structure prediction. In our calculation for the 2D structures, the base stacking energies are adopted from the Turner energy rules (Serra and Turner 1995). Second, based on the predicted 2D structure, we construct a 3D coarse-grained scaffold. In the coarse-grained structure, we use three atoms (P, C<sub>4</sub>, N<sub>1</sub> or N<sub>9</sub>) to represent a nucleotide. To construct a 3D scaffold, we model the predicted helix stems by A-form helices. For the loops/junctions, we use the fragments from the known PDB database. Specifically, we build a structural template database by classifying the structures according to the different motifs such as hairpin loops and internal/bulge loops, 3-way junctions, 4-way junctions, pseudoknots, etc. We then search the optimal structural templates for the

predicted loops/junctions from the structural template database. Third, we build the all-atom model from the coarse-grained scaffold by adding the bases to the virtual bond backbone. In the final step, we refine the all-atom structure by using AMBER energy minimization. We run 2000 steps minimization with 500.0 kcal/mol restraints for all the residues, followed by another 2000 steps minimization without restraints.

The computation involves two steps: (a) the prediction of the 2D structure and the construction of the coarse-grained 3D structure and (b) AMBER energy minimization. The computer times ( $T_a$ ,  $T_b$ ) for the two steps are (<1, 53) minutes, (<1, 81) minutes, and (26, 143) minutes, for the predictions of the dimer, the square, and the riboswitch domain, respectively. The first step calculation was performed on a desktop PC with Intel(R) Core(TM) 2 Duo CPU E8400 @ 3.00 GHz and the second step computation was carried out on a Dell EM64T cluster (Intel (R) Xeon(R) 5150 @ 2.66 GHz).

For predicting the dimer and a riboswitch structure (problems 1 and 3), we only relied on the sequence information and the 3D structures were generated by computer. No human interference was involved in the process. For the prediction of the square structure (problem 2), we used the experimentally determined structure for one strand to refine the other strand. The loops and the secondary structure of the square were predicted by the computer through the Vfold model (Cao and Chen 2011).

## **Das Group**

The Das lab employed a newly developed *ab initio* method called stepwise assembly (SWA) for recursively constructing atomic-detail biomolecular structures in small building steps. Each step involved enumerating several million conformations for each monomer, and we covered all step-by-step build-up paths in polynomial computational time. The method is implemented in Rosetta and uses the physically realistic Rosetta all-atom energy function [Das 2010, Das 2008]. We have recently benchmarked SWA on small RNA loop-modeling problems [Sripakdeevong and Das 2011]. We also applied *de novo* fragment assembly with full-atom refinement (FARFAR, also implemented in Rosetta), but did not submit those solutions as they either agreed with the SWA models (Problem 1; parts of Problems 2 and 3) or did not give converged solutions (other parts of Problems 2 and 3) [Das 2010].

Due to the deterministic and enumerative nature of SWA, the computational expense is high relative to stochastic and knowledge-based methods. The computational expense ranged from 20,000 (Problem 1) to 50,000 CPU-hours (Problems 2 and 3). Also, as we were developing code 'on-the-fly', we did not have time to fully optimize the run-time, but are doing so now.

The SWA modeling runs were fully automated. Manual input was used near the beginning to setup the runs, and near the end to ensure that models presented a diversity of base pairing patterns -- both of these steps could be easily automated, but the Stepwise Assembly method was still under development during the course of this community-wide RNA blind prediction experiment.

The lessons learned from the three models are the following.

Problem #1: SWA models 1 and 3 (out of five submitted) performed reasonably well on the 46-nucleotide homo-dimer, especially at the 9-nt L1 region (see Fig. 1C). Both models correctly predicted the non-canonical *cis* W.C/W.C C9-C37 base pair and the extra-helical bulge at U39. This accuracy was aided by a strategy that gave entropic bonuses to bulged nucleotides that make no other interactions; the bulges are 'virtualized' within Rosetta [Sripakdeevong and Das 2011]. In this L1 region, both models gave 1.0 Å all-heavy-atom RMSD to the crystallographic model, excluding the U39 extra-helical bulge. In contrast, none of five SWA models achieved atomic accuracy in the sequence-identical 9-nt L2 loop (see Fig. 1D; greater than 3.0 Å RMSD). Model 3 did correctly predict C14 to be an extra-helical bulge and C15 and C32 to be base-paired. However, the exact geometry of the predicted C15-C32 base pair and an additional U16-G31 base pair were incorrect. In the crystallographic model, the base of U16 bulged out and its phosphate formed hydrogen-bonding interactions with the base of G31; in our implementation at the time, we 'virtualized' the phosphates of any bulged nucleotides along with their bases. Partial virtualization of bulged bases and more rigorous modeling of conformational entropy are under investigation.

Problem #2: The SWA models performed well in the regions of the 100-nt "Self-assembling RNA square" within putatively regular secondary structure. We did not assume these to be ideal A-form helices, but modeled them from scratch. These regions are composed mainly of Watson-Crick base pairs but also included a non-canonical *cis* W.C/W.C base pair at corner E/F (see Fig. 3c) that SWA

model 1 correctly predicted. In contrast, the SWA models did not reach atomic accuracy for any of the 5-nt loops at each of the four corners of the square RNA. We partly expected this because the ‘corners’ of the nanosquare originated from a 5-nt bulge in HCV IRES domain IIa [PDB number:2PN4] [Zhao 2008], which happened to be part of our comprehensive SWA benchmark [Sripakdeevong and Das 2011]. There, it was possible to sample the crystallographic loop conformation but not to select it as the lowest-energy structure; the loop forms direct hydrogen bonds to metal ions, and these interactions are not yet modeled in Rosetta.

With this result in mind, after the nanosquare crystal structure was released, we compared it to the full ensemble of models generated by SWA. Loops in corner C/D and G/H were engaged in significant crystal contacts; but loops A/B and E/F should have been amenable to high-accuracy modeling. Indeed, for both of these loops, SWA sampled the crystallographic conformation of these loop regions with  $<1.0 \text{ \AA}$  RMSD, but these models had significantly worse Rosetta energy than our submissions. Again, these corners (and indeed all four corners) involved the binding of either divalent metal ions or cobalt hexamine (III). The lesson learned (or verified) from this puzzle is that approximations in the Rosetta all-atom energy function, especially with regards to metal ions, still remain too inaccurate to permit atomic-resolution RNA modeling on a consistent basis. This puzzle has inspired us to develop approaches to include metal ions during the *de novo* build-up of models.

**Problem #3:** Our recent research has focused on the prediction of high-resolution motifs as a stepping stone to modeling larger RNAs. This glycine riboswitch puzzle was thus currently out of range – its core 3-way junction and glycine binding site form an intricate noncanonical pairing network involving more than a dozen residues. Further, interactions across a dimer interface appear crucial for stabilizing the riboswitch conformation, but this information was not available to us. Our models were based on generating low-energy Rosetta SWA models for individual loops, two-way junctions, and three-way junctions, and then connecting them with ideal helices. Surprisingly, this basic approach, ignorant of higher order interactions, gave the best base pair recoveries (INF all, INF wc, INF nwc; see models 1 and 2 in Table 3) amongst submitted models. Other models of ours (models 4 and 5) gave the best RMSDs for the glycine-binding site. However, these were very far from atomic accuracy (2.8  $\text{\AA}$  and 2.9  $\text{\AA}$ ). Most critically, the global structure of the RNA was not recapitulated (RMSD and DI, Table 3). The helices formed the correct tuning-fork-like rearrangement but were twisted relative to the



crystallographic model (Tables 4 and 5). Globally correct solutions require global optimization, and this puzzle has motivated us to develop iterative hybrid high-resolution/low-resolution approaches to RNA modeling, analogous to the rebuild-and-refine method used in Rosetta template-based modeling [Qian 2007]. As a final note, in the paper describing this puzzle's crystal structure, a striking structural similarity of the glycine riboswitch core to a previously solved SAM-I riboswitch [\$\$\$] was noted. If such similarities could be inferred from sequence or multiple sequence alignments (analogous to fold recognition methods in protein modeling), we expect that substantially more accurate models could be built. We are therefore hopeful about by further development of RNA structural bioinformatics approaches such as Rmdetect [Cruz 2011] and FR3D [Sarver 2008].

### **Dokholyan group:**

The Dokholyan group adopted a multiscale molecular dynamics approach [Ding 2011]. Briefly, coarse-grained discrete molecular dynamics (DMD) simulations are used to sample the vast conformational space of RNA molecules. The representative structures are selected from the coarse-grained simulations based on energies and/or additional filters (i.e., radius of gyration and other experimentally known parameters). In the coarse-grained DMD simulations, RNA nucleotides are represented by three pseudo-atoms corresponding to the base, sugar, and phosphate groups [Ding 2008]. The neighboring beads along the sequence are constrained to mimic the chain connectivity and local chain geometry, including covalent bonds, bond angles, and dihedral angles. The parameters for bonded interactions mimic the folded RNA structure and are derived from high-resolution RNA structures. Non-bonded interactions include base-pairing, base-stacking, short-range phosphate-phosphate repulsion, and hydrophobic interactions. The interaction parameters are derived from the sequence-dependent free energy parameters of the individual nearest-neighbor hydrogen bond model (INN-HB) [Mathews 2009]. Given a coarse-grained RNA model, the corresponding all-atom model is reconstructed and further optimized with all-atom DMD simulations [Ding and Dokholyan, unpublished]. The all-atom DMD model of RNAs is the extension of the all-atom DMD model of proteins [Ding 2008b]. In DMD simulations the structural information of a given RNA, such as base pairs and distances between specific nucleotides, can be incorporated as constraints to guide the RNA folding [Gherghe 2009, Lavender 2010].

The CPU time for DMD simulations depends on RNA length. For the coarse-grained simulations, previously benchmark suggests a near linear dependence to RNA length [Ding 2008]. For example, for a RNA of ~80 nt (such as the 3rd puzzle) the total computational time for the coarse-grained DMD simulation is ~12 hours. The procedure to identify the representative structures using clustering algorithm is usually less than 1 hour. The CPU time of the all-atom DMD simulation also depends on RNA length,  $n$ , with the computational complexity of  $\sim n \ln(n)$ . For the 84 nt RNA (problem 3), the CPU time was approximately 18 h; and the CPU time for 100 nt RNA (problem) was approximately 24 h.

In the current three RNA puzzles, we included base-pairs either from previous knowledge (input from the experimentalists in puzzle 2; RNA secondary structure prediction combined with biochemical validates in the puzzle 3) or with biochemical intuition (puzzle 1). Once the structural information is gathered and prepared for the refinement simulations, the rest of the computational efforts are fully automated.

Our lesson is that inclusion of experimentally validated structure information as much as possible will help improve the prediction. In the case of uncertainties, it is better not to include them as constraints, but rather let the DMD simulations sample allowed conformational space, since the errors of the input constraints can bias the simulations toward unphysical structures. In our case of puzzle 1, we over-estimated the base-pairs in the middle of the monomer sequence based on the apparent knowledge of the “hybridize with blunt ends (C=G closing base pairs)”. As the result, our predicted structure of puzzle 1 had the highest RMSD among all the predictions. In a control simulation *post priori*, where only the GC pairs in the ends are constrained to form base pairs, the predicted model structure had a much smaller RMSD.

## **Flores Group**

For our 3D structure prediction we used RNABuilder (named MMB in a subsequent release), an internal coordinate mechanics code which allows the user to specify the flexibility, forces, constraints, and full or partial structural coordinates to model the structure and/or dynamics of an RNA molecule [Flores 2010]. Working in internal coordinates has the advantage that regions of the molecule whose

structure is known can be rigidified turning many atoms into a single body, eliminating the cost associated with solving the equations of motion for its internal degrees of freedom [Flores 2011]. Steric exclusion can be accounted for economically using collision-detecting spheres which are applied to a subset of atoms in user-specified residues. Any canonical or non-canonical base pairing interaction catalogued in [Leontis 2002], plus stacking and a “Superimpose” threading force can be enforced between any and all pairs of residues specified by the user. These features have been used for RNA threading [Flores 2010b] and for generating an all-atoms trajectory of ribosomal hybridization using structural and biochemical information [Flores 2011].

The processing time on a single core of an 3.0 GHz Intel processor was about 94 minutes. We note that this run was not optimized for speed, and also that a newer version of RNABuilder (named MMB) is at least 2X faster due to improvements in the underlying Simbody internal coordinate dynamics engine [Michael A. Sherman, Ajay Seth, Scott L. Delp. Simbody: multibody dynamics for biomedical research. *Procedia IUTAM* 2, 241-261 (2011)].

RNABuilder is intended to be easy to use, and this goal is supported by the use of a single free-format command file that is prepared using a relatively intuitive syntax using terms which may be recognized by a biologist. However the package also intends to give the user control over the flexibility, forces, and parameters of the model in order to be useful for a wide variety of applications, hence it is not automated. The human time required for preparing a run is thus dependent on the experience of the user and the complexity of the task. RNABuilder is designed to enable fast runtimes; most tasks we undertake require minutes to hours again dependent on the task. A trained user can also reduce the degrees of freedom and structure the problem to allow larger integration time steps for greater efficiency. Also in practice most users will do multiple calculations prior to coming to a biological conclusion. In our experience the human/computer time ratio is typically much greater than unity.

### **SantaLucia Group**

The details are as follow:

#### CASE 1 Prediction

- a. The machine calculation time ~ 25 Minutes on a 2.2GHz laptop that runs on windows vista.

- b. What was the manual input if any and at what stages of the process – No manual input
- c. How many people were involved (man-hours) – 1hr

#### CASE 2 Prediction

- a. The machine calculation time ~ 20 Minutes on a 2.2GHz laptop to construct one of the four “building blocks”. 10 Minutes to run Optimization algorithm to seal the daggling ends
- b. What was the manual input if any and at what stages of the process – Manual Input in aligning the building blocks onto the provided inner strand.
- c. How many people were involved (man-hours) – 1hr.

#### CASE 3 Prediction

We will not be submitting our prediction for this case. Our software is currently being developed to address multiloop predictions and as such it is not ready for this task.

#### ***PDB file normalization***

Both solution files and predicted model files, submitted in the PDB format, were normalized in order to comply with a common standard. Only the first model present in the file was considered. All records except for the 'ATOM' and 'TER' records were ignored. Only the four nucleotides A, C, U and G were considered. Modified nucleotides were treated as unmodified bases and extra atoms discarded (e.g. a 5-Bromouracil is treated as a normal Uracil and the extra bromine atom is discarded). The only atoms kept are the base ones (C2, C4, C6, C8, N1, N2, N3, N4, N6, N7, N9, O2, O4 and O6) and the sugar-phosphate backbone (C1', C2', C3', C4', C5', O2', O3', O4', O5', OP1, OP2 and P).

#### ***Stereochemical evaluation***

The stereochemical evaluation is performed using the MolProbity [Davis 2007] tool. In a first step hydrogen atoms are added to the model using the “reduce-build” command line utility, then, the clash score value is computed using the “oneline-analysis -nocbeta -norota -norama” command.

#### ***RMSD computation***

The RMSD is computed using the “Superimposer” class from the “Bio.PDB” package [Hamelryck 2003]. The “Superimposer” class translates and rotates the comparing model in order to minimize its

RMSD in respect to the reference model. It uses a singular value decomposition algorithm as described in [Golub and Van Loan 1989].

### ***Deformation Index and Deformation Profile computations***

The base-base interactions (BBI) of both solution and predicted models are extracted using the “MC-Annotate” [Gendron 2001] tool. The Interaction Network Fidelity (INF) value is computed as:

$$INF = \sqrt{\left(\frac{TP}{TP+FP}\right) \times \left(\frac{TP}{TP+FN}\right)},$$

where TP is the number of correctly predicted BBI, FP is the number of predicted BBI with no correspondence in the solution model and FN is the number of BBI in the solution model not present in the predicted model. The Deformation Index is then computed as:

$$DI = \frac{RMSD}{INF}.$$

Several partial INF (and respective DI) can be computed if one considers only the Watson-Crick (WC) base pairs ( $INF_{WC}$ ), the non Watson-Crick (NWC) base pairs ( $INF_{NWC}$ ), both WC and NWC base pairs ( $INF_{BPS}$ ) or the stacking interactions ( $INF_{STACK}$ ).

The Deformation Profile is computed using the “dp.py” command from the “SIMINDEX” package [Parisien 2009].

### ***P-value computation***

The P-value is computed as described in [Hajdin 2010] using:

$$P - value = \frac{1 + \operatorname{erf}\left(\frac{(RMSD - \langle RMSD \rangle) / 1.8}{\sqrt{2}}\right)}{2}, \text{ with } \langle RMSD \rangle = a \times N^{0.41} - b.$$

the constants  $a$  and  $b$  depend on whether the secondary structure base pairing information is provided ( $a=5.1$  and  $b=15.8$ ) or not ( $a=6.4$  and  $b=12.7$ ).

### ***Graphics***

Interactive molecular module images in the RNAPuzzles web site are produced with Jmol (<http://www.jmol.org>) and the secondary structures with VARNA [Darty 2009].

## **Acknowledgements**

JAC is supported by the Ph.D. Program in Computational Biology of the Instituto Gulbenkian de Ciência, Portugal (sponsored by Fundação Calouste Gulbenkian, Siemens SA, and Fundação para a Ciência e Tecnologia; SFRH/BD/33528/2008). TH is supported by the National Institutes of Health, grants AI72012 and CA132753. The work of the Bujnicki group was supported by the Polish Ministry of Science (HISZPANIA/152/2006 grant to J.M.B. and PBZ/MNiSW/07/2006 grant to MB), by the EU (6FP grant “EURASNET” LSHG-CT-2005-518238 and structural funds POIG.02.03.00-00-003/09), by the Faculty of Biology, Adam Mickiewicz University (PBWB-03/2009 grant to MR) and by the German Academic Exchange Service (grant D/09/42768 to KR). The work of the Chen group was supported by NIH grant GM063732 and NSF grants MCB0920067 and MCB0920411 to S.-J.C. The Weeks and Dokholyan groups were supported by the U.S. National Institutes of Health (grant GM064803).

## References

1. Berman, H.M. et al. The Protein Data Bank. *Nucleic acids research* **28**, 235-42 (2000).
2. Boniecki, M., Rotkiewicz, P., Skolnick, J. & Kolinski, A. Protein fragment reconstruction using various modeling techniques. *Journal of computer-aided molecular design* **17**, 725-38 (2003).
3. Cao, S. & Chen, S.-J. Free Energy Landscapes of RNA/RNA Complexes: With Applications to snRNA Complexes in Spliceosomes. *Journal of molecular biology* **357**, 292-312 (2006).
4. Cao, S. & Chen, S.-J. Predicting structures and stabilities for H-type pseudoknots with interhelix loops. *RNA (New York, N.Y.)* **15**, 696-706 (2009).
5. Cao, S. & Chen, S.-J. Predicting RNA pseudoknot folding thermodynamics. *Nucleic acids research* **34**, 2634-52 (2006).
6. Cao, S. & Chen, S.-J. Predicting RNA folding thermodynamics with a reduced chain representation model. *RNA (New York, N.Y.)* **11**, 1884-97 (2005).
7. Cao, S. & Chen, S.-J. Physics-Based De Novo Prediction of RNA 3D Structures. *The journal of physical chemistry. B* **115**, 4216-26 (2011).
8. Chen, S.-J. RNA folding: conformational statistics, folding kinetics, and ion electrostatics. *Annual review of biophysics* **37**, 197-214 (2008).
9. Cruz, J.A. & Westhof, E. sequence-based identification of 3d structural modules in RNA with rmdetect. *Nature Methods* (2011).doi:10.1038/nmeth.1603
10. Darty, K., Denise, A. & Ponty, Y. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25**, 1974-1975 (2009).
11. Das, R. & Baker, D. Macromolecular modeling with rosetta. *Annual review of biochemistry* **77**, 363-82 (2008).
12. Das, R., Karanicolas, J. & Baker, D. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nature methods* **7**, 291-4 (2010).
13. Davis, I.W. et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic acids research* **35**, W375-83 (2007).
14. Dibrov, S.M., McLean, J., Parsons, J. & Hermann, T. Self-assembling RNA square. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 6405-6408 (2011).
15. Dibrov, S., McLean, J. & Hermann, T. Structure of an RNA dimer of a regulatory element from human thymidylate synthase mRNA. *Acta crystallographica. Section D, Biological crystallography* **67**, 97-104 (2011).
16. Ding, F. & Dokholyan, N. Multiscale modeling of RNA structure and dynamics. *RNA Structure Prediction and Modelling* (2011).
17. Ding, F. et al. Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA (New York, N.Y.)* **14**, 1164-73 (2008).
18. Ding, F., Tsao, D., Nie, H. & Dokholyan, N. Ab initio folding of proteins using all-atom discrete molecular dynamics. *Structure* **16**, 1010-1018 (2008).
19. Ditzler, M. a, Otyepka, M., Sponer, J. & Walter, N.G. Molecular dynamics and quantum mechanics of RNA: conformational and chemical change we can believe in. *Accounts of chemical research* **43**, 40-7 (2010).
20. Flores, S.C. & Altman, R. Structural insights into pre-translocation ribosome motions. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 205-11 (2011).at <<http://www.ncbi.nlm.nih.gov/pubmed/21121048>>

21. Flores, S.C. & Altman, R.B. Turning limited experimental information into 3D models of RNA. *RNA (New York, N.Y.)* (2010).doi:10.1261/rna.2112110
22. Flores, S.C., Sherman, M.A., Bruns, C.M., Eastman, P. & Altman, R.B. Fast Flexible Modeling of RNA Structure Using Internal Coordinates. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **8**, 1247-1257 (2011).
23. Flores, S.C., Wan, Y., Russell, R. & Altman, R.B. Predicting RNA structure by multiple template homology modeling. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 216-27 (2010).at <<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2872935&tool=pmcentrez&rendertype=abstract>>
24. Gardner, P.P. et al. Rfam: updates to the RNA families database. *Nucleic acids research* **37**, D136-40 (2009).
25. Gendron, P., Lemieux, S. & Major, F. Quantitative analysis of nucleic acid three-dimensional structures. *Journal of molecular biology* **308**, 919-36 (2001).
26. Gherghe, C.M., Leonard, C.W., Ding, F., Dokholyan, N.V. & Weeks, K.M. Native-like RNA Tertiary Structures Using a Sequence-Encoded Cleavage Agent and Refinement by Discrete Molecular Dynamics. *Journal of American Chemical Society* **131**, 2541-2546 (2009).
27. Golub, G. & Van Loan, C. *Matrix Computations*. (Johns Hopkins University Press: 1989).
28. Hamelryck, T. PDB file parser and structure class implemented in Python. *Bioinformatics* **19**, 2308-2310 (2003).
29. Hofacker, I.L. et al. Fast folding and comparison of RNA secondary structures. *Monatshefte fur Chemie Chemical Monthly* **125**, 167-188 (1994).
30. Huang, L., Serganov, A. & Patel, D.J. Structural Insights into Ligand Recognition by a Sensing Domain of the Cooperative Glycine Riboswitch. *Molecular cell* **40**, 774-786 (2010).
31. Jonikas, M. a, Radmer, R.J. & Altman, R.B. Knowledge-based instantiation of full atomic detail into coarse-grain RNA 3D structural models. *Bioinformatics (Oxford, England)* **25**, 3259-66 (2009).
32. Jossinet, F., Ludwig, T.E. & Westhof, E. Assemble: an interactive graphical tool to analyze and build RNA architectures at the 2D and 3D levels. *Bioinformatics (Oxford, England)* **26**, 2057-2059 (2010).
33. Kosinski, J. et al. A "Frankenstein's monster" approach to comparative modeling: Merging the finest fragments of Fold-Recognition models and iterative model refinement aided by 3D structure evaluation. *Proteins: Structure, Function, and Bioinformatics* **53**, 369-379 (2003).
34. Lavender, C. a, Ding, F., Dokholyan, N.V. & Weeks, K.M. Robust and generic RNA modeling using inferred constraints: a structure for the hepatitis C virus IRES pseudoknot domain. *Biochemistry* **49**, 4931-3 (2010).
35. Leontis, N.B., Stombaugh, J. & Westhof, E. The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic acids research* **30**, 3497-531 (2002).
36. Leontis, N.B. & Westhof, E. Geometric nomenclature and classification of RNA base pairs. *RNA (New York, N.Y.)* **7**, 499-512 (2001).
37. Marti-Renom, M. a, Madhusudhan, M.S., Fiser, A., Rost, B. & Sali, A. Reliability of assessment of protein structure prediction methods. *Structure (London, England: 1993)* **10**, 435-40 (2002).
38. Martinez, H.M., Maizel, J.V. & Shapiro, B. a RNA2D3D: a program for generating, viewing, and comparing 3-dimensional models of RNA. *Journal of biomolecular structure & dynamics* **25**, 669-83 (2008).



39. Mathews, D.H., Sabina, J., Zuker, M. & Turner, D.H. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of molecular biology* **288**, 911-40 (1999).
40. Matthews, B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure* **405**, 442-451 (1975).
41. Metropolis, N. & Ulam, S. The Monte Carlo Method. *Journal of American Statistical Association* **44**, 335-341 (1949).
42. Michel, F. & Westhof, E. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *Journal of Molecular Biology* **216**, 585-610 (1990).
43. Moulton, J. Rigorous performance evaluation in protein structure modelling and implications for computational biology. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **361**, 453-8 (2006).
44. Parisien, M., Cruz, J.A., Westhof, E. & Major, F. New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA (New York, N.Y.)* **15**, 1875-85 (2009).
45. Parisien, M. & Major, F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **452**, 51-5 (2008).
46. Qian, B. et al. High-resolution structure prediction and the crystallographic phase problem. *Nature* **450**, 259-64 (2007).
47. Reuter, J.S. & Mathews, D.H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC bioinformatics* **11**, 129 (2010).
48. Rother, K. et al. Template-based and template-free modeling of RNA 3D structure: inspirations from protein structure modeling. *RNA Structure Prediction and Modelling* (2011).
49. Rother, K., Rother, M., Boniecki, M., Puton, T. & Bujnicki, J.M. RNA and protein 3D structure modeling: similarities and differences. *Journal of molecular modeling* (2011).doi:10.1007/s00894-010-0951-x
50. Rother, M., Rother, K., Puton, T. & Bujnicki, J.M. ModeRNA: a tool for comparative modeling of RNA 3D structure. *Nucleic acids research* 1-16 (2011).doi:10.1093/nar/gkq1320
51. Sarver, M., Zirbel, C.L., Stombaugh, J., Mokdad, A. & Leontis, N.B. FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *Journal of mathematical biology* **56**, 215-52 (2008).
52. Sripakdeevong, P. & Das, R. An enumerative stepwise ansatz enables atomic-accuracy RNA loop modeling. *submitted*
53. Surles, M.C., Richardson, J.S., Richardson, D.C. & Brooks, F.P. Sculpting proteins interactively: continual energy minimization embedded in a graphical modeling system. *Protein science: a publication of the Protein Society* **3**, 198-210 (1994).
54. Tinoco, I. & Bustamante, C. How RNA folds. *Journal of molecular biology* **293**, 271-81 (1999).
55. Word, J.M. et al. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *Journal of molecular biology* **285**, 1711-33 (1999).
56. Zemla, A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Research* **31**, 3370-3374 (2003).
57. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702-10 (2004).

58. Zhao, Q., Han, Q., Kissinger, C.R., Hermann, T. & Thompson, P. a Structure of hepatitis C virus IRES subdomain IIa. *Acta crystallographica. Section D, Biological crystallography* **64**, 436-43 (2008).
59. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research* **31**, 3406-3415 (2003).

## Legends for Figures

### Figure 1

Problem 1 - (a) Secondary structure of the reference RNA molecule. X-ray structures of the reference RNA molecule (green) and the predicted models of lowest RMSD (blue) for (b) the full molecule of Das model 3, (c) details of loop L1 of Das model 1 and (d) details of loop L2 of Das model 3.

### Figure 2

Problem 1 - Deformation Profile values for each of the 5 domains of the homodimer. Color lines represent the DP values for the two predicted models with lowest RMSD, Das model 3 (dark red) and Das model 1 (dark green), and for the predicted model with higher RMSD, Dokholyan model 1 (dark blue). Radial red lines indicate the minimum, maximum and mean DP values for each domain.

### Figure 3

Problem 2 - (a) Secondary structure of the reference RNA molecule. X-ray structures of the reference RNA molecules (green) and the predicted models of lowest RMSD (blue) for (b) the full molecule and Bujnicki model 2, (c) details of helices H1, H2 and H4 of Das model 1 and helix H3 of Bujnicki model 2 and (d) details of loops L1 and L2 of Santalucia model 1, loop L3 of Dokholyan model 1 and loop L4 of Bujnicki model 3.

### Figure 4

Problem 2 - Deformation Profile values for the three predicted models with lowest RMSD: Bujnicki model 2 (dark red), Bujnicki model 3 (dark green) and Das model 1 (dark blue). Radial red lines indicate the minimum, maximum and mean DP values for each domain.

### Figure 5

Problem 2 - Detail of the helix H1 of the X-ray structure (green) and of the lowest RMSD model, Das model 1 (blue).

**Figure 6**

Problem 3 - (a) Secondary structure of the reference RNA molecule. (b) X-ray structure of the reference RNA molecule (P1 red, P2 orange, P3, P3a, P3b yellow, active site green) and the predicted model of lowest RMSD Chen model 1 (blue). (c) Detail of the active site for the X-ray structure (green) and the predicted model of lowest RMSD Chen model 1 (blue).

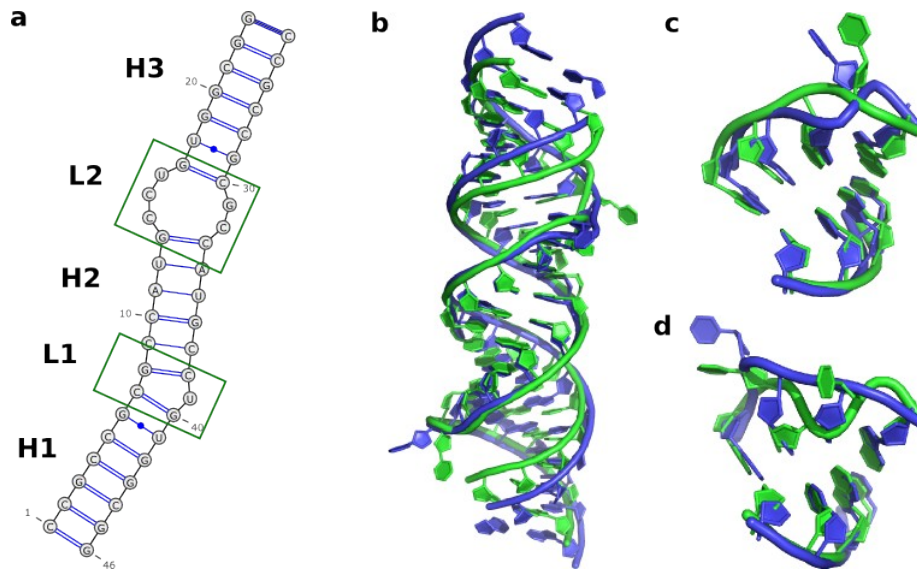
**Figure 7**

Problem 3 - Deformation Profile values of the pairwise helical interdomains (P1, P2, P3, P3a and P3b) for the three predicted models with lowest RMSD: Chen model 1 (dark red), Dokholyan model 2 (dark green) and Das model 5 (dark blue). Radial red lines indicate the minimum, maximum and mean DP values for each intradomain pair.

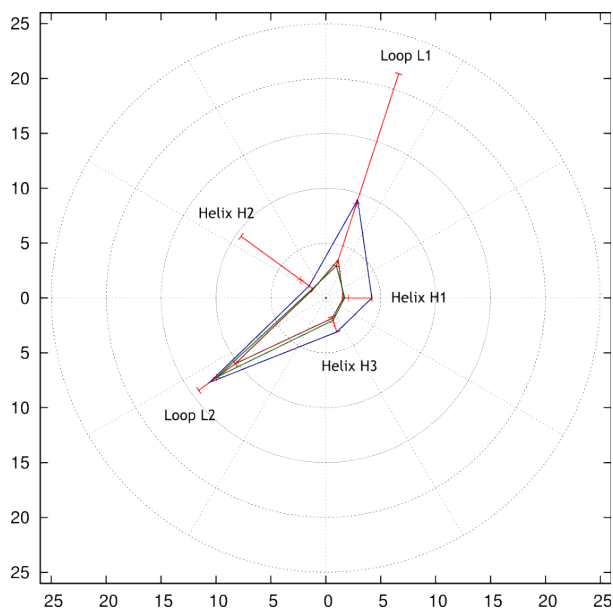
**NOTE: Image resolutions and table aspect will be upgraded in the final version!**

## Figures

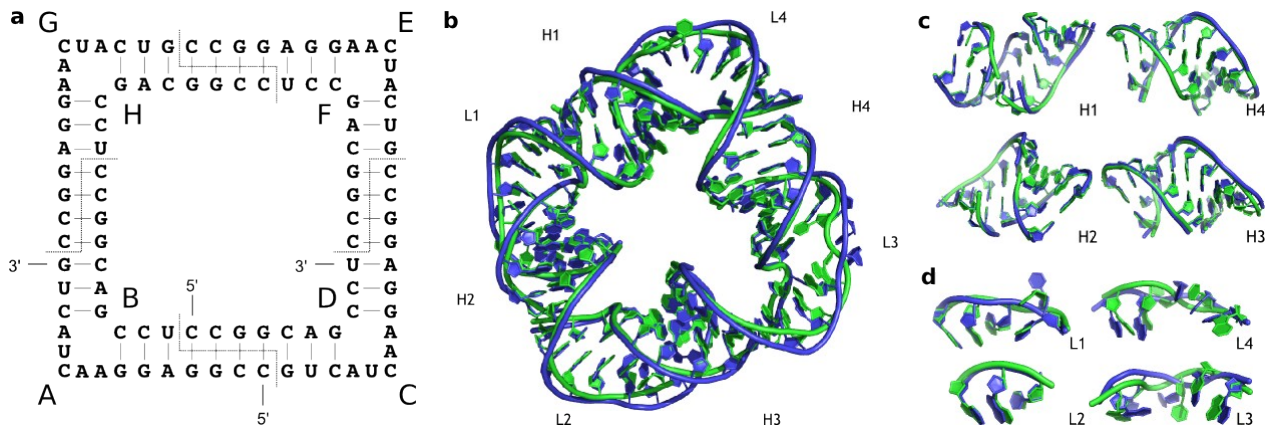
**Figure 1**



**Figure 2**



**Figure 3**



**Figure 4**

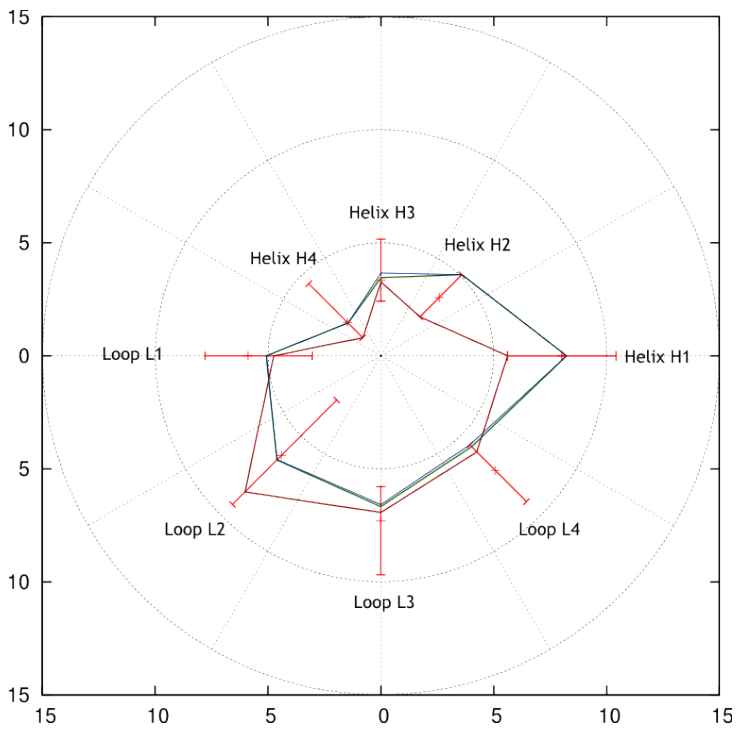


Figure 5

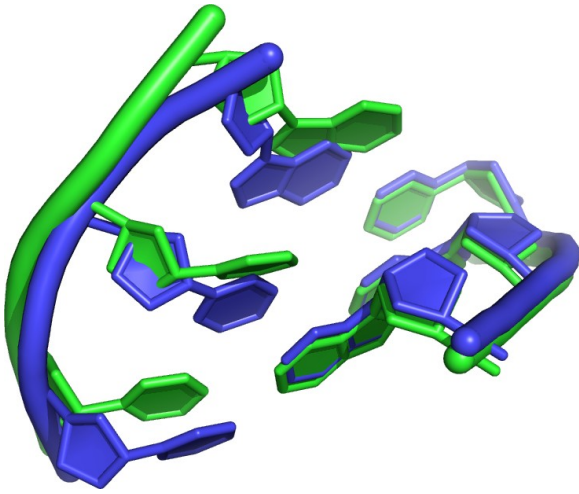
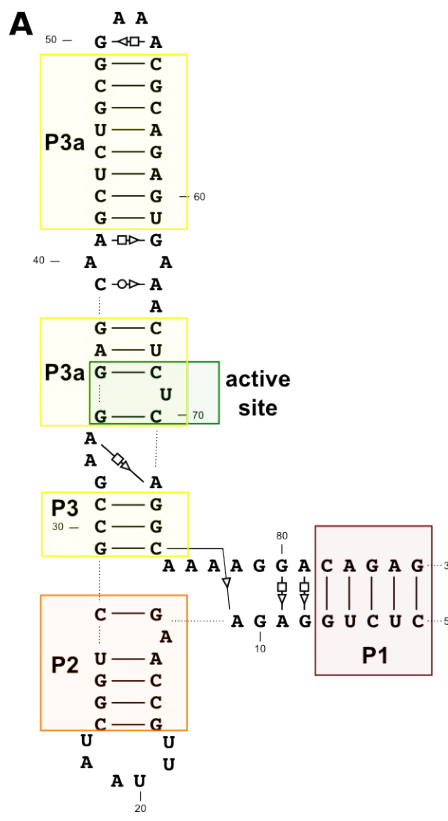
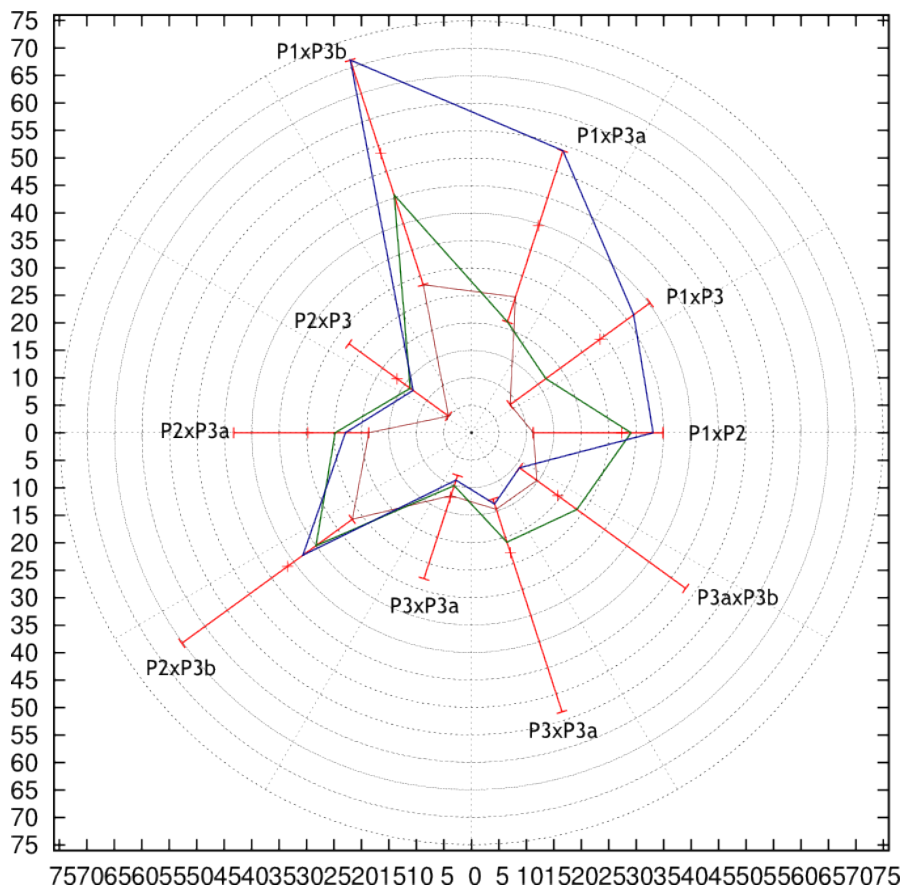


Figure 6



**Figure 7**





## Legends for Tables

### Table 1

Summary of the results of problem 1. Values in each row correspond to a predicted model. a) Name of the research group that submitted the model; b) number of the model among all the models of the same group; c) RMSD of the model compared with the reference crystal structure; d) Columns signaled with “#” indicate the rank of the model in respect to the left hand column metric; e)  $DI_{all}$  – Deformation Index taking into account all interactions (stacking, Watson-Crick and non Watson-Crick); f)  $INF_{all}$  – Interaction Network Fidelity taking into account all interactions; g)  $INF_{wc}$  – Interaction Network Fidelity taking into account only Watson-Crick interactions; h)  $INF_{stack}$  – Interaction Network Fidelity taking into account only Stacking interactions; i) Clash score as computed by the MolProbity suite [Davis 2007]; j) Significance of the predicted model.

### Table 2

Summary of the results of problem 2. Values in each row correspond to a predicted model. a) Name of the research group that submitted the model; b) number of the model among all the models of the same group; c) RMSD of the model compared with the reference crystal structure; d) Columns signaled with “#” indicate the rank of the model in respect to the left hand column metric; e)  $DI_{all}$  – Deformation Index taking into account all interactions (stacking, Watson-Crick and non Watson-Crick); f)  $INF_{all}$  – Interaction Network Fidelity taking into account all interactions; g)  $INF_{wc}$  – Interaction Network Fidelity taking into account only Watson-Crick interactions; h)  $INF_{nwc}$  – Interaction Network Fidelity taking into account only non Watson-Crick interactions; i)  $INF_{stack}$  – Interaction Network Fidelity taking into account only Stacking interactions; j) Clash score as computed by the MolProbity suite [Davis 2007]; k) Significance of the predicted model.

### Table 3

Summary of the results of problem 3. Values in each row correspond to a predicted model. a) Name of the research group that submitted the model; b) number of the model among all the models of the same group; c) RMSD of the model compared with the reference crystal structure; d) Columns signaled with “#” indicate the rank of the model in respect to the left hand column metric; e)  $DI_{all}$  – Deformation Index taking into account all interactions (stacking, Watson-Crick and non Watson-Crick); f)  $INF_{all}$  – Interaction Network Fidelity taking into account all interactions; g)  $INF_{wc}$  –

Interaction Network Fidelity taking into account only Watson-Crick interactions; h)  $INF_{nwc}$  – Interaction Network Fidelity taking into account only non Watson-Crick interactions; i)  $INF_{stack}$  – Interaction Network Fidelity taking into account only Stacking interactions; j) Clash score as computed by the MolProbity suite [Davis 2007]; k) Significance of the predicted model.

**Table 4**

Pairwise interdomain Deformation Profile values for the helical domains P1, P2, P3, P3a and P3b from problem 3. In red are all DP values less than 15.

**Table 5**

The RMSD values between the active site of the X-ray model and each predicted model of problem 3.

## Tables

**Table 1**

Problem 1															
Group <sup>a</sup>	Num <sup>b</sup>	RMSD <sup>c</sup>	# <sup>d</sup>	DI all <sup>e</sup>	#	INF all <sup>f</sup>	#	INF wc <sup>g</sup>	#	INF stack <sup>h</sup>	#	Clash Score <sup>i</sup>	#	P-value <sup>j</sup>	#
Das	3	3.41	1	3.66	1	0.93	1	0.95	2	0.92	1	0.00	5	2.22E-016	1
Das	1	3.58	2	3.89	2	0.92	3	0.95	1	0.91	2	0.00	3	4.44E-016	2
Das	4	3.91	3	4.31	3	0.91	4	0.91	8	0.91	4	0.00	4	1.94E-015	3
Major	1	4.06	4	4.57	4	0.89	5	0.95	6	0.87	5	66.40	11	3.83E-015	4
Chen	1	4.11	5	5.01	6	0.82	9	0.87	11	0.80	8	0.68	6	4.77E-015	5
Das	2	4.34	6	4.70	5	0.92	2	0.95	4	0.91	3	1.36	7	1.28E-014	6
Das	5	4.56	7	5.36	7	0.85	7	0.88	10	0.84	7	0.00	2	3.31E-014	7
Bujnicki	3	4.66	8	5.75	9	0.81	11	0.95	3	0.74	14	54.73	10	4.99E-014	8
Bujnicki	4	4.74	9	6.59	11	0.72	14	0.65	14	0.75	13	83.33	14	6.87E-014	9
Bujnicki	5	4.89	10	6.26	10	0.78	13	0.78	13	0.80	9	81.98	13	1.30E-013	10
Bujnicki	1	5.07	11	5.75	8	0.88	6	0.93	7	0.86	6	0.00	1	2.76E-013	11
Bujnicki	2	5.43	12	6.75	12	0.80	12	0.90	9	0.77	12	71.57	12	1.15E-012	12
Santalucia	1	5.69	13	6.75	13	0.84	8	0.95	5	0.79	11	39.86	9	3.20E-012	13
Dokholyan	1	6.94	14	8.55	14	0.81	10	0.86	12	0.79	10	31.74	8	3.31E-010	14
<b>Mean</b>		<b>4.67</b>		<b>5.56</b>		<b>0.85</b>		<b>0.89</b>		<b>0.83</b>					
<b>Standard deviation</b>		<b>0.93</b>		<b>1.34</b>		<b>0.06</b>	N	<b>0.09</b>		<b>0.07</b>					
										<b>X-Ray Model</b>		<b>1.35</b>			

**Table 2**

Problem 2																	
Group <sup>a</sup>	Num <sub>b</sub>	RMSD <sub>c</sub>	# <sup>d</sup>	DI all <sup>e</sup>	#	INF all <sup>f</sup>	#	INF wc <sup>g</sup>	#	INF nwc <sub>h</sub>	#	INF stack <sup>i</sup>	#	Clash Score <sup>j</sup>	#	P-value <sup>k</sup>	#
Bujnicki	2	2.3	1	2.83	1	0.81	8	0.92	9	0	13	0.79	7	14.54	2	0.00E+000	4
Bujnicki	3	2.33	2	2.9	3	0.8	10	0.91	10	0	2	0.77	9	0.62	1	0.00E+000	2
Das	1	2.5	3	2.9	2	0.86	2	0.96	5	0	8	0.85	2	19.8	5	0.00E+000	3
Dokholyan	1	2.54	4	3.09	5	0.82	6	0.9	11	0	1	0.8	5	19.85	6	0.00E+000	1
Bujnicki	1	2.65	5	2.99	4	0.89	1	0.96	4	0	3	0.86	1	15.47	3	5.55E-017	5
Chen	1	2.83	6	3.74	9	0.76	13	0.9	12	0	9	0.69	13	18.66	4	1.11E-016	6
Das	4	2.83	7	3.46	6	0.82	7	0.97	3	0	12	0.78	8	23.82	8	1.11E-016	7
Major	1	2.98	8	3.82	10	0.78	12	0.95	7	0	10	0.71	12	134.26	12	2.22E-016	8
Das	3	3.03	9	3.67	7	0.83	5	0.97	1	0	6	0.8	6	25.37	10	2.78E-016	9
Das	2	3.05	10	3.69	8	0.83	4	0.97	2	0	7	0.81	3	23.51	7	2.78E-016	10
Das	5	3.46	11	4.18	11	0.83	3	0.96	6	0	11	0.81	4	24.75	9	1.89E-015	11
Flores	1	3.48	12	4.4	12	0.79	11	0.89	13	0	5	0.77	10	165.57	13	2.00E-015	12
Santalucia	1	3.65	13	4.54	13	0.81	9	0.92	8	0	4	0.75	11	25.73	11	4.27E-015	13
<b>Mean</b>		<b>2.90</b>		<b>3.55</b>		<b>0.82</b>		<b>0.94</b>		<b>0.00</b>		<b>0.78</b>					
<b>Standard deviation</b>		<b>0.44</b>		<b>0.59</b>		<b>0.03</b>		<b>0.03</b>		<b>0.00</b>		<b>0.05</b>					
												<b>X-Ray Model</b>		<b>36.10</b>			

**Table 3**

Problem 3																	
Group <sup>a</sup>	Num <sup>b</sup>	RMSD <sup>c</sup>	# <sup>d</sup>	DI all <sup>e</sup>	#	INF all <sup>f</sup>	#	INF wc <sup>g</sup>	#	INF nwc <sup>h</sup>	#	INF stack <sup>i</sup>	#	Clash Score <sup>j</sup>	#	P-value <sup>k</sup>	#
Chen	1	7.24	1	9.84	1	0.74	2	0.86	5	0	6	0.73	1	1.1	3	2.01E-05	1
Dokholyan	2	11.46	2	16.1	2	0.71	6	0.82	9	0	9	0.71	6	41.21	10	3.90E-02	2
Das	5	11.97	3	16.42	3	0.73	5	0.9	1	0.36	5	0.71	3	1.1	4	6.92E-02	3
Bujnicki	1	12.19	4	17.49	5	0.7	7	0.82	10	0	10	0.7	7	14.72	8	8.71E-02	4
Das	2	12.2	5	16.6	4	0.74	3	0.86	6	0.4	2	0.73	2	0.74	2	8.83E-02	5
Major	2	13.7	6	23.33	10	0.59	11	0.67	11	0	8	0.61	10	93.52	12	3.03E-01	6
Bujnicki	2	14.06	7	22.51	7	0.62	10	0.83	8	0	7	0.59	11	5.15	7	3.75E-01	7
Das	1	15.48	8	20.9	6	0.74	1	0.87	4	0.57	1	0.71	5	0	1	6.81E-01	8
Dokholyan	1	15.92	9	23.28	9	0.68	9	0.9	2	0	12	0.66	9	39.37	9	7.629E-01	9
Das	3	16.95	10	23.17	8	0.73	4	0.89	3	0.4	3	0.71	4	1.47	5	9.02E-01	10
Das	4	18.3	11	26.55	11	0.69	8	0.85	7	0.38	4	0.67	8	2.21	6	9.79E-01	11
Major	1	22.99	12	45.27	12	0.51	12	0.39	12	0	11	0.59	12	75.11	11	1.00E+00	12
<b>Mean</b>		<b>14.37</b>		<b>21.79</b>		<b>0.68</b>		<b>0.80</b>		<b>0.18</b>		<b>0.68</b>					
<b>Standard deviation</b>		<b>3.99</b>		<b>8.69</b>		<b>0.07</b>		<b>0.14</b>		<b>0.22</b>		<b>0.05</b>					
												<b>X-Ray Model</b>		<b>1.83</b>			

**Table 4**

	P1xP2	P1xP3	P1xP3a	P1xP3b	P2xP3	P2xP3a	P2xP3b	P3xP3a	P3xP3b	P3axP3b
3_bujnicki_1.dat	19.8	22.5	37.6	46.2	<b>7.0</b>	22.2	41.2	<b>11.2</b>	27.2	17.7
3_bujnicki_2.dat	27.3	37.3	49.4	68.2	20.1	32.3	47.4	<b>8.1</b>	21.5	16.0
3_chen_1.dat	<b>11.3</b>	<b>8.7</b>	26.0	28.3	<b>5.1</b>	18.6	26.8	<b>12.0</b>	<b>14.6</b>	<b>14.8</b>
3_das_1.dat	31.7	36.6	48.9	71.2	18.3	35.6	65.1	<b>11.9</b>	30.1	18.9
3_das_2.dat	30.7	32.9	34.1	34.0	24.9	32.9	27.0	<b>10.4</b>	<b>12.6</b>	<b>13.4</b>
3_das_3.dat	29.9	33.7	43.9	59.1	23.2	35.9	45.1	<b>13.3</b>	21.9	<b>13.5</b>
3_das_4.dat	33.1	36.5	54.0	71.3	<b>13.1</b>	22.9	37.9	<b>9.1</b>	<b>13.6</b>	<b>10.8</b>
3_das_5.dat	30.4	34.2	39.1	45.6	25.9	35.1	43.4	<b>8.9</b>	<b>13.5</b>	<b>11.8</b>
3_dokholyan_1.dat	34.9	21.5	32.9	59.4	<b>12.1</b>	28.3	32.9	<b>14.7</b>	26.8	25.9
3_dokholyan_2.dat	29.0	16.8	21.2	45.5	<b>13.8</b>	24.8	35.0	<b>10.1</b>	20.9	23.8
3_major_1.dat	23.0	40.3	44.4	46.4	27.5	43.3	56.5	27.9	53.4	48.2
3_major_2.dat	27.6	26.8	44.2	66.1	<b>9.6</b>	25.1	37.9	<b>10.4</b>	19.4	18.5

**Table 5**

	Problem 3
3_das_5	2.842888
3_das_4	2.928573
3_bujnicki_2	3.042605
3_chen_1	3.703777
3_das_2	3.915769
3_dokholyan_2	4.138209
3_bujnicki_1	4.253633
3_major_2	4.447882
3_das_1	4.554707
3_das_3	4.681876
3_dokholyan_1	5.821877
3_major_1	17.289223

## Chapter 4

# Annotation of ncRNA genes

Genome annotation, i.e., the process of attributing a function to a DNA segment localized in a genome<sup>1</sup>, is a major task of any full genome sequencing project. The annotation process consists of localizing the precise genomic location of each gene, classifying them by family and function, and identifying their products by homology with closely related species. The accurate and comprehensive annotation of a genome is an invaluable resource to the research community as it provides the data to support posterior studies, e.g., the presence or absence of a single gene can provide useful hints about the cell lifestyle, its metabolic pathways and functioning mechanisms; the evolutionary history of a species – its origins, genome dynamic and relationship with other species – can be largely informed by the order of genes in synteny groups and their distribution across chromosomes.

The concept of gene has been changing over time (Noble, 2008) and the annotation process has followed this change. A gene is no longer – if it ever was – a strict synonym of “protein coding region”, thus, annotation projects, must do more than locating the coding components of the genome. They must identify a plethora of other components such as, non coding transcripts, transposable elements, repetitive elements, introns and so on. The changes in the gene concepts must be accompanied by changes in annotation tools and methodologies. While automatic protein coding gene finding is a well established field (Harrow et al., 2009), reliable tools to discover other types of genes – in particular ncRNA genes – are relatively recent and their integration in full genome annotation pipelines is not yet a common place.

This chapter describes a pipeline for ncRNA gene discovery that integrates publicly available tools in order to improve genome annotation projects with homology and *de novo* gene discovery capabilities. The proposed pipeline was applied to 15 yeast genomes in the context of the Génolevures

---

<sup>1</sup>Here I deliberately simplified the definition of annotation as many other significant genomic features are localized during the annotation process (e.g., centromeres, telomeres, regulatory elements and other genomic structures). Those features, however, are beyond the scope of the present work.

project (Souciet et al., 2000) and the corresponding results will be described in Chapter 5.

## 4.1 Introduction

In recent years several new specific ncRNA discovery tools became available. As I will describe in this chapter, each of these tools focuses either on a particular discovery approach or on a specific ncRNA family. If, on one hand, this “*approach vs. family*” specialization improves the performance of each tool, on the other hand, it narrows their domain of application. As whole genome annotation projects require the discovery of as many as possible genes, the practical solution is to combine several available tools in order to maximize the gene discovery.

Another important issue to be addressed in whole genome annotation projects has to do with the total human effort spent on the process. Automatic tools usually produce many thousands of candidates, a number much larger than the real number of genes to be found. The classification of each candidate as a *bona fide* gene or a false positive prediction will be, at the very end, an human expert decision. Therefore, automation becomes essential to reduce the human effort. Some examples of tasks amenable to automation include: batch running of the available search tools; parsing the results and collecting the produced candidates; filtering and selecting the best among those candidates; and summarizing all the results in a human readable fashion, allowing the curator to take a quick, yet informed, decision. This chaining of automatic tasks is commonly known as an “annotation pipeline” (in the remaining of the text, for economy sake, I will refer to it simply as “pipeline”).

Although several ncRNA discovery pipelines have been proposed before (Yao et al., 2007; Noirot et al., 2008), the following particularities of the present project led us to explore a specific approach:

- (i) From the start of the project 10 yeast genomes were available (to which 5 new ones were added) and more than 30 new genomes are being sequenced. Consequently I was interested in applying some newly developed *de novo* search techniques (Washietl et al., 2005) that take advantage of multiple sequence alignments of related species;
- (ii) It is known that some yeast ncRNAs present particularly large extended domains (Kretzner et al., 1990; Kachouri et al., 2005). A particular treatment of the candidate genes is required in order to avoid missing those sequences and detecting their correct limits;
- (iii) Previous knowledge about the organization of ncRNA genes on yeasts – synteny relationships and polycistronic gene clusters (Souciet et al.,



2009) – should be incorporated in the validation part of the pipeline in order to improve the automatic selection of candidates.

## 4.2 The Pipeline

The developed pipeline consists of five main components (see Figure 4.1):

- **Data collection:** Retrieves, from external sources, the available data required for the annotation, e.g., previously existing annotations, ncRNA alignments, taxonomic information, ...;
- **Homology search:** Performs ncRNA gene search based on homologous sequence information;
- ***de novo* search:** Performs ncRNA gene search based on characteristic features of ncRNA sequence, with no homolog sequence comparisons;
- **Hit processing and candidate generation:** Collects all the produced hits, join them in candidates;
- **Candidate validation:** Performs an automatic validation preparing the candidates for the final manual validation.

In the following sections I will describe each of the enumerated components.

### 4.2.1 Terminology

At this point a terminology remark is necessary. In the present text some terms are used with a specific meaning that can, eventually, differ from the common language meaning of these words.

**Contig:** Continuous sequence of nucleotides resulting, usually, from genome sequencing. A whole genome is composed by several contigs. Ideally each contig will correspond to a chromosome, but in many cases, for technical reasons of the assemblage process, a chromosome is broken into several contigs.

**Hit:** Genome location identified by a search tool as an eventual ncRNA gene (or part of a gene). Hits are characterized by: (i) the name of the tool that produced it; (ii) the name of the contig, strand, start and end coordinates in which the hit was found; (iii) a score and/or an expected value of the occurrence which are statistical measures of the “quality” of the hit (see Appendix A).

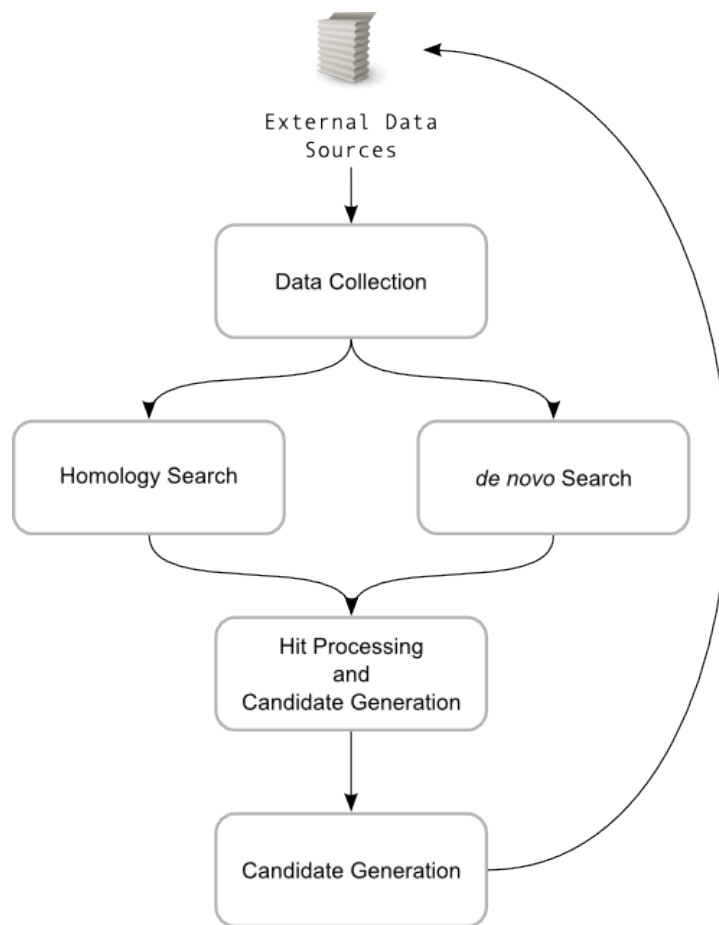


Figure 4.1: The five main components of the annotation pipeline.

**Candidate:** Set of hits corresponding to the same ncRNA gene that occur close together in the same region of the genome. Like a hit, a candidate is defined by: (i) the name of the contig, strand, start and end coordinates corresponding to the region where the contained hits were found; (ii) an E-value that corresponds to the highest E-value of all of the contained hits. It is frequent for a candidate to contain a single hit.

**Valid candidate:** A candidate that was considered a *bona fide* gene at the end of the validation process.

**Target sequence:** A sequence in which ncRNAs gene will be searched. In particular, the target genome is the genome being annotated.

**Query:** a model of a ncRNA gene (e.g., homologous sequence, covariation model, ...) that will be searched for in the target sequence.

#### 4.2.2 Data Collection

The main goal of the data collection component is to obtain and keep up-to-date all the data required for posterior phases of the pipeline such as the homology search and candidate validation. The main data sources, for the proposed pipeline, are the Rfam database (Gardner et al., 2009) and the Génolevures Consortium database (Sherman et al., 2004) from which the following data is retrieved:

- Rfam database:
  - ncRNA sequence alignments and sequence information;
  - Taxonomic information about the sequences in the alignments;
  - Updated versions of the covariance models available at Rfam. These models will be used for homology search with **INFERNAL** (Nawrocki et al., 2009).
- Génolevures Consortium database:
  - Genome files;
  - Coordinates of ncRNA genes resulting from previous annotations;
  - Raw ncRNA sequences of yeast genomes to be used in for homology search with **BLAST** (Altschul et al., 1990);
  - Open reading frame coordinates in order to improve the gene validation process (see Section 4.2.6).

### 4.2.3 Homology search

Homology search is a type of gene discovery strategy that uses the information from homologous genes (the query) to search the genome of interest (the target) for possible candidates, e.g., The BLAST sequence search tool (Altschul et al., 1990) is arguably the best known example of an homology search tool. When using BLAST the queries are sequences of homologous genes. In some other cases different kind of models can be used as queries, such as covariance models (Eddy and Durbin, 1994) that rely on homologous structural information to perform the search.

A general advantage of the homology search approach is that if the E-value of a candidate is low enough one can be quite sure that the candidate belongs to the family of the query model. Although this statement seems obvious, it does not stand for the *de novo* approach where the nature of a candidate is frequently not obvious and requires experimental validation. On the other hand, the drawback of the homology search is that one will only find what one is looking for, i.e., no new ncRNA family can ever be found with this approach. As expected, this is precisely the advantage of the *de novo* search.

As described in the introduction, ncRNA genes, present a number of specific features that render the homology search particularly difficult:

**(i) Low sequence conservation between homologous ncRNAs:** ncRNA conservation is constrained mainly by structural requirements allowing an important sequence variation even between homologous genes in closely related species. Pure sequence comparison based homology search will fail if sequence conservation is too low.

A number of specific search tools for ncRNA search has been developed to address this issue (see Table 4.1). Family specific tools resort to explicit sequence and secondary structure knowledge, about each ncRNA family, in addition to sequence information, to find genes of that particular family. General purpose tools use covariation information, gathered from multiple sequence alignments, to infer covariance models (Eddy and Durbin, 1994) that are used as queries to search the target genomes. Although any particular covariance model is also family specific, the tool can be used to search for any ncRNA given the appropriate model.

An extreme example of such a low sequence conservation, for which no family specific tool is available at the moment, is the telomerase ncRNA (see Section 5.5.1)

**(ii) Strong sequence conservation limited to short regions:** In many ncRNAs a strong sequence conservation occurs only in very short

Class	Tool	Purpose / Approach
Family specific	tRNAScan-SE (Lowe and Eddy, 1997) snoScan (Lowe, 1999) snoGPS (Schattner et al., 2004) snoReport (Hertel et al., 2008) SRPScan (Regalia et al., 2002)	tRNA snoRNA C/D snoRNA H/ACA snoRNA both SRP
General purpose	INFERNAL (Nawrocki et al., 2009)	Covariation Models

Table 4.1: ncRNA homology search tools.

regions related to inter molecular interactions. If the conserved region is too small, the produced hits will also have a small E-value and risk to be discarded by the filtering steps of the algorithm. A good example are snoRNAs in which the guide sequence presents more than 90% of conservation in a region of a few bases (see Table 4.2).

Several approaches can be combined to address this issue: use ncRNA specific tools as in the previous point; use a generic sequence search tool but establish families specific E-value thresholds; incorporate synteny information whenever possible.

### (iii) Large sequence insertions/deletions in homologous molecules:

Some ncRNA molecules have the ability to conserve a core three dimensional structure in spite of large sequence insertions/deletions (see Section 5.6). Searching target sequences, which contain those large insertions, using queries from homologous genes without those insertions, will produce – in the best cases – a series of relatively low E-value hits scattered along a much larger extension than the query itself. Simple hit clustering techniques that will group those scattered hits in single candidates can be used to overcome this problem (see Section 4.2.5).

#### 4.2.4 *de novo* Search

As mentioned in the previous section, the homology search strategy will discover genes from ncRNA families for which at least one homologous gene is already known. On the other hand, the goal of the *de novo* search strategy is to search for ncRNAs for which no homologous gene or family are given.

In the field of *de novo* protein gene discovery, the search tools resort to a series of statistical features of protein coding genes to discriminate between coding and non-coding regions (Harrow et al., 2009). Those signals can be: the occurrence of long open reading frames between start and stop codons; a significant statistical bias in the codon usage; differences on nucleotide

Organism	Seq. id. (%)	100% length	BLAST E-value	INFERNAL E-value
<i>Candida glabrata</i>	45%	14	n/a	$9.0 \times 10^{-19}$
<i>Zygosaccharomyces rouxii</i>	62%	17	n/a	$3.0 \times 10^{-34}$
<i>Saccharomyces kluyveri</i>	64%	20	$1.8 \times 10^{-3}$	$2.1 \times 10^{-47}$
<i>Kluyveromyces thermotolerans</i>	63%	18	n/a	$2.8 \times 10^{-41}$
<i>Eremothecium gossypii</i>	60%	18	n/a	$8.9 \times 10^{-39}$
<i>Kluyveromyces lactis</i>	63%	19	$2.7 \times 10^{-2}$	$6.6 \times 10^{-45}$
<i>Debariomyces hansenii</i>	59%	19	$7.6 \times 10^{-3}$	$1.0 \times 10^{-41}$
<i>Pichia sorbitophila</i>	59%	16	$1.8 \times 10^{-1}$	$2.3 \times 10^{-36}$
<i>Arxula adeninovorans</i>	55%	18	$1.1 \times 10^{-1}$	$2.7 \times 10^{-25}$
<i>Yarrowia lipolytica</i>	53%	18	$2.0 \times 10^{-1}$	$4.1 \times 10^{-19}$

Table 4.2: Low homology example: Result of BLAST and INFERNAL searches for snR43 snoRNA gene. The snR43 presents low overall pairwise sequence identity between *S. cerevisiae* and each of the other budding yeasts. Although the guide sequence is strongly conserved it is not enough to produce significant BLAST hits. For 4 of the yeasts BLAST will not return any hit, while for another 3 the E-value is too high to be selected ( $>10^{-1}$ ). INFERNAL search, on its turn, will find the snoRNA in all genomes (Columns – “Organism”: name of the budding yeast being compared; “Seq. id (%)”: percentage of sequence identity between the organism and *S. cerevisiae*; “100% length”: length of the longest 100% conserved sequence; “BLAST E-value”: E-value of best true positive hit from BLAST; “INFERNAL E-value: E-value of best true positive hit from INFERNAL.)

composition between coding and non-coding regions. Unfortunately, ncRNA genes do not present those types of features.

To tackle this problem *de novo* ncRNA search tools rely on some observations and assumptions:

- (i) ncRNA belonging to the same family will share the same core secondary structure;
- (ii) Due to compensatory mutations in helical regions of the molecule, the alignment of homologous ncRNAs sequences will reveal co-variation between the column of the alignment corresponding to canonical base pairs of the helical regions, which, consequently, allows the inference of the secondary structure of the molecule;
- (iii) The minimum free energy (MFE) obtained by folding an ncRNA sequence with standard secondary structure folding algorithms will be lower than the average MFE obtained by folding random sequences with the same nucleotide composition.

Notice that none of these assumptions is an absolute truth, in particular, the third one has been the motivation for an interesting debate (Rivas and Eddy, 2000; Clote et al., 2005). However, when used together they can be good indicators of the occurrence of a ncRNA in a given region of the target genome. The first tool to approach this problem, QRNA (Rivas and Eddy, 2001), uses the observed pattern of covariance in pair-wise alignments to classify the alignment as “coding”, “structural RNA” or “something else”. A more recent tool, RNAz (Washietl et al., 2005), also uses covariance information, but instead of a pair-wise comparison it uses multiple sequence alignments. In addition, RNAz computes the MFE Z-score (see Appendix A) of a given alignment in comparison with the MFE distribution of random sequences of the same nucleotide composition and relies in a support vector machine to discriminate between coding and non-coding sequences.

The proposed pipeline performs a whole genome multiple sequence alignment, on all available related species, using the Threaded Blockset Aligner (Blanchette et al., 2004), and use the resulting multiple sequence alignments to feed the RNAz tool. The obtained candidates are filtered in order to exclude:

- Candidates with an RNAz score <50%;
- Candidates totally contained in known protein coding regions.

#### 4.2.5 Hit processing and candidate generation

After collecting all the hits produced in the search stage one has to filter and cluster them to obtain the final candidates for validation. The filtering step consists in discarding all the hits that:

- Present an E-value smaller than a determined threshold;
- Have a genome location that overlaps an annotated open reading frame (ORF).

The E-value threshold depends on the tool that produced the hit (in Chapter 5 we present the analysis of the E-value thresholds for both BLAST and INFERNAL with real world data). The ORF overlapping criteria was imposed as it is not expected that a ncRNA would occur in a protein-coding region. Notice that only the hits overlapping the region between the start and stop codons are discarded. This way, the hits that overlap the 5' and 3' UTR regions are kept for they can correspond to regulatory regions of protein coding genes and, eventually, to *bona fide* candidates according to our definition of ncRNA (see 1.3). Notice that hits that overlap an ORF, but at the same time present a particularly low e-value, e.g., e-value  $<10^{-5}$ , must be analyzed manually in order to detect eventual annotation errors. The hits that survive the filtering step are clustered according to Algorithm 1.

---

**Algorithm 1** Hit clustering
 

---

```

# Assigns to each hit its own individual cluster;
for  $hit_i$  IN  $hitlist$  do
  #  $cluster_i$  will assume the gene family and coordinates of  $hit_i$ 
   $cluster_i \leftarrow hit_i$ 
end for
# Performs a pairwise comparison between all clusters:
repeat
   $merge = \text{false}$ 
  for  $cluster_i$  IN  $clusterlist$  do
    for  $cluster_j$  IN  $clusterlist$  do
      # Both clusters have the same family, contig and strand properties?
      # Clusters are less than 2048 bases apart?
      if  $Properties(cluster_i) == Properties(cluster_j)$  AND
         $Distance(cluster_i, cluster_j) < 2048$  then
         $merge = \text{true}$ 
         $cluster_i \leftarrow cluster_j$ 
         $cluster_j \leftarrow \emptyset$ 
      end if
    end for
  end for
until  $merge == \text{false}$ 

```

---

Each cluster that results from this procedure will constitute a candidate gene. This way, hits corresponding to conserved regions of a gene, separated



by big insertions (or deletions in the query gene), can be correctly assigned to the same candidate, simplifying the validation procedure and largely improving the determination of the flanking coordinates of the gene (see Figure 4.2).

#### 4.2.6 Automatic Candidate Validation

Once all of the hits, produced by the search tools, have been filtered and clustered in their respective candidate genes it is time to choose which of the candidates will constitute the genome annotation as *bona fide* genes and which of them will be discarded by lack of support. The complete validation process occurs in two steps, a first automatic validation process and a final manual verification of the remaining candidates. Here I describe the automatic validation, while the manual validation is described in the next section.

The automatic candidate validation applies a number of knowledge-based criteria to perform a last filtering step before manual validation. In short, I try to discard clearly negative candidates according to some previous knowledge about these genes in order to ease the manual validation step. The acceptance criteria used in this step depends strongly on the species and type of genome being annotated and should be adapted accordingly.

**Strong conservation in local short regions:** Many ncRNA molecules have at the same time a overall low sequence conservation and local high conservation in very short regions. Those regions are essential to the molecule function and it is highly unlikely that they would diverge significantly in a single specie. Thus, the selected candidates must display a strong conservation in those regions.

**Conservation in complementary regions:** Some important classes of ncRNAs interact by base pairing, in trans, with other RNA molecules, e.g. snoRNAs and rRNAs; U2 and U6 snRNAs. Those interaction sequences are usually conserved in closely related species but can diverge in a longer time span. These sequences, however, tend to co-evolve in order to maintain the base pairing. Thus, if the target sequence is known (which is frequently the case as most of the targets are located in rRNAs and other easily found genes) the guide sequence must be complementary and easily validated.

**Synteny and Polycistronic Clusters:** Synteny, i.e., the co-localization of genes in the same region of a genome can be conserved across closing species. Although the precise reason for this co-localization is unknown, polycistronic genes, i.e., genes expressed in the same transcript (Souciet et al., 2009) or genes participating in the same interaction network (Teichmann and Veitia, 2004) are known to be syntenic. When synteny is con-

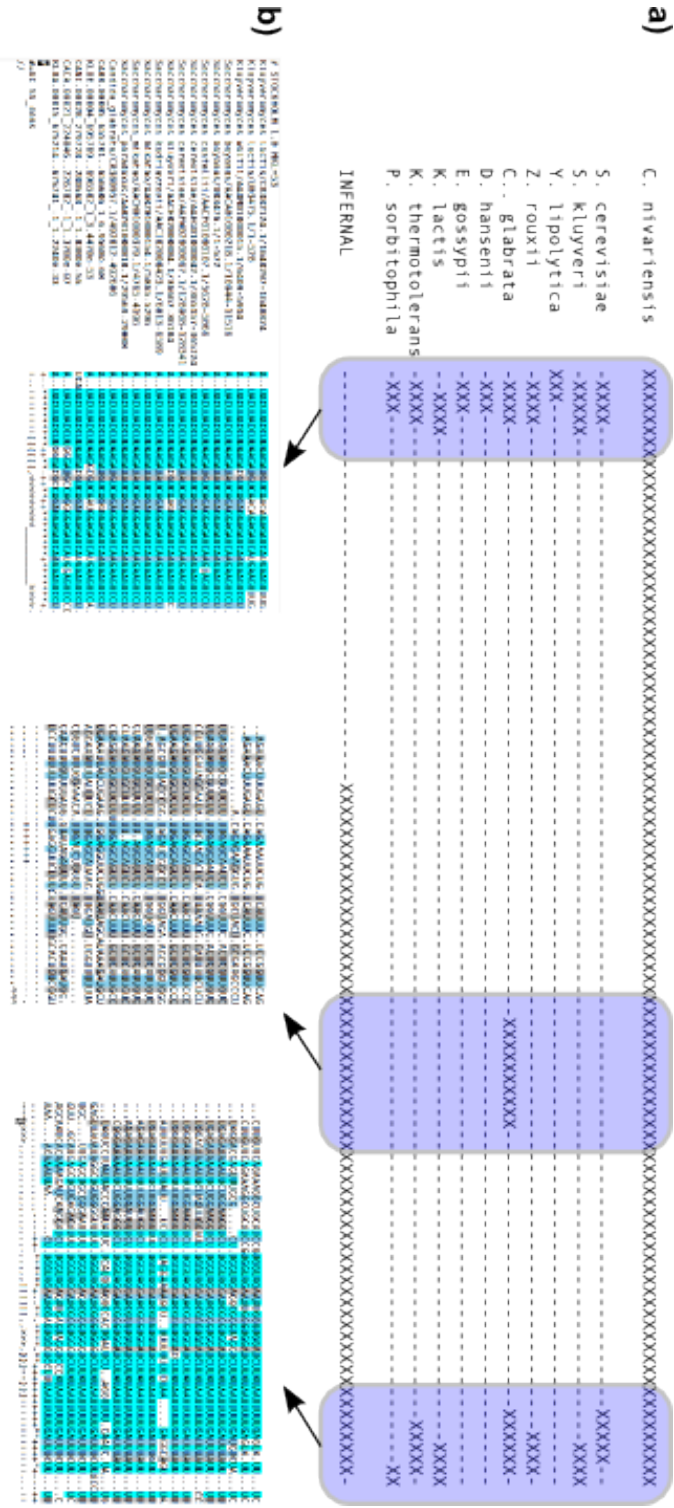


Figure 4.2: Hit cluster of the U1 snRNA of *C. niuariensis* candidate. Due to long insertions (see Section 5.6) The 19 obtained hits cover only part of the complete molecule. a) Schematic representation of the BLAST and INFERNAL hits. First line: target sequence (845 nts); Lines starting with species names: hits produced by BLAST for each query sequence (the U1 homologue of each species was used as query); Last line: INFERNAL hit. b) Snapshots of the full alignment showing the high sequence identity in relative short regions. The alignment in the center presents high sequence identity only for some sequences.

served across many species belonging to the same clade it can be used as an additional support to candidate validation and missing gene identification by narrowing the region of search to the syntenic region.

### 4.3 Manual candidate validation

In every scientific field experimental data is absolutely required as the definitive validation of any prediction. This premise also applies to genome annotation in the sense that all annotated genes are no more than “mere” predictions, until an eventual posterior experimental validation. Given the huge difference between the rate of new gene annotations and the ability to experimentally validate them, most of these annotations will remain “experimentally unvalidated” for long periods of time, although available to the community in public databases. If this is not a problem for the majority of the annotations (that are statistically very likely to represent real genes), in some cases errors occur. Thus, the help of a human expert (or curator) is essential for the quality of the final annotation in two ways: First, it reduces the number of incorrect annotations, and second, it allows the detection of errors in the steps upstream to the validation phase, e.g., pipeline software errors, sequencing errors, genome assemblage problems . . .

The manual validation of candidates is the last step of the annotation pipeline. Manual validation implies a one-by-one verification of all candidates and requires a significant human effort.

Two broad types of validation are addressed by the curator in this phase:

- Validate the selected candidates as *bona fide* ncRNA genes;
- Verify if there are any expected/obvious ncRNA genes (e.g., an essential and highly conserved gene, . . . ) missing and why.

As it is not possible to specify a detailed and objective recipe for manual candidate validation (otherwise it would be possible to implement it as an automatic procedure), the pipeline generates two data views to ease and speed up the validation process:

- Hit location maps;
- Multiple sequence alignments with structural information.

The hit location maps provide a simple view on how the hits that constitute a candidate are distributed along the sequence (see Figure 4.3). The curator can easily assess how many hits support the candidate and which are their E-values, scores and which regions of the candidate are more or less supported by hits.

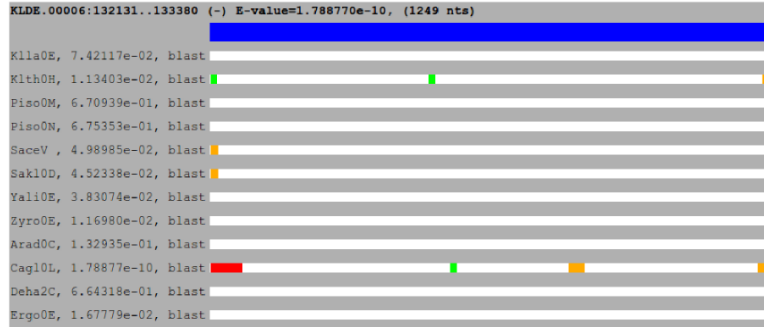
**Candidates for: RF00009****(NAME:RNaseP\_nuc|GNLVRS:1|CLAN:2|MASTER:-)**

Figure 4.3: Example of a hit location map. The *K. delphensis* RNaseP candidate is shown. Each hit is identified by the source species, E-value and search tool. The colors of the hits indicate the E-value (red: lower; orange: medium; green: higher).

The multiple sequence alignments allow the curator to compare the candidate sequence with all available homologous sequences. The comparison is done in terms of sequence conservation and structure compatibility (see Figure 4.4). Although sometimes a manual realignment is required, in most cases the provided automatic alignment is enough to confirm or discard a candidate.

#### 4.4 Technical implementation

The annotation pipeline was implemented as a series of Python scripts serving as logical glue between: the external data sources; the internally developed algorithms; and the third-party tools. The developed scripts are organized in seven functional modules according to their responsibilities. In this section I enumerate each of these modules.

**Raw data retrieval and preprocessing:** As referred in Section 4.2.2, almost all of the data used in the pipeline originates from two major data sources: the Rfam database and the Génolevures consortium. While some of the data is ready for use immediately after download, some other data needs to be processed before it can be used. Downloading the data and preprocessing it are the main tasks of this module.

*Sequence alignments:* Sequence alignments are a basic resource for sequence comparison and analysis. In order to be analyzed and accepted as a *bona fide* gene, all candidate sequences must be included in a sequence alignment with their respective homologues. The most complete source of

```

# STOCKHOLM 1.0
Athaliana  AAGCACCAGGGGUCUU.G.....C.AGGCUGAGAA.AGUCCUUGAACUCUGAACAGGGUAAUGCUCGGCAGGGA.GUGUGC
Arad_1     GGUUUC.ACGGGUGCCU.AGG.....CUA.GUGCUGAGAUGAUACCGU.UGAACUCGAUCAAGGUCUAUCUUGCGUGAGAA.AGCACC
Arad_2     GACACU.GCGGGUGCCA.....UGGCUGAGAUAUACCGUUGAACUUGAUGGGGUUAAAGACCUACGGAAGAAUUGUGUC
Arad_3     CGCAUG.GCCGGUGCCU.....GUGCUGAGAUAUACGGC.AAAACUUGAUCAAGUAAUUAUCUAGCGAAAGAA.CAUGCG
Yali_1     CGCACU.GCGGGUCUU.G.....C.AGGCUGAGAAAACACCGC.CGAACUCGACCAAGAUAAUUCUUGCGUGAGAA.AGUGCA
Yali_2     UGCACU.GCGGGUGUU.UUG.....CAA.AGGCUGAGAUAUACCGC.CACACUCGAACAAGAUAAUUCUUGCGUGAGAA.AGUGCG
Deha_1     GAGACUAGCGGGUGUU.UUC.....GAA.AGA.UGAGAAUACCGUUUGAACUCGAUCAAGUUUGAUUCUUGCGUGAGAA.AGUUUU
Ecoli_1    ACGACU.CGGGGUGCC.UUC.....GAA..GGCUGAGAA.AUACCGUAUACUCUGAUUGGAUAAUUCGCGUAGGGG.GCCAAA
Ecoli_2    ACGACU.CGGGGUGCC.UUC.....GAA..GGCUGAGAA.AUACCGUAUACUCUGAUUGGAUAAUUCGCGUAGGGG.AGUCAC
Ecoli_3    UCUCAA.CGGGGUGC.CACGC.....GCGUGCGUCGAGAAAUAACCGUCGAAACUCGAUCCGGUAAACCGCGCGAAGGGA.UUUGAG
#=GC SS_cons  <<<<<<. <<<<*<<<<<<<<<<<<<<<<. >>>>>>*****.,>. *>>>>.....*<<<<<<. *.....>>>>. *>>>>
#
//

# STOCKHOLM 1.0
Athaliana  AAGCACACGGGGUCUU.G.....C.AGGCUGAGAA.AGUCCUUGAACUCUGAACAGGGUAAUGCUCGGCAGGGA.GUGUGC
Arad_1     GGUUUC.ACGGGUGCCU.AGG.....CUA.GUGCUGAGAUGAUACCGU.UGAACUCGAUCAAGGUCUAUCUUGCGUGAGAA.AGCACC
Arad_2     GACACU.GCGGGUGCCA.....UGGCUGAGAUAUACCGUUGAACUUGAUGGGGUUAAAGACCUACGGAAGAAUUGUGUC
Arad_3     CGCAUG.GCCGGUGCCU.....GUGCUGAGAUAUACGGC.AAAACUUGAUCAAGUAAUUAUCUAGCGAAAGAA.CAUGCG
Yali_1     CGCACU.GCGGGUCUU.G.....C.AGGCUGAGAAAACACCGC.CGAACUCGACCAAGAUAAUUCUUGCGUGAGAA.AGUGCA
Yali_2     UGCACU.GCGGGUGUU.UUG.....CAA.AGGCUGAGAUAUACCGC.CACACUCGAACAAGAUAAUUCUUGCGUGAGAA.AGUGCG
Deha_1     GAGACUAGCGGGUGUU.UUC.....GAA.AGA.UGAGAAUACCGUUUGAACUCGAUCAAGUUUGAUUCUUGCGUGAGAA.AGUUUU
Ecoli_1    ACGACU.CGGGGUGCC.UUC.....GAA..GGCUGAGAA.AUACCGUAUACUCUGAUUGGAUAAUUCGCGUAGGGG.GCCAAA
Ecoli_2    ACGACU.CGGGGUGCC.UUC.....GAA..GGCUGAGAA.AUACCGUAUACUCUGAUUGGAUAAUUCGCGUAGGGG.AGUCAC
Ecoli_3    UCUCAA.CGGGGUGC.CACGC.....GCGUGCGUCGAGAAAUAACCGUCGAAACUCGAUCCGGUAAACCGCGCGAAGGGA.UUUGAG
#=GC SS_cons  <<<<<<. <<<<*<<<<<<<<<<<<. >>>>>>*****.,>. *>>>>.....*<<<<<<. *.....>>>>. *>>>>
#
//

```

Figure 4.4: Show an example of the same alignment in sequence and structure views. Snapshots of the RaLee alignment editor (Griffiths-Jones, 2005).

ncRNA sequence alignments is the Rfam database (Gardner et al., 2009). The Rfam provides comprehensive sequence alignments for all known families of ncRNAs<sup>2</sup>. Browsing the Rfam web site, the user can explore the available alignments in many different views and formats. An automatic pipeline, however, will require local copies of the alignments to speed up its tasks. Rfam provides the full set of alignments in a single compressed text file (“Rfam.full.gz”) that must be broken into individual files – one for each family – and normalized – one sequence per line – to be of use.

*Covariance models:* Covariance models for all Rfam families are available for download from the Rfam web site. The INFERNAL package provides tools to build (`cmbuild`) and calibrate (`cmcalibrate`) new models directly from structural alignments. While the building process is fast even for big alignments, users should be aware that the calibration process (required to produce hit E-values in search results) is particularly time consuming.

*Genome files:* The genome files can be downloaded from the Génolevures web site. The target genomes are not usually available for download and must be obtained separately. Genomes are simple fasta files with one sequence per contig.

*ncRNA and Open Reading Frames coordinates:* Any previous annotation information is key to obtain search queries and validation data. This data can be downloaded from the Génolevures web site in `gff` format<sup>3</sup>.

<sup>2</sup>The role of Rfam as a reference for ncRNA families is such that I usually use the “Rfam family code” as a synonym of the family name. The Rfam code has the following format: RFNNNNN (RF followed by 5 digits), e.g., RF00004: U2 snRNA; RF0127: small nucleolar RNA snRN37.

<sup>3</sup>See <http://genome.ucsc.edu/FAQ/FAQformat.html> for details on the GFF format.

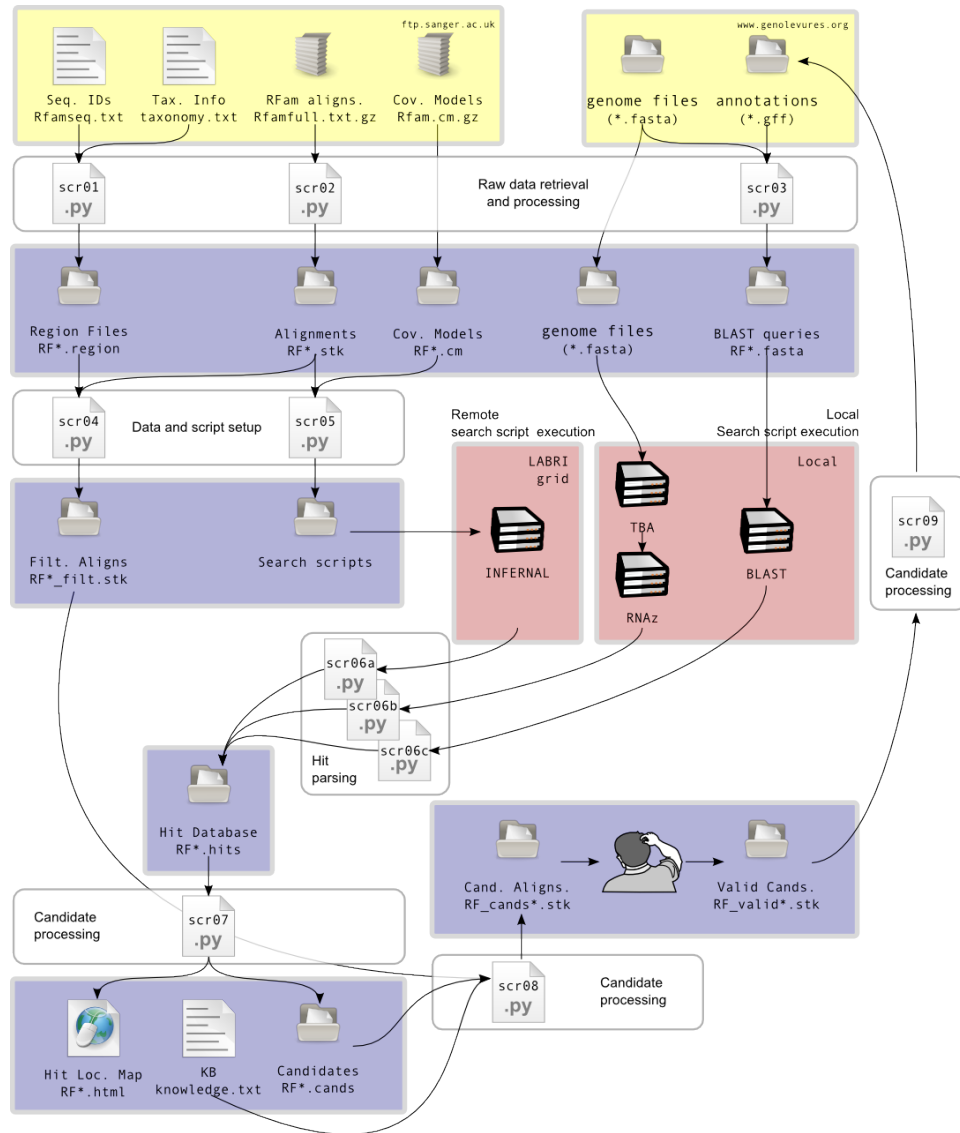


Figure 4.5: Complete flow of the annotation pipeline

Any additional data sources can be easily integrated in the pipeline as long as they could provide data on either `stockholm`, `gff` or `fasta` formats.

**Data and script setup:** The search and validation processes require a number of data files and scripts that must be prepared case by case depending on the actual query models being searched. As part of the search is performed in a cluster<sup>4</sup> running the *Sun Grid Engine*<sup>5</sup>(SGE) workload management solution, a number of SGE scripts must be built to execute the INFERNAL searches. Those scripts are produced on demand according to the genomes to search and the available covariance models.

Another setup task performed by this module is to produce filtered alignments containing sequences only from phylogenetically related species (from the Dikarya subkingdom in the present case). These filtered alignments will be used to compare the obtained candidates for validation.

**Hit parsing:** Each search tool produces hit data in its own format. This data must be parsed and converted to an homogeneous format so it can be further processed. The currently supported hit data formats are: BLAST, INFERNAL and RNAz.

**Hit processing:** This module performs all the tasks described in detail on Section 4.2.5. The goal is to produce, from the initial hits, the final candidates and the hit location maps in HTML for user validation.

**Candidate processing:** The step immediately before human intervention is to align each candidate with its respective homologues. This module will produce, for each ncRNA family, a sequence alignment containing the homologue sequences from closely related sequences (the closest one can obtain from Rfam) and the candidates for that family. User can then evaluate each candidate in the context of the appropriate alignment.

**Annotation update:** After human validation, the last step of the pipeline is to produce the ncRNA annotation data in a format that could be integrated in the databases for future use. The candidates validated by the human curator are stored in a GFF file with their coordinates and relevant information and sent back to the Génolevures consortium so it can be part of the next annotation phases.

---

<sup>4</sup>Cluster time was kindly provided by LABRI – *Laboratoire Bordelais de Recherche en Informatique* (<http://www.labri.fr>).

<sup>5</sup>See <http://www.oracle.com/technetwork/oem/grid-engine-166852.html> for more information on SGE.

## Chapter 5

# Annotation of ncRNAs in Budding Yeasts

The previous chapter sets the stage for the effective usage of a ncRNA annotation pipeline, describing its motivations, functioning and technical details. The present chapter will describe the application of the pipeline to the annotation of several yeast genomes in the context of the Génolevures project. The Génolevures consortium (Sherman et al., 2004) brings together experts in genomics, yeast and bioinformatics with the goal of providing the biology community with annotated sequence data and annotation for a number of species of *Hemiascomycetous* yeasts. The consortium web site provides: “(...) *genetic element pages, orthologs defined by syntenic homology, protein families, a genome browser for interspecies comparison, and data sets for downloading.*”<sup>1</sup>.

At the moment, the consortium is concluding its activities. Started in 2000 the Génolevures gives now place to a new yeast sequencing project – the Dikaryome – which aims to sequence and annotate thirty new genomes of the Dikarya sub kingdom using the latest sequencing and high throughput techniques.

As members of the Génolevures consortium, our main responsibility was the discovery and annotation of ncRNAs. Due to the different delivery dates of each group of genomes, the annotation occurred in two phases: First I performed the annotation of the ten budding yeasts from Génolevures 2, Génolevures 3, and *E. gossypii*. Second I annotated five additional yeasts of the Nakaseomycetes clade sequenced in a separate sequencing program led by a Génolevures consortium member, C. Fairhead (Université Paris-Sud 11).

---

<sup>1</sup>Quoted from <http://genolevures.org>.



Organism	Year	Institution
<i>Saccharomyces cerevisiae</i>	1996	
<i>Ashbya gossypii</i>	2004	Biozentrum, Basel, Switzerland
<i>Candida glabrata</i>	2004	Génolevures
<i>Debaryomyces hansenii</i>	2004	Génolevures
<i>Kluyveromyces lactis</i>	2004	Génolevures
<i>Yarrowia lipolytica</i>	2004	Génolevures
<i>Lodderomyces elongisporus</i>	2007	Broad Institute
<i>Meyerozyma guilliermondii</i>	2007	Broad Institute
<i>Scheffersomyces stipitis</i>	2007	DOE Joint Genome Institute
<i>Vanderwaltozyma polyspora</i>	2007	Trinity College, Dublin, Ireland
<i>Candida tropicalis</i>	2009	Broad Institute
<i>Clavispora lusitaniae</i>	2009	Broad Institute
<i>Candida dubliniensis</i>	2009	U. of Aberdeen, Scotland
<i>Kluyveromyces thermotolerans</i>	2009	Génolevures
<i>Zygosaccharomyces rouxii</i>	2009	Génolevures
<i>Pichia pastoris</i>	2009	U of Gent, Belgium
<i>Candida albicans</i>	2009	Stanford Genome Technology Center

Table 5.1: Publicly available complete genomes of budding yeast (as of March 2011).

## 5.1 Budding yeasts

Budding yeasts are unicellular eukaryotes classified in the Fungi kingdom that constitute an important group of species from the scientific, economical and health care perspectives (Kurtzman and Fell, 2006). Budding yeasts belong to the *Saccharomycotina* subphylum (also known as Hemiascomycetes).

Yeasts are easily cultured in laboratory and offer a number of powerful manipulation and analysis techniques. Their small and compact genomes make them ideal models for comparative and evolutionary genomic studies (Souciet et al., 2000; Dujon, 2010). Since ancient times yeasts have been essential to the manufacturing of bread, beer and wine. Today, they are used in the industrial synthesis of many product (e.g. vitamins, ethanol, lipids, ...) (Sherman et al., 2004). Some yeast species are opportunistic human pathogens. For example, invasive candidiasis, caused by several species of the genus *Candida*, is a serious threat for hospitalized and immunocompromised patients (Pfaller and Diekema, 2007; Butler et al., 2009).

It is therefore not surprising that *Saccharomyces cerevisiae* was the first eukaryotic organism to be completely sequenced back in 1996 (Goffeau et al., 1996). To this day 17 budding yeast complete genomes are publicly available in the NCBI GenBank Database (Benson et al., 2008) (see Table 5.1).

## 5.2 The Annotated Genomes

During the ten years period of the Génolevures consortium, seven complete genomes were sequenced, annotated and made publicly available; two other genomes are about to be published. The Nakaseomyces species *C. glabrata* was sequenced and annotated by the Génolevures consortium and more recently five new Nakaseomyces species have been sequenced and annotated in a separate program (see Figure 5.1).

The sequenced species were chosen based on their phylogenetic representativeness, industrial interest, biomedical importance and the existence of significant genetic and molecular studies regarding them (Souciet et al., 2000). From an evolutionary point of view, the *Saccharomycotina* separated from the *Pezyzomicotina* (another *Ascomycetes* subphylum) between 400 Myr and 1000 Myr ago (Dujon, 2010), which makes them an interesting model for comparative genomic studies.

### 5.2.1 Data sources

The genome sequences of *Candida glabrata*, *Debariomyces hansenii*, *Kluyveromyces lactis*, *Kluyveromyces thermotolerans*, *Saccharomyces kluyveri*, *Yarrowia lipolytica*, *Zygosaccharomyces rouxii* and their respective annotations were obtained from the Génolevures web site (Sherman et al., 2004).

The annotations for the *Eremothecium gossypii* were obtained from the “Ashbya Genome Database” web site (<http://agd.vital-it.ch>) (Gattiker et al., 2007) and its complete genome is available at the NCBI GenBank (<http://www.ncbi.nlm.nih.gov>) (Benson et al., 2008) with the accessions NC\_005782 to NC\_005788.

The genome sequence of *Pichia sorbitophila* was provided as a personal communication from Veronique Leh Louis (Université de Strasbourg).

The genome sequence of *Arxula adeninovorans* was provided as a personal communication from Cécile Neuvéglise (INRA - AgroParisTech).

The genome sequences of *Candida bracarensis*, *Candida castelii*, *Candida nivariensis*, *Kluyveromyces bacillisporus*, *Kluyveromyces delphensis* were provided as a personal communication from Cécile Fairhead (Université Paris-Sud 11).

## 5.3 *Saccharomyces cerevisiae* – Reference Genome

Before starting an annotation project it is legitimate to ask the following – somehow naive – question “How many ncRNAs exist in a genome?”. The answer to this question should be the reference value to evaluate the success of any annotation. Unfortunately, this is an increasingly difficult question.

Estimates made for several higher eukaryotes indicate that the ratio between non-coding<sup>2</sup> and coding regions goes from 27:1 in human to 1.1:1 in

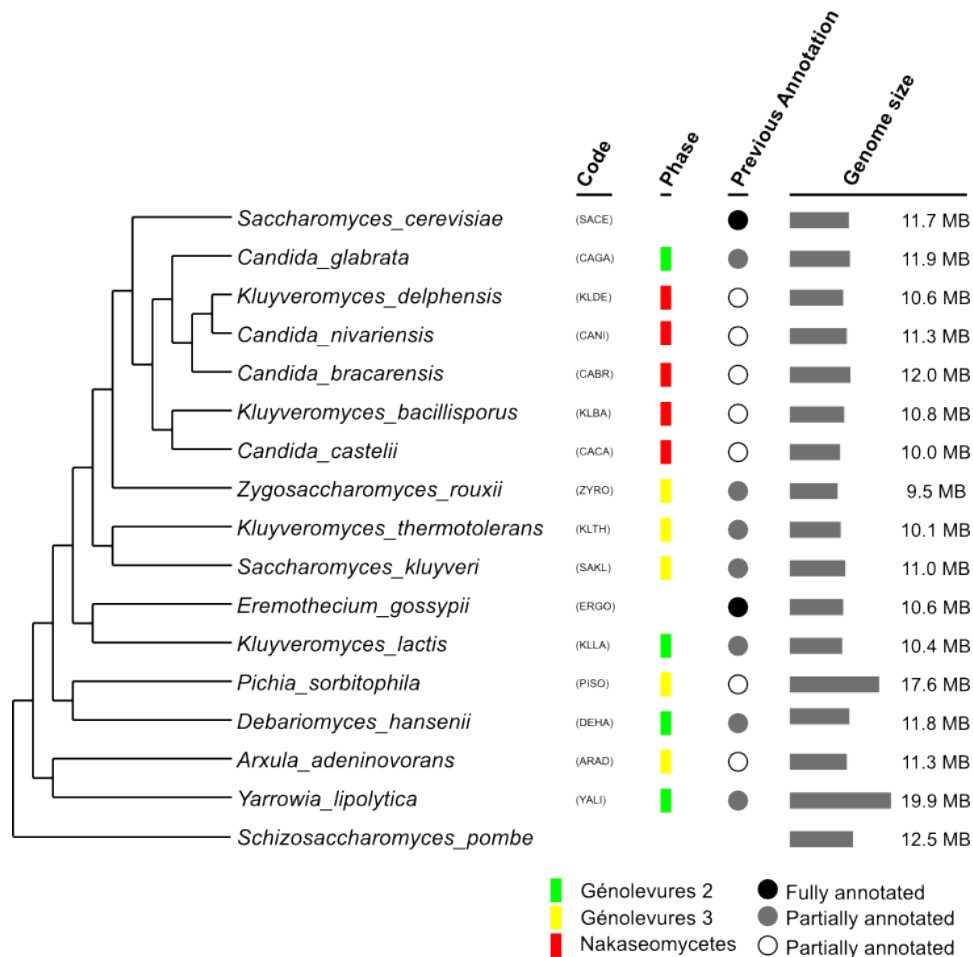


Figure 5.1: Annotated genomes. Phylogenetic relationship between the annotated genomes. “Code” is the working abbreviation of the genome, “Phase” is the Génolevures project phase in which the genome was available, “Previous Annotation” is the annotation status before the application of the pipeline and “Genome Size” is the size in millions of base pairs of each genome.

nematode (Frith et al., 2005). Additionally, the ENCODE project (ENCODE, 2007), estimates that 4.9% of the human genome is under selective pressure and 40% of it do not have any known function associated. Further experimental work with mice showed that some of those sequences are expressed in the brain and are regulated in a tissue-specific fashion during development (Mehler and Mattick, 2006; Dinger et al., 2008; Mercer et al., 2008; Amaral et al., 2009), raising interesting questions about possible functions of large, transcribed, non-coding regions of higher eukaryotic genomes.

In the absence of a definitive answer I chose a pragmatic approach and took the set of annotated ncRNAs from *S. cerevisiae* as reference. The relative success of the annotation pipeline was then measured as the proportion of *S. cerevisiae* ncRNA homologues found in each genome. This option is justified by the fact that the *S. cerevisiae* genome is small (two hundred times smaller than the human genome), very compact (about 3/4 of it corresponds to annotated protein genes) and, arguably, the most studied and well annotated buddy yeast (if not eukaryote) genome. Thus the current number of annotated ncRNAs in *S. cerevisiae* could be considered a reasonable proxy for the real number of ncRNAs in yeast in general.

It is important to remember that not all of *S. cerevisiae* ncRNAs would inevitably have an homologue in all studied species and, as will be shown, that some ncRNA with no known homologue in *S. cerevisiae* are indeed found in other species. Those cases, though, are restricted in number.

## 5.4 The annotation process

I separately describe the two annotation phases as the pipeline described in Chapter 4 was applied in a slightly different fashion between them. In the first phase I annotated the genomes from Génolevures 2, Génolevures 3, and *E. gossypii* and I performed a ROC (Brown and Davis, 2006) analysis of the obtained E-value in this phase. In the second phase I annotated five yeasts of the Nakaseomycetes clade and applied the hit clustering step. Additionally, the computer resources and time for the analysis in the second phase allowed us to extend the homology search to all Rfam (Gardner et al., 2009) families.

### 5.4.1 Phase 1: Génolevures 2, Génolevures 3 and *E. gossypii*

As referred above some of the species of our study have been partially annotated for ncRNA genes. While the rRNA, tRNA and snRNAs families were fully annotated in all genomes the coverage of the remaining families

---

<sup>2</sup>By “non-coding region”, Frith et al. (Frith et al., 2005) mean all non-repetitive, transcribed genomic regions that do not code for proteins.

Family	SACE	ERGO	CAGL	DEHA	KLLA	KLTH	SAKL	YALI	ZYRO
RNase P	1	1	1	1	1	1	1	1	1
RNase MRP	1	-	-	-	-	-	-	-	-
SRP	1	1	1	1	1	1	1	1	1
Telomerase	1	1	1	-	1	-	1	-	-
snRNA	5	5	5	5	5	5	5	5	5
snoRNA C/D	45	42	1	1	42	42	42	1	41
snoRNA H/ACA	28	26	-	-	-	-	-	-	-
<i>S. cerevisiae</i> spec.	4	-	-	-	-	-	-	-	-
<b>Total</b>	<b>86</b>	<b>75</b>	<b>9</b>	<b>8</b>	<b>50</b>	<b>49</b>	<b>50</b>	<b>8</b>	<b>48</b>

Table 5.2: Number of ncRNAs annotated in *S. cerevisiae* and Génolevures genomes, grouped by families.

differed between species (see Table 5.2). We applied the pipeline on the genomes from Génolevures 2 and 3 phases and *E. gossypii* with the following results described below:

**1. BLAST homology search:** The 383 sequences used as queries for the BLAST search correspond to all annotated ncRNAs<sup>3</sup>. The search produced 1540 hits with E-values <0.1. After validation 554 of the hits were selected as ncRNA genes.

**2. INFERNAL homology search I:** A first INFERNAL search was performed using 83 covariance models corresponding to all Rfam families for which at least one budding yeast annotated homologue was found. The search produced 1250 hits with E-values <0.5. After validation 602 of the hits were selected as ncRNA genes.

**3. ROC analysis:** Given the hits resulting from the two previous steps I performed a simple ROC analysis (see A) of the obtained results in order to better understand the role of the E-value as a discriminator of hits.

As can be seen from Figure 5.2, INFERNAL’s E-values are better discriminant than BLAST ones. For a given True Positive Rate (TPR) the False Positive Rate (FPR) of BLAST is higher. A TPR of 99% is achieved for E-values of 0.07 for INFERNAL and 0.08 for BLAST (which is a quite similar result). The FPR however is of 0.35 for INFERNAL and 0.61 for BLAST which means that:

<sup>3</sup>In this section, the term “all annotated ncRNAs” refers to the annotated ncRNAs in the 10 species considered with exception of tRNA and rRNAs.

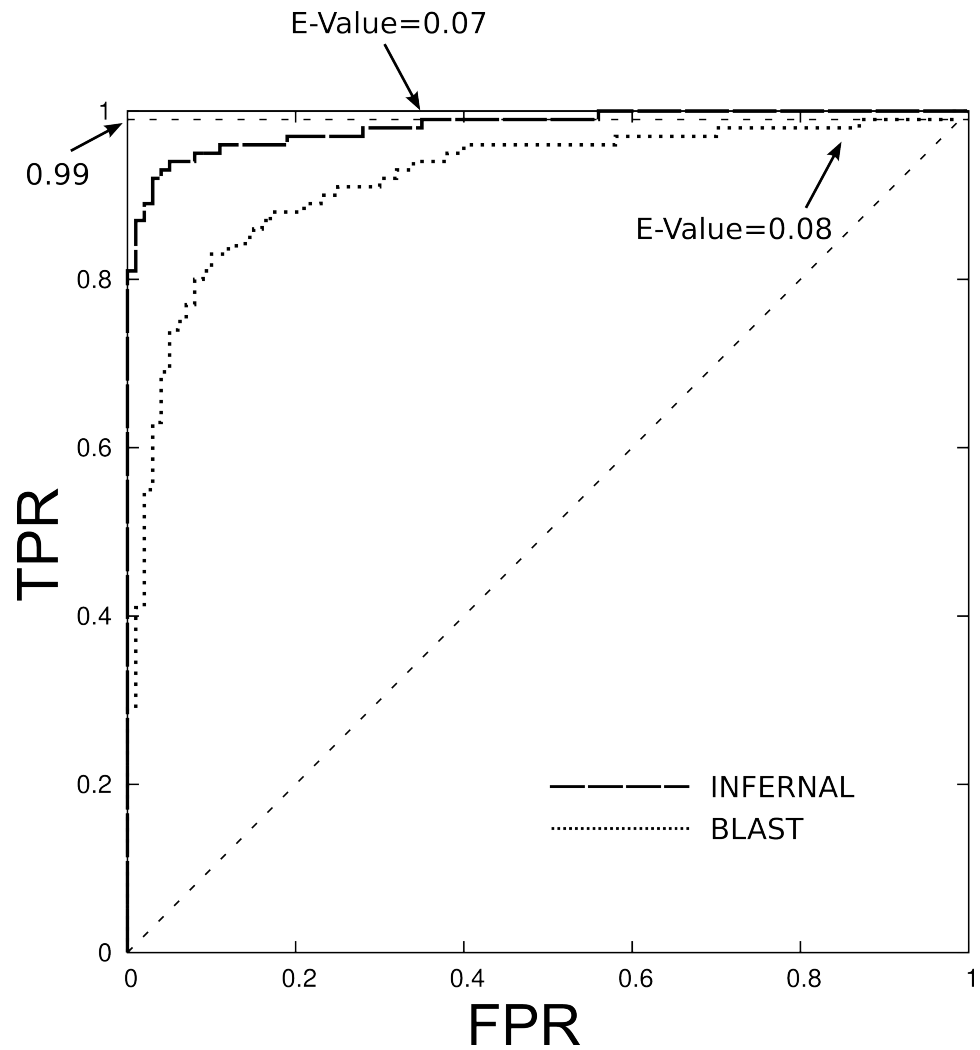


Figure 5.2: ROC analysis of the INFERNAL and BLAST E-values. Figure extracted from (Cruz and Westhof, 2011a).

1. If one is willing to loose up to 1% of all genes in order to save manual validation time a E-value cutoff of 0.1 for both **INFERNAL** and **BLAST** can be applied;
2. Assuming this E-value cutoff 35% of **INFERNAL** hits and 61% of **BLAST** hits are expected to be False Positives.

**4. INFERNAL homology search II:** A second **INFERNAL** search was performed using the remaining 719 covariance models from Rfam. This number corresponds to all Rfam families except the viral ncRNAs, the miRNAs (and, of course, the 83 from the first search). 1360 hits were obtained applying an E-value cutoff of 0.1 (a cutoff of 0.5 would produce 5578 hits). From these hits only 41 were selected as ncRNA genes. This low number should not surprise if one takes into account that the great majority of the searched families do not have homologues in yeast, thus most of the hits are False Positives with high E-value.

**5. *de novo* search I results:** The *de novo* search was performed as described in Chapter 4. Specific selection criteria were applied to *de novo* candidates:

- Candidates must occur in intergenic regions with no previous annotations or, at most, with a small overlap with annotated genes;
- The pairwise similarity between all candidate sequences must be higher than 50%;
- Candidates must display potential secondary structure shared between sequences. This secondary structure should be supported by co-variation (or, at least, by some compensatory mutations).

The 630 selected candidates were distributed as shown in Table 5.3. The difficult task of *de novo* gene finding is obvious if one considers that just a minor fraction of the existent ncRNAs were found. The *de novo* search, however, is far from a futile exercise for I was able to find a few ncRNAs that escaped the homology search and some interesting totally new candidates.

**Final results:** Applying the pipeline to the 10 genomes indicated above allowed us to find, in overall, 81% of the reference ncRNA genes, even though the coverage varies greatly from species to species (see Table 5.4).

Type of ncRNA	Found	%
Previously annotated tRNAs or rRNAs	273	43%
Previously annotated ncRNAs from other families	103	16%
ncRNAs with homologues in yeasts not found in the previous searches	25	4%
False positives (e.g. repetitive elements wrongly classified as ncRNAs by RNAz)	210	33%
Candidates selected as new putative ncRNA genes	19	3%

Table 5.3: *de novo* search results.

Genome	snRNAs	snoRNA C/D Box	snoRNA H/ACA Box	Other	Total	%
CAGL	5	42	24	3 †	74	87%
ZYRO	5	43	23	3 †	74	86%
SAKL	5	44	25	3 †	77	91%
KLTH	5	44	24	3 †	76	90%
KLLA	5	44	24	3 †	76	91%
ERGO	5	42	24	3 †	74	86%
DEHA	5	39	17	1 *	62	77%
PISO	5	35	17	2 §	59	73%
YALI	5	32	13	2 §	53	62%
ARAD	5	33	14	3 †	55	64%

Table 5.4: Phase 1 ncRNA annotations. Legend: †(RNaseP + RNase MRP + SRP); §(RNaseP + RNase MRP); \*(RNaseP).



Genome	snRNAs	snoRNA C/D Box	snoRNA H/ACA Box	Other	Total	%
CABR	5	42	23	3 †	74	89%
CACA	5	40	26	4 ‡	74	89%
CANI	5	40	24	3 †	73	88%
KLBA	5	38	23	4 ‡	69	83%
KLDE	5	41	21	3 †	71	86%

Table 5.5: Phase 2 ncRNA annotations (Nakaseomycetes genomes). Legend: ‡(RNaseP + SRP + RNase MRP + Telomerase RNA); †(RNaseP + SRP + RNase MRP). (Table from C. Fairhead et al., *in preparation*)

### 5.4.2 Phase 2: Nakaseomycetes

In phase 2 I applied the annotation pipeline, as described in Chapter 4, to the five nakaseomycetes genomes. The BLAST search, using 814 query sequences, produced 16892 hits of E-value <1.0 and the INFERNAL search, using the full 1362 Rfam families, produced 66282 hits of E-value <1.0. All hits were clustered in the respective candidates and only the 3113 clusters with E-value <0.1 were kept for analysis. Of those 1041 correspond to tRNAs and 1152 correspond to miRNAs and were not included in further analysis. The remaining 1020 candidates were aligned in their respective families and after manual validation I obtained 358 ncRNA annotations corresponding to 83% of the reference ncRNAs.

## 5.5 Some Noticeable Annotation Results

### 5.5.1 The telomerase ncRNA

The RNA component of the telomerase complex (TER) is a ncRNA particularly difficult to annotate in yeasts due to its poor sequence conservation – e.g., 7 aligned saccharomyces TERs present a mean pairwise sequence identity of 65%<sup>4</sup>. The longest hit obtained by BLAST, using the four known TER sequences as queries, had 20 nts long and an E-value of 0.01, which was not enough to produce a reliable alignment of the candidate – together with long insertions or deletions between conserved structural domains (Kachouri-Lafond et al., 2009). Our annotation pipeline was able to find the TER ncRNA only in three Nakaseomycetes genomes (*C. bracarenensis*, *C. nivariensis* and *K. delphensis*) whose Telomeric Repeat Sequence (TRS) were identical to the TRS of *C. glabrata*. In all other cases classical

sequence comparison approaches are helpless.

Although out of reach for the pipeline some observations could contribute to build an effective approach to find TERs (in yeast genomes at least).

It is known from structural and comparative studies that TER have some secondary structure features conserved across all known yeast homologues (Gunisova et al., 2009):

- A template region complementary to the TERS;
- A pseudknot with two uridine-rich and a adenine-rich strands;
- The yKu complex binding site (helical region);
- The EST1 complex binding site (helical region);
- Sm complex binding site (uridine-rich region);

I expect that the genome wide search for any of those features using available bioinformatics approaches (e.g., BLAST, position weight matrices, secondary motif searches, covariance model) will yield an enormous amount of noise. However, the conjugation of the individual candidates or each feature search, in the correct order and in a limited regions of the genome could filter part of the noise providing a reasonable small number of exploitable candidates. Unfortunately, testing all possible combinations of candidates at the genomic scale is computationally prohibitive.

A possible way around could be to reduce the search space from a full genome to a sub set of sequences. After analyzing the regions flanking the known TER genes I noticed a recurrent synteny between TER and the snR161 H/ACA Box snoRNA:

The synteny between TER and snR161 is striking. If one assumes it to be conserved across all yeasts one could define a relatively short region to apply the combinatorial search approach.

### 5.5.2 The synteny between RNase MRP and snR66

In budding yeast genomes five polycistronic ncRNA clusters (i.e., a unique transcript region containing several distinct ncRNA genes) have been identified (Souciet et al., 2009). Those clusters are characterized by the occurrence of 2 to 7 ncRNAs in a short region. The intervening genes, their relative order and the short distance between them are conserved across all yeasts with some minor exceptions (e.g., a few clusters lack one gene which can be

---

<sup>4</sup>Aligned sequences of *S. cerevisiae*, *S. paradoxus*, *S. kudriavzevii*, *S. cariocanus*, *S. bayanus*, *S. pastorianus* and *S. mkatae* obtained from (Podlevsky et al., 2008)

Genome	Contig	Strand	TER		intergenic	snR161	
SACE	II	-	307587	<		307185	>
ERGO	A	-	678331	>	DAD1	677570	>
CAGL	I	+	420932	<		421748	>
KLLA	B	-	611456	<	tRNA(Ile) – DAD1	609614	>
SAKL	D	-	349043	<	tRNA(Ile) – DAD1	345294	>
PISO	I	+	?		DAD1	577256	>
ARAD	D	-	?			3199765	>
KLTH	C	-	?		tRNA(Ile) – DAD1	699940	>
YALI	F	-	?			2693200	>
ZYRO	A	-	?		DAD1	295835	>
CABR	37	-	587451	>	n/a	585926	<
KLDE	3	+	363531	<	n/a	364450	>
KLBA	29	-	?		n/a	320229	>
CANI	19	+	694447	>	n/a	693152	<
CACA	36	+	?		n/a	486116	>

Table 5.6: Telomerase (TER) vs. snR161 syntenic region

due to incomplete annotation). From our annotation I observed a syntenic cluster containing the RNase MRP and the snR66 snoRNA, which occur together in all genomes where they were identified at a maximum distance of 230 nts. The conserved synteny and the short distance between those genes suggests a potential new polycistronic cluster. An interesting particularity of this potential cluster is that the genes involved are from different families, contrary to all other known clusters in which all genes are snoRNAs.

### 5.5.3 The TPP riboswitch

Riboswitches are regulatory domains present in the 5' UTR, 3' UTR and introns of mRNAs that regulate gene expression by binding to a specific ligand and changing the normal processing of the mRNA (Nahvi et al., 2002; Mironov et al., 2002). Particularly common in bacteria where they regulate gene expression by transcription terminating or translation inhibition, riboswitches are rarer in fungi and plants where they are present in introns and modulate the alternative splicing of the pre mRNA (Bocobza et al., 2007; Kubodera et al., 2003; Cheah et al., 2007).

Riboswitches are beautiful examples of RNA structure modularity: The three-dimensional structure of the aptamer domain is responsible for binding the metabolite. Each specific aptamer structure recognizes a particular ligand and many ligands are known to be recognized by riboswitches, such as adenosylcobalamin (AdoCbl), flavin mononucleotide (FMN); glucosamine-6-

Genome	Contig	Strand	RNase MRP - snR66
SACE	II	+	64
KLLA	E	+	230
SAKL	E	+	122
PISO	F	+	94
ARAD	D	-	100
KLTH	G	-	143
YALI	D	-	85
ZYRO	A	-	117
CABR	29	+	157
KLDE	27	-	175
KLBA	23	+	109
CANI	17	-	162
CACA	5	+	148

Table 5.7: RNase MRP vs. snR66 syntenic region

phosphate (GlcN6P), S-adenosylmethionine(SAM), thiamin pyrophosphate (TPP) and many others (Barrick and Breaker, 2007). The presence/absence of the ligand causes a change in conformation allowing an interaction between the aptamer and an expression platform which activates/inactivates the gene expression. In bacteria, virtually all combinations of aptamers and expression platforms can be observed. On the contrary, in eukaryotes, only TPP riboswitches are observed either in plants (Bocobza et al., 2007) as in filamentous fungi (Kubodera et al., 2003). In budding yeasts, as far as I know, no riboswitch has been reported up to this day.

Five genes containing a TPP riboswitch were found by the pipeline in *A. adeninovorans* (3 genes), *D. hansenii* (1 gene) and *Y. lipolytica* (2 genes). The alignment against the crystal structures of the TPP riboswitches from *E. coli* (Edwards and Ferré-D'Amaré, 2006) and *A. thaliana* (Thore et al., 2006) revealed a striking similarity of the key structural nucleotides (see Figure 5.3 A).

All of those newly found riboswitches occur in the 5' UTR of their respective genes. Determining how they function, however, is beyond the bioinformatic approach and will require direct experimentation.

The distribution of this regulatory domain across the Saccharomycotina phylogeny (see Figure 5.3 B), that is totally absent in the *S. cerevisiae* branch, raises interesting questions about the evolution of this type of regulatory mechanisms, and shows the utility of extending homology search beyond the closest group of species.



#### 5.5.4 A snoRNA candidate in *Yarrowia lipolytica*

A curious snoRNA candidate was found in the genome of *Y. lipolytica* using the covariance model of an archaeal snoRNA from the *Pyrococcus* family (Rfam code RF00095). The lack of similarity between the candidate and the *Pyrococcus* guide sequences (and even among the different *Pyrococcus* sequences themselves) clearly excludes the hypothesis of homology. In fact I believe that the RF00095 family, in contrast with other snoRNA families, includes several different *Pyrococcus* snoRNAs that were grouped together because of the length and the C/D boxes conservation.

Several arguments support the *Y. lipolytica* as a snoRNA candidate (see Figure 5.4):

- The length, Box C/D sequences and the distance between them are compatible with a snoRNA;
- Both putative guide sequences are complementary with the ribosome sequences on 10 base pairs.

I aligned both *Y. lipolytica* and *S. cerevisiae* ribosomal sequences in order to determine which *S. cerevisiae* nucleotides are the homologues of the *Y. lipolytica* nucleotides potentially modified by this snoRNA. No modifications are reported in the “The Yeast snoRNA Database” (Piekna-Przybylska et al., 2007) in either *S. cerevisiae* nucleotide. This absence raises doubts that only experimental validation would clarify.

#### 5.5.5 *de novo* Candidates

Most of the ncRNA candidates found by the *de novo* search are short alignments of only two sequences which prevents further considerations on conservation and potential function. The genomic location of the candidates, however, suggests potential regulatory roles as 14 of the 19 genes occur less than 200 from the 5' or 3' UTR of known genes. Some of those genes are homologues of ribosomal related *S. cerevisiae* genes. Figure 5.5 shows three examples of the found ncRNAs (for a complete list of all *de novo* candidates found in this search see (Cruz and Westhof, 2011a)).

### 5.6 The extremely large ncRNAs in the Nakaseomycetes family

Structural ncRNAs present the ability to accept large sequence insertions / deletions at specific structural domains without compromising the overall three dimensional structure of the molecule. The U1 and U2 snRNAs of *S. cerevisiae* are striking examples of it. The yeast U1 snRNA presents three yeast specific domains adding up to 568 nts (compared to 164 nts

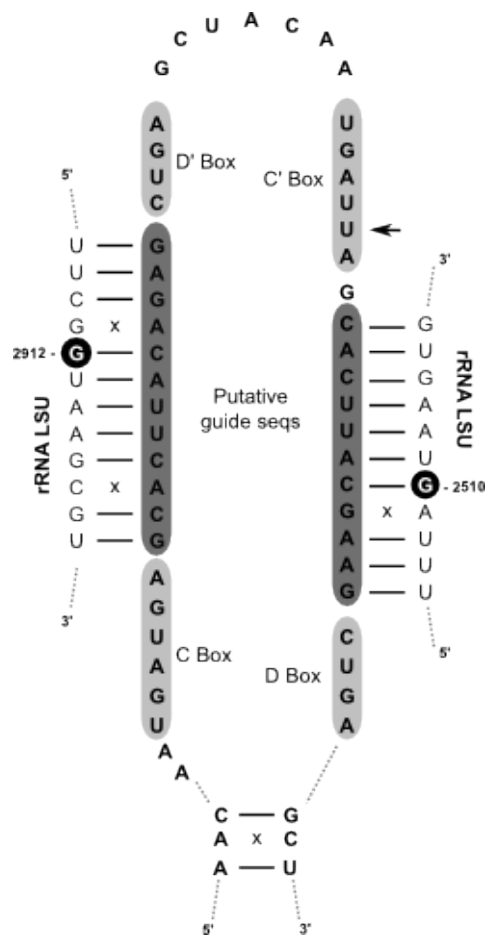
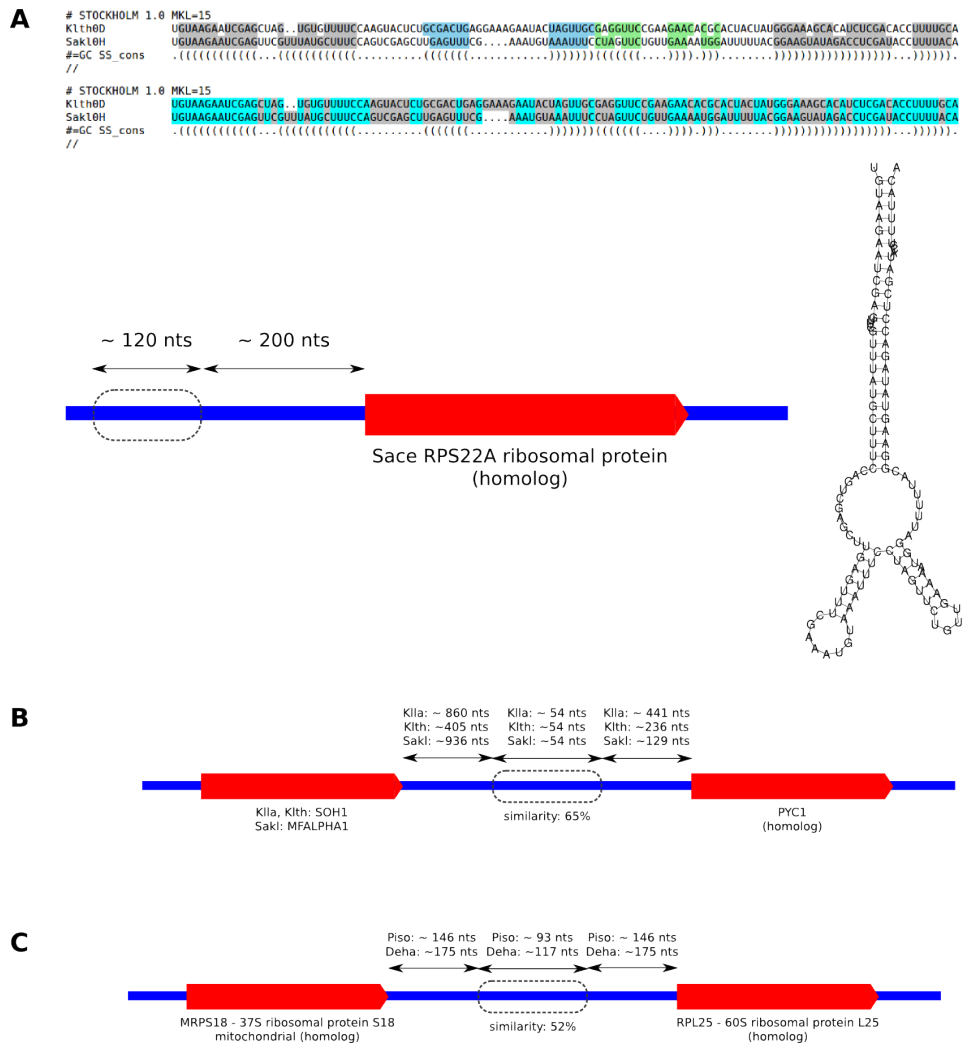


Figure 5.4: *Y. lipolytica* snoRNA candidate. Sequence alignment and secondary structure. Figure extracted from (Cruz and Westhof, 2011a).

Figure 5.5: Examples of three *de novo* candidates



in *H. sapiens*) (Kretzner et al., 1987; Kretzner et al., 1990). The yeast U2 snRNA (1175 nts), on its turn, is six times longer than the *H. sapiens* one (187 nts) (Igel and Ares, 1988). Surprisingly, those additional yeast domains were shown to be non essential (Kretzner et al., 1987; Liao et al., 1990). More recently, the RNaseP *C. glabrata* was also shown to contain particularly large extra domains (Kachouri et al., 2005).

The nakaseomycetes yeasts add new examples to this list of particularly large molecules with an extremely large RNaseP, such as the one from *K. delphensis* that exceeds *C. glabrata* RNase P by more than 200 nucleotides. (see Figure 5.6) and U1 snRNA from *C. nivariensis* that extends the three additional domains present in Budding Yeasts in more than 1100 nucleotides and presents a curious new additional domain on an extremely conserved helix. This new extension domain replaces a single bulged nucleotide that is conserved across eukaryotes (see Figure 5.7).

Figure 5.8 depicts the distribution and the length of the additional domains across the budding yeast genomes.

## 5.7 Conclusions

Here I described the work done on the development of an ncRNA annotation pipeline and its application to 15 yeast genomes. With this pipeline I was able to discover and annotate 83% of the ncRNA genes expected from comparison with the *S. cerevisiae* genome and a number of few other genes with no new homologue in *S. cerevisiae*. As expected, evolutionary distance between species is a key factor to the annotation success, and the availability of new, close species will surely contribute to better and denser annotation in the future.

These results show that the available tools make the automatic annotation of ncRNAs in yeast genomes a practical approach to be included in the global annotation pipelines of the next sequencing projects like the Dikaryome project.

Our original goal in the beginning of this work was to discover, using solely bioinformatics techniques, the maximum number of ncRNAs present in the Genolevures genomes. This goal has a clearly subjective aspect as one cannot evaluate “the maximum number” without an idea about the real number. As a practical reference I chose the best annotated yeast genome to compare our results with. The question, however, remains (see Section 5.3): “How many ncRNAs are there?” Recent observations from *S. cerevisiae* studies, such as the existence of transcribed intergenic regions with no annotated functions [Nagalakshmi 2008] and the detection of long ncRNAs of unknown function [Kavanaugh 2009], show how pertinent this question is. The correct identification of possible new ncRNAs will surely require the synergy between pure sequence analysis methods, high-throughput techniques

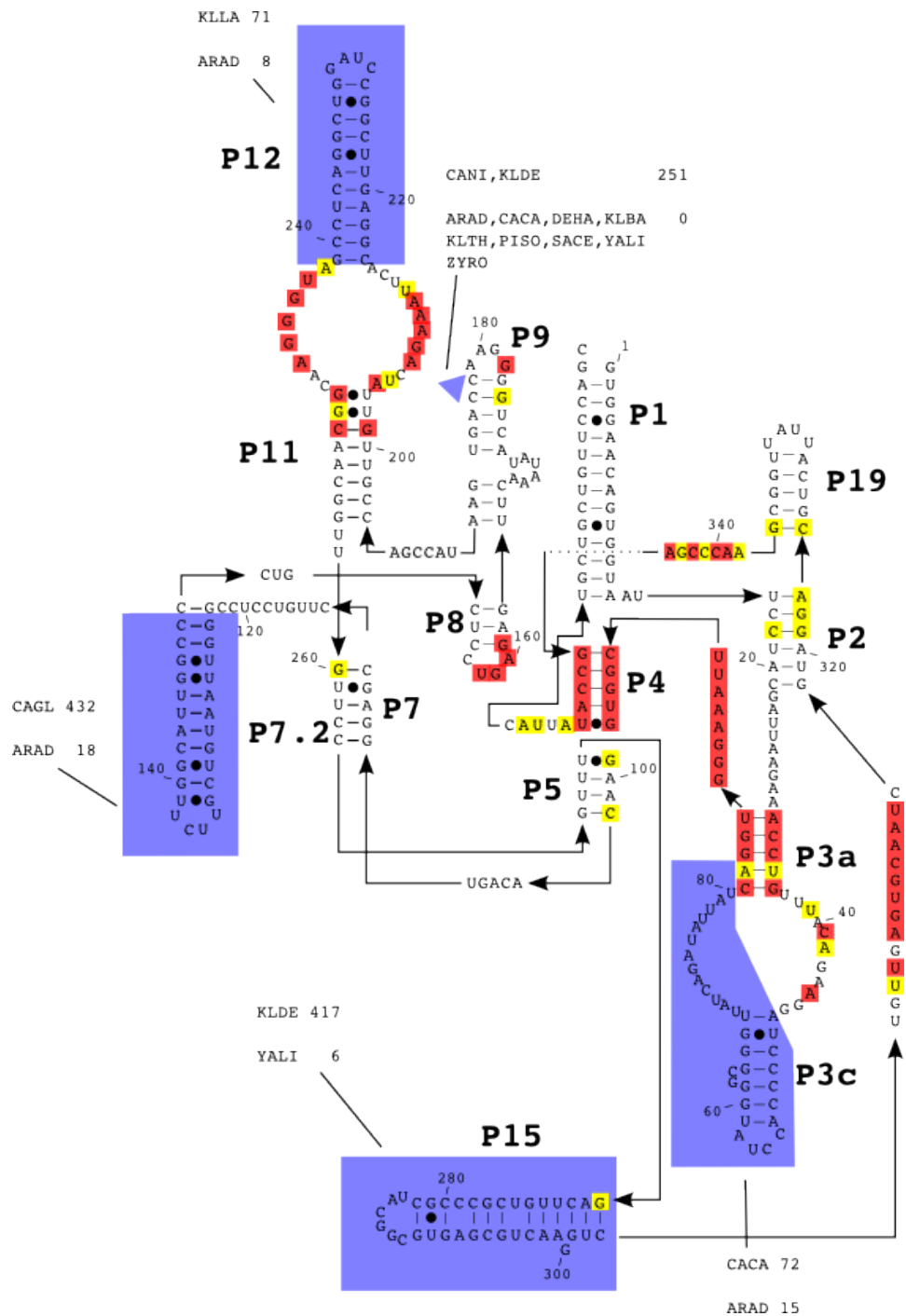


Figure 5.6: RNase P extensions. The sequence corresponds to the *S. cerevisiae* RNase P. Conserved positions are signaled with red (> 90%) and yellow (> 75%) squares. Yeast specific extensions are represented as shaded blue areas. Numbers close to the extensions indicate the species with the largest and smallest extension and the respective sizes (Figure from C. Fairhead et al., *in preparation*).

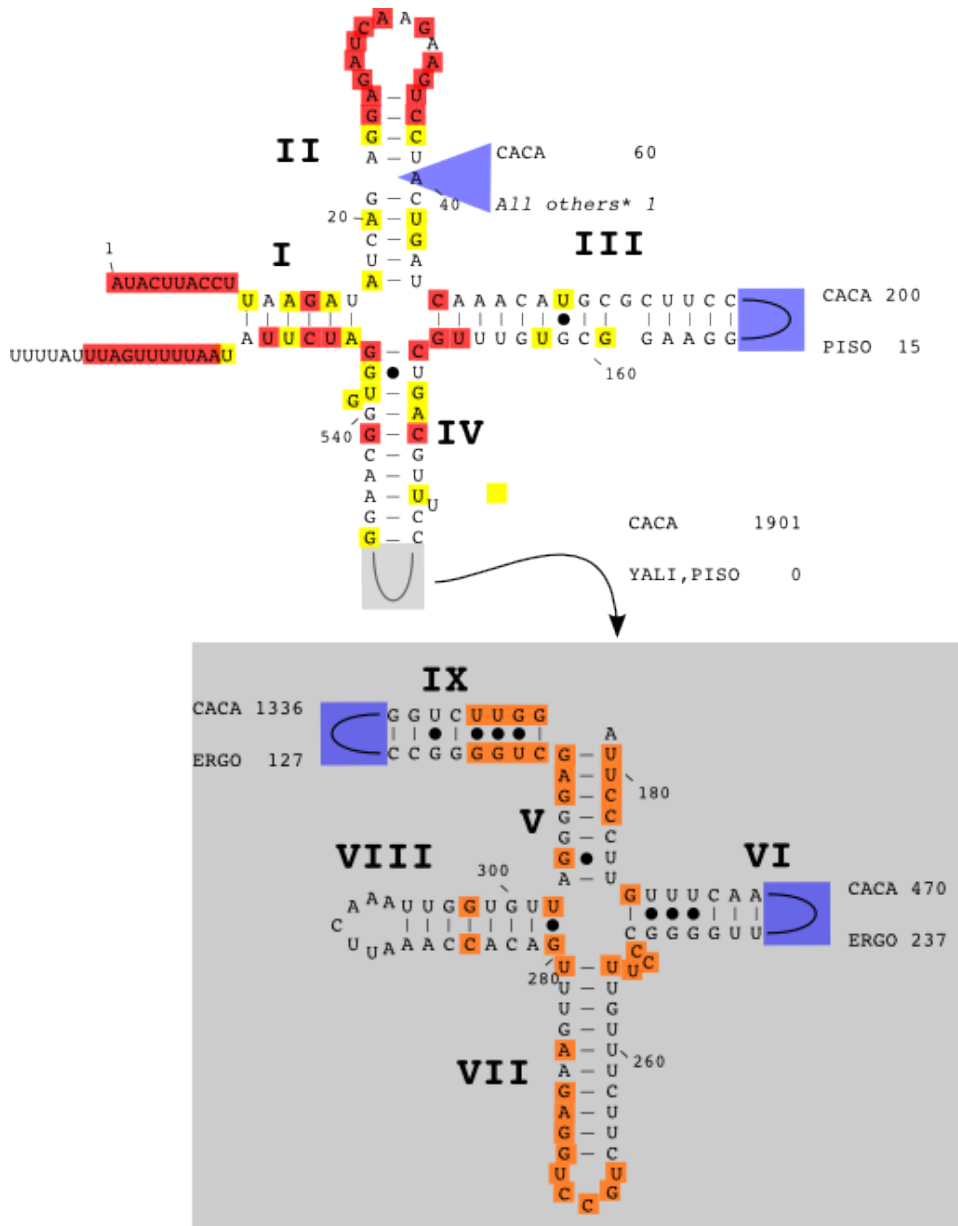


Figure 5.7: U1 snRNA extensions. The sequence corresponds to the *S. cerevisiae* U1 snRNA. Conserved positions are signaled with red (> 90%) and yellow (> 75%) squares. Yeast specific extensions are represented as shaded blue areas. Numbers close to the extensions indicate the species with the largest and smallest extension and the respective sizes. The gray shaded area corresponds to domains not present in *D. hansenii*, *P. sorbitophila*, *A. adeninovorans* and *Y. lipolitica*. Orange squares correspond to positions conserved in > 90% of the species that present the extension domains (Figure from C. Fairhead et al., *in preparation*).

	RNase P					Total	snRNA U1				Total	Genome Size (MB)
	P3c	P7.2	P9	P12	P15		II	III	VI	IX		
Saccharomyces_cerevisiae	33	39	0	26	34	358	1	93	320	206	565	11.7
Candida_glabrata	42	432	78	65	248	1141	1	118	327	440	836	11.9
Kluyveromyces_delphensis	58	392	251	24	417	1359	1	119	371	316	711	10.6
Candida_nivariensis	47	279	251	24	220	1048	1	120	368	442	838	11.3
Candida_bracarenensis	42	323	219	24	237	1059	1	162	378	463	900	12.0
Kluyveromyces_bacillisporus	38	312	0	25	328	925	1	91	267	129	525	10.8
Candida_castellii	72	153	0	16	142	609	60	200	470	1336	1935	10.0
Zygosaccharomyces_rouxii	40	45	0	15	35	336	1	82	340	237	554	9.5
Kluyveromyces_thermotolerans	18	33	0	33	26	325	1	72	278	157	483	10.1
Saccharomyces_kluyveri	35	31	3	16	28	327	1	59	252	128	445	11.0
Eremothecium_gossypii	19	41	3	66	35	371	1	107	237	127	484	10.6
Kluyveromyces_lactis	20	44	1	71	28	374	0	104	284	175	525	10.4
Pichia_sorbitophila	31	46	0	37	21	321	0	15	0	0	135	17.6
Debaryomyces_hanseni	29	25	0	16	22	304	0	20	0	0	137	11.8
Arxula_adeninovorans	15	18	0	8	14	254	1	20	0	0	134	11.3
Yarrowia_lipolytica	22	25	0	30	6	288	1	20	0	0	124	19.9
Schizosaccharomyces_pombe												

Figure 5.8: Lengths of the yeast specific extensions by species, for each extension domain of RNase P and U1 snRNA (Table from C. Fairhead et al., *in preparation*).

of sequencing [Wang 2009] as well as the application of structural knowledge to ncRNA search.

## 5.8 Article – Identification and Annotation of Non-Coding RNAs in Saccharomycotina

The present chapter and the previous one are extended summaries of the following article:

Cruz, J. A. and Westhof, E. (2011). *Identification and annotation of non-coding RNAs in Saccharomycotina*. *Comptes rendus – Biologies* 334, 671-678.



Contents lists available at ScienceDirect

## Comptes Rendus Biologies

www.sciencedirect.com



Molecular biology and genetics / Biologie et génétique moléculaires

Identification and annotation of noncoding RNAs in *Saccharomycotina*

José Almeida Cruz\*, Eric Westhof

*Architecture et réactivité de l'ARN, institut de biologie moléculaire et cellulaire du CNRS, université de Strasbourg, 15, rue René-Descartes, 67084 Strasbourg cedex, France*

## ARTICLE INFO

## Article history:

Received 8 November 2010

Accepted after revision 23 March 2011

Available online 6 July 2011

## Keywords:

Noncoding RNAs  
Genome sequencing  
*Saccharomycotina*

## ABSTRACT

The importance of ncRNAs in biological processes makes their annotation an essential component of any genome-sequencing project. The identification of ncRNAs in genomes requires specific expertise and tools that are distinct from the traditional protein gene annotation tools. Here, we describe the assembly of two automatic annotation pipelines, integrating publicly available tools, for homology and *de novo* ncRNA search in genomes. We applied both pipelines to 10 *Saccharomycotina* genomes and were able to find and annotate 693 ncRNA genes, corresponding to 81% of the ncRNAs expected for those genomes assuming the number of ncRNAs in *Saccharomyces cerevisiae* (86) as a reference. Several new ncRNAs, not yet known in the *Saccharomycotina* clade, were also detected. The results show the feasibility of automatic search for ncRNAs in full genomes and the utility of such approaches in large multi-genome sequencing and annotation projects.

© 2011 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.

## 1. Introduction

Noncoding RNAs (ncRNAs) form an important class of macromolecules participating in key cellular mechanisms such as protein synthesis, gene splicing, telomere elongation, regulation of gene expression (e.g., riboswitches, miRNAs, and other small regulatory RNAs), and gene silencing. In general, ncRNAs are specific transcripts that often participate in complex regulatory mechanisms [1–3].

Finding ncRNAs in genomes is far from trivial. First, the ncRNA genes lack the characteristic features of protein genes such as start and stop codons, splicing sites, codon frequency bias [4]. Second, structured ncRNAs are, in general, more conserved in structure than in sequence due to base covariations in helices and neutral substitutions in tertiary interactions [5,6]. Third, insertions of long sequences occur frequently [7–9]. These characteristics reduce the effectiveness of searches based on pure sequence comparison. Finally, the curation of ncRNA gene predictions

is time consuming and demand expertise. In spite of these difficulties, the increasing recognition of the importance of ncRNAs in biological processes makes their annotation an essential component of any genome sequencing project.

To guarantee a timely and effective annotation of ncRNAs, multi-genome sequencing projects, such as the Génolevures project [10], require, as far as possible, automatic gene annotation and a set of simplifying procedures and tools for fast and accurate manual curation.

The problem of genomic ncRNA annotation has been tackled by several authors [11–21]. However, the specificities of each project, such as the differences in ncRNA biology between species, genome size, sequence and structure divergence, demand a careful analysis before applying any known method.

Here, we present our approach of ncRNA annotation in the context of the Génolevures consortium. We assembled a pipeline integrating publicly available tools for ncRNA search in sequences and applied it to the annotation of 9 budding yeast genomes from the Génolevures Database [22] and the *Ashbya* Genome Database [23]. We were able to annotate automatically, with a relatively small human validation effort, 693 ncRNAs that correspond to 81% of

\* Corresponding author.

E-mail address: J.Cruz@ibmc.u-strasbg.fr (J.A. Cruz).

what we would expect taking the 86 annotated ncRNAs of the *Saccharomyces cerevisiae* genome as a reference.

## 2. Results

### 2.1. Homology pipeline

Homology search consists in searching for new members of known ncRNAs families. The important, and increasing, number of publicly available ncRNA annotations makes homology search the first step in any ncRNA annotation process.

Search tools based on sequence alignment, such as BLAST [24], allow for very fast searches of sequences (query sequences) in genomes (target sequences). The sequence similarity between query and target sequences plays a major role in the success rate of BLAST searches. Significant candidates (i.e. those with E-values < 0.1) will present sequence similarities above 84% with a minimum length of 16 nts. Thus, the use of BLAST for homology search becomes more effective for ncRNA families with large conserved sequences, such as rRNA or snRNAs, or when searching within closely related species. When the target and query sequences come from distant species, or present low sequence conservation, with potentially large insertions, pure sequence alignment methods loose efficiency. A way to get around this limitation is to explore the statistical signals imprinted on the sequence by the RNA structural constraints.

ncRNAs present complex three-dimensional structures of packed double-stranded helical regions connected by tertiary interactions that are generally mediated by single-stranded regions [25]. The set of double-stranded and

single-stranded regions is called secondary structure. RNA helices consist of stacks of A–U; G–C and G–U base pairs. Those base pairs are structurally equivalent, and the substitution of one base pair by another one has, frequently, minimal or no impact on the molecular structure. The accumulation of base pair substitutions in a RNA sequence can render two homologous sequences very dissimilar. However, when observed from the point of view of a multiple sequence structural alignment, it generates a pattern of covariation between the base-paired positions that can be detected in the respective columns of the sequence alignment [26,27]. This dependency between paired positions in alignments is used to build covariance models, used by ncRNA search tools such as INFERNAL [28], to search for ncRNAs in large sequences. Additionally, if sufficiently diverse sequences are included in the alignments, the known positions of insertions can also be included into the models. Covariance models, for most of the known ncRNAs families, are curated and maintained in the RFam database [29] and could be readily applied in our search.

The results produced by any search tool must be automatically filtered in order to exclude candidates less likely to be real ncRNAs. Candidates conflicting with known annotations or with low score should be discarded from the candidate list. Additionally, all retained candidates should be structurally aligned with known homologues in order to facilitate human validation, required as the last step of the annotation process (Fig. 1A).

The Génolevures database [22] contains 9 budding yeast genomes in which only tRNAs and rRNAs were fully annotated and were not considered in this work. Some other ncRNA families were partially annotated (see Table 1,

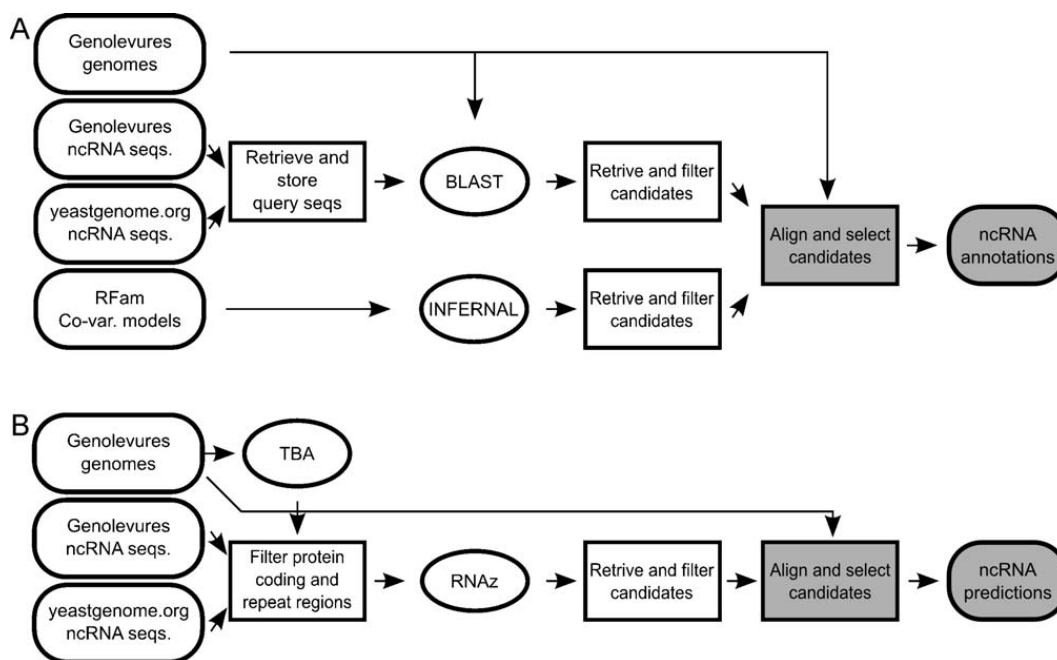


Fig. 1. Workflow of the two pipelines for the ncRNA search and annotation. The pipelines integrate external sources of data (round white squares), external tools (ellipsis), automatic (white squares) and manual (shaded squares) processing steps to produce final ncRNA predictions and annotations (round shaded squares). A. Homology search pipeline. B. *De novo* search pipeline.



**Table 1**

Number of ncRNAs found with the homology pipeline. Rows contain species numbers, columns contain numbers for each ncRNA family. First row (*sace*) displays the reference number of annotated ncRNA in the *Saccharomyces cerevisiae* genome. Original annotation column refers to ncRNAs annotated previously to the present work. “*Sace* specific” column refers to 4 ncRNA annotated in the *S. cerevisiae* genome.

Species	Original Annotation	Rnase P	SRP	Rnase MRP	Telomerase	snRNA	snoRNA C/D	snoRNA H/ACA	<i>Sace</i> specific	TOTAL
<i>sace</i>	86	1	1	1	1	5	44	29	4	86
		Number of found ncRNAs								
<i>cagl</i>	9	1	1	1	0	5	42	24	0	74
<i>zyro</i>	48	1	1	1	0	5	43	23	0	74
<i>sakl</i>	50	1	1	1	0	5	44	25	0	77
<i>klth</i>	49	1	1	1	0	5	44	24	0	76
<i>klla</i>	50	1	1	1	0	5	44	24	0	76
<i>ergo</i>	75	1	1	1	0	5	42	24	0	74
<i>deha</i>	8	1	0	0	0	5	39	17	0	62
<i>piso</i>	0	1	0	1	0	5	35	17	0	59
<i>yali</i>	8	1	0	1	0	5	32	13	1	53
<i>arad</i>	0	1	1	1	0	5	33	14	0	55
<i>Total</i>	297	10	7	9	0	50	398	205	1	680

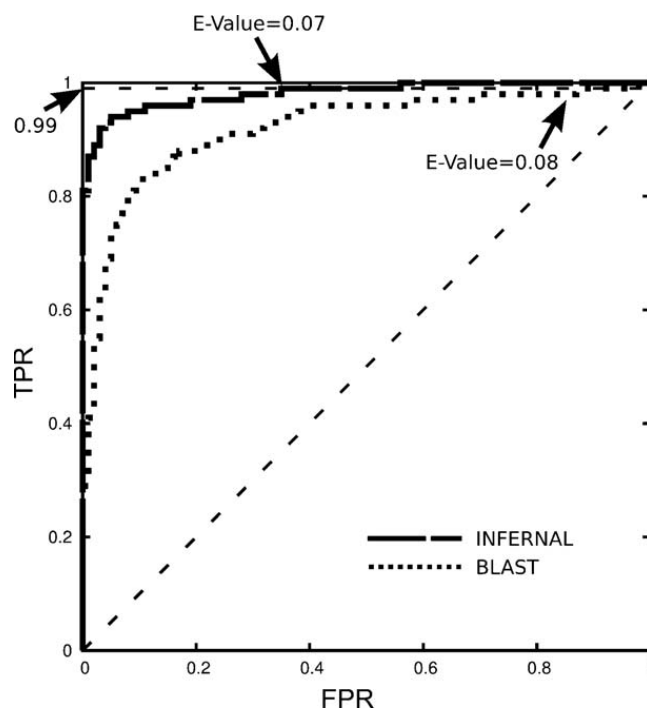
column “Original Annotation”). The Ashbya Genome Database [23] contains the fully annotated genome of the *Ashbya Gossypii* (also known as *Eremothecium gossypii*) yeast.

We performed a BLAST search on those 10 genomes using as queries the *S. cerevisiae* ncRNAs and all the originally annotated ncRNAs from the original databases. This set of query sequences corresponds to the closest species for which we had reliable and ready to use ncRNA annotations. We believe that using a larger and more distant set of ncRNA sequences would increase the low score candidates with little improvement on the amount of genes found (e.g. of the 86 *S. cerevisiae* query sequences 37% produced true positive candidates when BLASTed against *Candida glabrata*, but only 6% when BLASTed against the more distant *Debaryomyces hansenii* and *Yarrowia lipolytica*—see Table SI 1). The BLAST search produced 1540 candidates with E-values < 0.1. A first INFERNAL search, using the 83 covariance models corresponding to the ncRNA families present in *S. cerevisiae* genome, produced 1250 candidates with E-values < 0.5. The candidates were included in the structural alignment of the corresponding family and manually validated according to the criteria described in the section “Candidate Validation”. After validation we retained 554 BLAST candidates and 602 INFERNAL candidates as ncRNAs.

We were interested on how E-values behave as classifiers in this specific data set. Fig. 2 shows the ROC curve [30], after the manual classification of True and False positives (TP and FP), for both tools. INFERNAL E-values were slightly better discriminators than the E-values of BLAST. While both tools will achieve 99% of True Positive Rates (TPR) at similar E-values (0.07 and 0.08 for INFERNAL and BLAST, respectively), the False Discovery Rates (i.e., the proportion of FPs in selected candidates) is significantly lower for INFERNAL (0.35) than for BLAST (0.61).

A second INFERNAL search was performed including all the remaining Rfam families with exception of viral and miRNA families (719 Rfam families). According to the sensitivity analysis performed previously (Fig. 2), we reduced the expected value cutoff to < 0.1 obtaining 1360 candidates (applying the less restrictive cutoff, E-values < 0.5, we would have obtained 5578 candidates). From this search, only 41 candidates were selected. This

surprisingly low number can be explained by the fact that most of the 719 Rfam families do not have homologues in yeasts, thus the resulting candidates are mostly FPs with high E-values (see Supplementary material SI 1). Curiously, all except two of the selected candidates correspond to homologous ncRNAs not found in the first search. A representative example is the box C/D snoRNA snR47 that was identified in *Y. lipolytica*, *D. hansenii*, *K. thermotolerans* and *Z. rouxii* with the covariance model of the snoRNA SNORD36 that is the mammalian ortholog of snR47. The remaining two candidates were, until now, unknown in the *Saccharomycotina* clade: a Box C/D snoRNA found in *Y. lipolytica* (Fig. 3) and a TPP riboswitch found in the 5'



**Fig. 2.** Receiver Operating Characteristic (ROC) curves for INFERNAL (dashed curve) and BLAST (dotted curve) E-values. INFERNAL E-values are slightly better discriminants (Area Under the Curve [AUC] = 0.98) than BLAST E-values (AUC = 0.92). A True Positive Rate (TPR) of 0.99 implies a False Positive Rate (FPR) of 0.34 for INFERNAL (E-Value = 0.07) and 0.87 for BLAST (E-Value = 0.08).

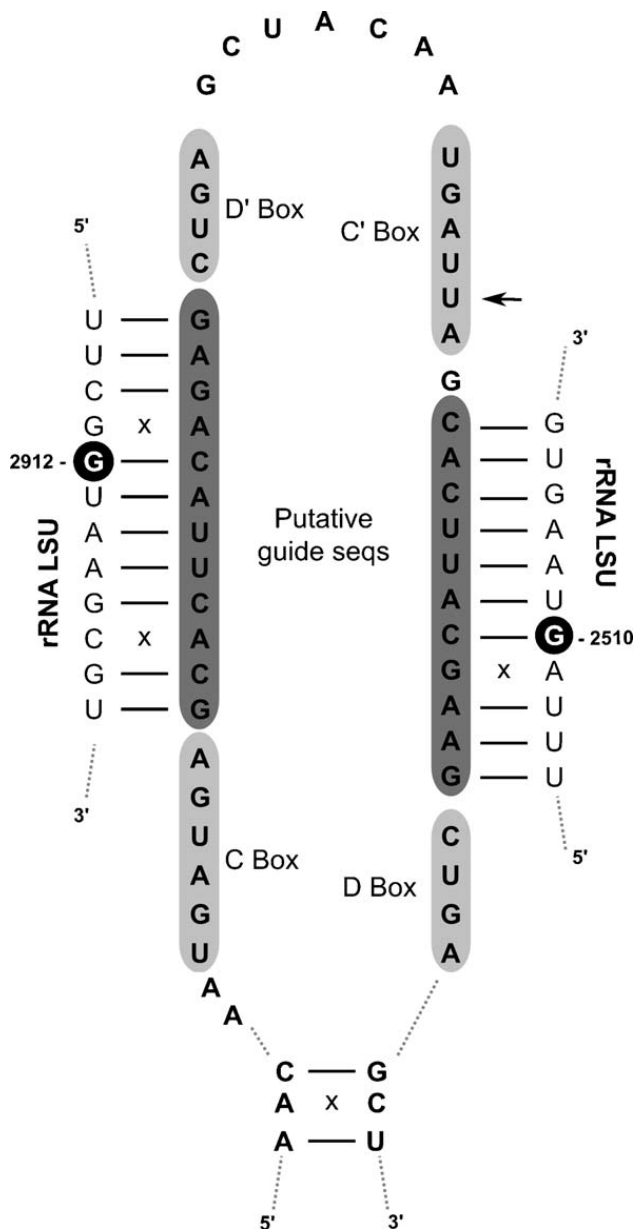


Fig. 3. Secondary structure of the C/D box snoRNA candidate found in *Y. lipolytica*. The size, sequence and position of the putative C/D and C'/D' boxes are compatible with a typical C/D snoRNA (black arrow points to a deviation from a canonical C' Box). The putative guide sequences are complementary to two regions of the Ribosomal Large Subunit of *Y. lipolytica* (with pairs missing Watson-Crick complementarity indicated by small 'x'). The predicted modified positions (white nucleotides in black circles) are not known to be modified in yeast.

UTR of protein coding genes in three different species (Fig. 4). The snoRNA candidate was found by INFERNAL using a covariance model of an archaeal snoRNA from the *Pyrococcus* family. The conservation of the characteristics Box C/D, and the complementarity of the putative guide sequences with the ribosome prevent a rapid exclusion of this candidate. On the other hand, the fact that the homologues of the putative ribosomal target sites are not modified in *S. cerevisiae* raises doubts about the nature of this candidate that only experimental validation could confirm. The TPP riboswitches are the only riboswitches found in eukaryotes [31,32] and the structural alignment

with the crystal structures of the bacterial (*Escherichia coli*) [33] and plant (*A. thaliana*) [34] TPP riboswitches reveals a striking similarity of key structural nucleotides. The distribution of this regulatory domain across the *Saccharomycotina* phylogeny (Fig. 4B), totally absent in the *S. cerevisiae* branch, while present in *D. hansenii*, *A. adeninovorans*. and *Y. lipolytica*, raises interesting questions about the evolution of this type of mechanisms, and show the utility of extending homology search beyond the close group of species.

Table 1 presents the results of the homology pipeline distributed by species and ncRNAs families. The total retained candidates correspond to 79% of all ncRNAs that were expected assuming the *S. cerevisiae* database as the reference for the ncRNA families present on yeast. At least 60% of the expected ncRNAs were found in all species. Unsurprisingly the homology search was much more effective in the species from the upper branch (from *S. cerevisiae* to *A. gossypii*), as 90% of the sequences used as queries came from that branch. Comparing the proportion of ncRNAs found by INFERNAL and BLAST for each species (Fig. 5B), we can observe that BLAST was as sensitive as INFERNAL as long as the searched genomes were close enough. In more distant species, while both tools loose performance, INFERNAL is much more sensitive.

Finally, it was not possible to find the RNA component of the Telomerase complex with any of the used tools. This ncRNA presents a challenge to automatic search programs due to minimal sequence and secondary structure conservation, extensive insertions/deletions and variable size between species [9,35,36]. The telomerase ncRNAs contains some structural features that are common to most known yeasts [36] such as the template region complementary to the Telomeric Repeat Sequence (TRS), a characteristic uridine-rich pseudoknot, two helical regions known to be the binding sites of the yKu and EST1 complexes, and a uridine-rich Sm binding domain. The template region can be detected with a simple BLAST using the TRS of the organism as a BLAST query if this TRS is long enough to produce meaningful hits [9]. If the TRS is too short or unknown, there is no simple way to find the telomerase ncRNA with bioinformatics analysis alone. In this case, the combined search for all structural features occurring in the correct order in the same region of the genome could, eventually, be an alternative approach.

## 2.2. De novo pipeline

Contrary to the homology search pipeline, a *de novo* search involves looking for what we do not know, i.e., ncRNAs for which no homologous are available *a priori*. One of the limitations of the *de novo* searches is that even the most promising candidates will require experimental evidence to be validated as *bona fide* ncRNAs. In general, *de novo* ncRNA search tools [37,38] rely on the same set of assumptions: (i) the homologous sequences of a ncRNA share the same overall secondary structure; (ii) alignments of ncRNAs reveal the covariation patterns resulting from compensatory mutations and, consequently, allow the inference of the secondary structure; (iii) when applying standard folding algorithms to the alignment,

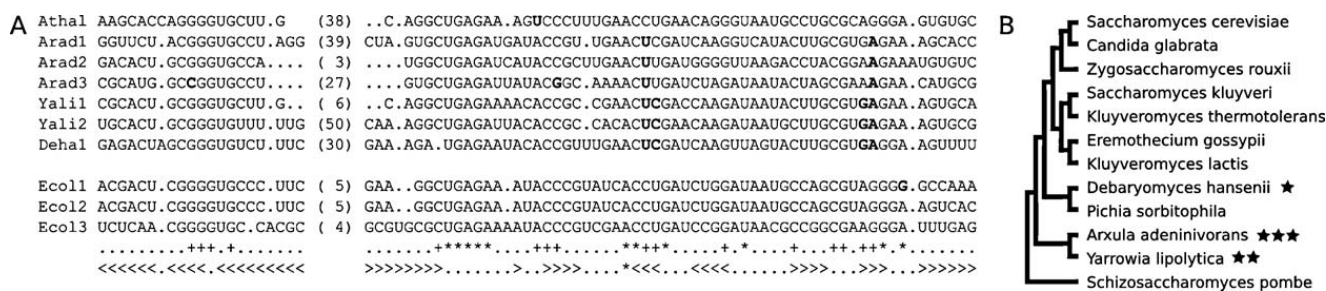


Fig. 4. TPP riboswitch candidates found by homology. A. Multiple structural sequence alignment of the seven predicted candidates. The '+' and '\*' indicate columns conserved in at least 90 and 99% of known sequences respectively. In bold are the nucleotides deviating from the consensus. The last row represents the secondary structure in bracket notation. Candidate sequences are compatible with the observed sequence conservation and secondary structure. B. Phylogenetic tree of the searched species. Stars represent the number of TPP ncRNAs found in each organism.

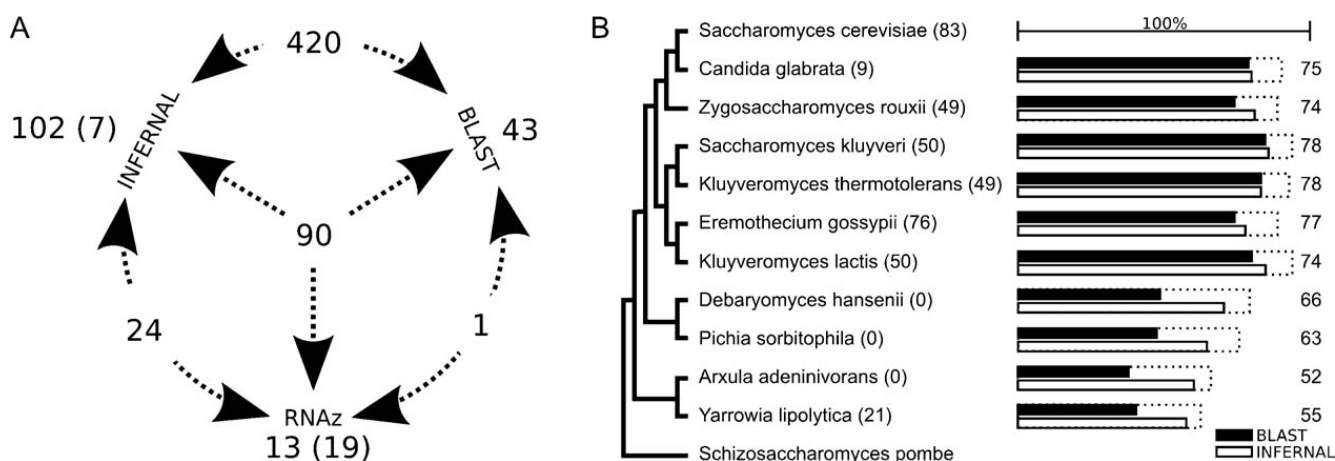


Fig. 5. Number of ncRNAs found using each of the tools. A. The numbers beside each tool name correspond to ncRNAs found exclusively by that tool. Numbers between two arrows are ncRNAs found by both tools. The number of ncRNAs found by the three tools is at the center. In parenthesis are the found ncRNAs with no homologue in *Saccharomyces cerevisiae*. B. Proportion of ncRNA found in each genome by INFERNAL and BLAST. While both tools decrease performance with increasing evolutionary distance, INFERNAL is less affected than BLAST. Numbers in parenthesis indicate the number of ncRNAs originally annotated and used in the homology search. Dotted squares and numbers correspond to the total proportion and absolute number of found ncRNAs after running both pipelines.

the resulting minimum free energy (MFE) will be lower than the average MFE obtained by folding random sequences with the same nucleotide composition. Although none of these assumptions is always true—in particular the third one [39,40], together they can be considered as good indicators for the acceptance of a predicted candidate as a ncRNA.

In the present *de novo* search pipeline, we performed a whole genome Multiple Sequence Alignment (MSA) between the ten budding yeast genomes, which resulted in a set of local MSAs. The MSAs were filtered to discard too small or known protein coding regions. We then applied RNAz [37] to select the MSAs with higher probability of belonging to a ncRNA. Each selected MSA was then evaluated within its genomic context. Unannotated sequences, belonging to selected MSAs for which at least one of the sequences fall into an annotated region, are automatically given the same annotation. MSAs with no known annotation (those falling in intergenic regions for example) must be manually validated and will, eventually, represent new ncRNAs (Fig. 1B).

The automatic steps of the *de novo* pipeline, applied to the 10 genomes, produced 630 candidates distributed in the following way: 376 (60%) previously annotated ncRNA

genes (273 of which correspond to rRNAs or tRNAs), 13 (2%) new ncRNA annotations of expected genes not found with the homology pipeline and 210 (33%) repetitive elements wrongly classified as ncRNA. The remaining 19 (3%) candidates were considered putative new ncRNA genes. Those candidates occur in intergenic regions with no previous annotations, they have a pairwise similarity higher than 50% between all sequences and display potential secondary structures supported by covariation or, at least, some compensatory mutations. The genomic location of some candidates suggests a potential regulatory role, 14 of the 19 genes occur less than 200 from the 5' or 3' UTR of known genes. Notice that all, except one of the candidates, were identified in only two species, a fact which prevents a more detailed analysis of the sequence variation (see Supplementary File SI 1).

Confirmation of the candidates requires experimental validation. However, the fact that 64% of the candidates could be confirmed as real ncRNA supports the assumption that, at least, some of the putative candidates correspond to real ncRNAs genes or regulatory elements. Curiously, the immediate utility of the *de novo* pipeline was the discovery of genes of already known families, functioning as a complement to the homology pipeline.

### 3. Candidate validation

Both search strategies produce many more candidates than expected. Many of the candidates (mainly those with low scores) are FPs that display some sequence or secondary structure resemblance to *bona fide* ncRNAs. Search tools usually assign, to each candidate, a log likelihood score that measures the ratio between the probability of obtaining the candidate using a specific model and the probability of obtaining the same candidate just by chance:

$$score = \log_2 \left( \frac{P(candidate|Model_{ncRNA})}{P(candidate|Model_{random})} \right)$$

Additionally, some tools provide also an E-value for the candidate; it corresponds to the number of candidates with a score better than one would expect to obtain by chance in a sequence with the same characteristics (length and nucleotide, or di-nucleotide, composition).

Scores and E-values provide general guidance to accept or reject a candidate in a first approximation. However, they are not perfect discriminators in the sense that one cannot find a specific value of score or E-value that totally separates FP from TP candidates. In real world applications, any chosen cutoff values of score or E-value will imply a number of FP and FN. It is easy to see that the choice of cutoff value is of great importance. Choosing too high a cutoff value will discard too many positive candidates, while choosing a low value will produce a large number of FPs that will have to be manually validated one-by-one.

The final decision about each candidate must be taken over by the human curator on the basis of a combination of candidate features analysis and experience, a task often difficult or impossible to automate. To systematize the process of human validation of candidates, we established a list of acceptance criteria that must be checked: (i) Extensive sequence similarity on known conserved sequences; (ii) Candidates of families with known guide sequences (such as snoRNAs) should also present the guide sequences compatible with the targets in the same genomes; (iii) Conserved homologous synteny should be observed (similarly, known polycistronic genes should be

occurring together); (iv) Known (or predicted) secondary structures should be supported by covariation and compensatory mutations in the structural alignments. The failure to comply with one or more of the above criteria would not discard a candidate per se, but it would demand stronger evidence for its acceptance.

### 4. Conclusions

Here we described the assembly and application of two automatic ncRNA annotation pipelines to 10 complete genomes of *Saccharomycotina* yeasts. Two annotation strategies were followed, a homologous search and a *de novo* search of ncRNAs. The assembled pipelines are based on publicly available tools and information obtained from ncRNA sequence databases. In total, we were able to find 81% of the expected ncNRAs (693 unique ncRNAs) on the searched genomes, more than doubling the 297 originally annotated ncRNAs, and 26 new candidates with no homologues in the reference species *S. cerevisiae* (Table 2).

The analysis of the ncRNA annotation coverage, species-by-species (Fig. 5B), reveals that the less covered species are those more distant from the *S. cerevisiae* group. This observation shows the importance of close related species queries for homology search and suggests the need for a denser taxonomic sampling in regions of the phylogeny less represented in future genomic sequencing projects [41,42]. As an alternative hypothesis we cannot exclude that some of the ncRNAs that were not found do not exist at all. However, this hypothesis does not allow a direct bioinformatic validation. Comparing the ncRNAs found by each tool (Fig. 5A) we observe that 158 (23%) candidates are found by only one search tool, stressing the complementarity between the used methods.

Although manual validation is still needed, the human effort involved in the annotation process was strongly reduced and focused only on the validation of the automatically selected candidates and not on the search itself.

The real number of ncRNAs present in genomes is an open question [43]. In particular, data from human studies indicate that the number of potential ncRNAs could be much

**Table 2**

Proportion of ncRNAs found with both homology and de novo pipelines, assuming the ncRNAs present in the reference genome *Saccharomyces cerevisiae* as 100%. Rows contain numbers of species, columns contain the percentages of ncRNA found for each family. Rows and columns legends have the same meaning as in Table 1.

	Original Annotation	Rnase P	SRP	Rnase MRP	telomerase	snRNA	snoRNA C/D	snoRNA H/ACA	<i>Sace</i> specific	TOTAL (count)	TOTAL
<i>sace</i>	86	1	1	1	1	5	44	29	4		
		Fraction of found ncRNAs									
<i>cagl</i>	0.10	1.00	1.00	1.00	0.00	1.00	0.95	0.86	0.00	75	0.87
<i>zyro</i>	0.56	1.00	1.00	1.00	0.00	1.00	0.98	0.79	0.00	74	0.86
<i>sakl</i>	0.58	1.00	1.00	1.00	0.00	1.00	1.00	0.90	0.00	78	0.91
<i>klth</i>	0.57	1.00	1.00	1.00	0.00	1.00	1.00	0.86	0.00	77	0.90
<i>klla</i>	0.58	1.00	1.00	1.00	0.00	1.00	1.00	0.90	0.00	78	0.91
<i>ergo</i>	0.87	1.00	1.00	1.00	0.00	1.00	0.95	0.83	0.00	74	0.86
<i>deha</i>	0.09	1.00	1.00	1.00	0.00	1.00	0.89	0.66	0.00	66	0.77
<i>piso</i>	0.00	1.00	1.00	1.00	0.00	1.00	0.82	0.66	0.00	63	0.73
<i>yali</i>	0.09	1.00	0.00	1.00	0.00	1.00	0.73	0.45	0.25	53	0.62
<i>arad</i>	0.00	1.00	1.00	1.00	0.00	1.00	0.75	0.48	0.00	55	0.64
<i>Total</i>		1.00	0.90	1.00	0.00	1.00	0.91	0.74	0.03	693	0.81

bigger than the currently annotated ones [44]. Although yeasts have very compact genomes (72% of the genomic sequence corresponds to protein genes) that are, on average, two hundred times smaller than the human genome, the possible existence of a number of yet unidentified ncRNAs cannot be discarded. Several recent observations such as the existence of expressed intergenic regions with no annotated function [45], the detection of several long ncRNAs of unknown function in the *S. cerevisiae* [16] and the extreme difficulty in identifying some elusive ncRNAs (e.g. the RNA component of the Telomerase), raise the question of how many ncRNAs are still to be found. The correct identification of possible new ncRNAs will surely require synergy between pure sequence analysis methods, high throughput techniques of sequencing [46] as well as the application of structural knowledge to ncRNA search.

## 5. Materials and methods

### 5.1. Data sources

The *A. gossypii* genome was obtained from the *Ashbya* Genome Database ([agd.vital-it.ch](http://agd.vital-it.ch)) [23]. The *A. adeninovorans* genome was obtained from (C. Neuvéglise, personal communication) and the *P. sorbitophila* from (V. Leh, personal communication). All other genomes and the original annotations corresponding to 297 ncRNA sequences that were used as BLAST queries are from the Génolevures Database ([www.genolevures.org](http://www.genolevures.org)) [22]. From the yeast genome database ([www.yeastgenome.org](http://www.yeastgenome.org)) [47] we obtained 86 *S. cerevisiae* ncRNAs also used as BLAST queries. From the RFam-ncRNA families database ([rfam.sanger.ac.uk](http://rfam.sanger.ac.uk)) [29] (version 9.1) we downloaded 802 covariance models for INFERNAL search.

### 5.2. Homology pipeline BLAST search

The 383 query sequences were BLASTed against each one of the 10 genomes using the “blastall -p blastn” command (version 2.2.21) with default parameters. No query sequence was BLASTed against its own genome. The obtained candidates with E-Values < 0.1 were retained.

### 5.3. Homology pipeline INFERNAL search

The homology search using INFERNAL (version 1.0) was performed in two steps. First, 83 RFam covariance models, corresponding to the ncRNAs families already known in yeasts were searched and the obtained candidates with E-Values < 0.5 were retained. Second, 719 of the remaining RFam families (corresponding to all the remaining families with exception of the viral and miRNAs families) were searched and the obtained candidates with E-Values < 0.1 were retained. All INFERNAL homology searches were performed using the “cmsearch” command with default parameters.

### 5.4. Candidate selection

The retained candidates were automatically aligned with the known homologous fungal sequences using the

“cmbuild” and “cmalign” commands from the INFERNAL package. Each alignment was manually validated according to the criteria described in the “Candidate Validation” section above.

### 5.5. De novo search

For the *de novo* search, a whole genome MSA was performed using the TBA tool [48] according to the protocol described in “A Practical Guide to Using TBA” ([www.bx.psu.edu/miller\\_lab](http://www.bx.psu.edu/miller_lab)) with the following tree: “((((((sace cagl) zyro) (saki klth)) (klla ergo)) (deha piso)) (yali arad))”. The MSAs smaller than 50 nts or overlapping coding regions were discarded. The remaining MSAs were split using a sliding window of 120 nts with a step of 40 nts. We searched the resulting MSAs with RNAz and retained all candidates with reported probability higher than 0.5. Retained candidates were evaluated according to sequence conservation, secondary structure prediction, possible secondary structure signals such as covariation and compensatory mutations and genomic location with respect to neighboring genes.

### Disclosure of interest

The authors declare that they have no conflicts of interest concerning this article.

### Acknowledgements

The authors are very grateful to Bernard Dujon and Jean-Luc Souciet for help and numerous discussions, and for leading and animating the Génolevures Consortium. JAC is supported by the Ph.D. Program in Computational Biology of the Instituto Gulbenkian de Ciência, Portugal (sponsored by Fundação Calouste Gulbenkian, Siemens SA, and Fundação para a Ciência e Tecnologia; SFRH/BD/33528/2008).

### Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.crv.2011.05.016.

### References

- [1] R.R. Breaker, Complex riboswitches, *Science* (New York N. Y.) 319 (2008) 1795–1797.
- [2] L.S. Waters, G. Storz, Regulatory RNAs in bacteria, *Cell* 136 (2009) 615–628.
- [3] C.P. Ponting, P.L. Oliver, W. Reik, Evolution and functions of long noncoding RNAs, *Cell* 136 (2009) 629–641.
- [4] J. Harrow, A. Nagy, A. Reymond, T. Alioto, L. Patthy, S.E. Antonarakis, et al., Identifying protein-coding genes in genomic sequences, *Genome Biol.* 10 (2009) 201.
- [5] N.B. Leontis, E. Westhof, Geometric nomenclature and classification of RNA base pairs, *RNA* (New York, N. Y.) 7 (2001) 499–512.
- [6] J.Y. Dutheil, F. Jossinet, E. Westhof, Base pairing constraints drive structural epistasis in ribosomal RNA sequences, *Mol. Biol. Evol.* 27 (2010) 1868–1876.
- [7] L. Kretzner, A. Krol, M. Rosbash, *Saccharomyces cerevisiae* U1 small nuclear RNA secondary structure contains both universal and yeast-specific domains, *Proc. Natl. Acad. Sci. U. S. A.* 87 (1990) 851–855.

- [8] R. Kachouri, V. Stribinski, Y. Zhu, K.S. Ramos, E. Westhof, Y. Li, A surprisingly large RNase P RNA in *Candida glabrata* 11 (2005) 1064–1072.
- [9] R. Kachouri-Lafond, B. Dujon, E. Gilson, E. Westhof, C. Fairhead, M.T. Teixeira, Large telomerase RNA, telomere length heterogeneity and escape from senescence in *Candida glabrata*, FEBS Lett. 583 (2009) 3605–3610.
- [10] J. Souciet, M. Aigle, F. Artiguenave, G. Blandin, M. Bolotin-Fukuhara, E. Bon, et al., Genomic exploration of the hemiascomycetous yeasts: 1. A set of yeast species for molecular evolution studies, FEBS Lett. 487 (2000) 3–12.
- [11] J.P. McCutcheon, Computational identification of non-coding RNAs in *Saccharomyces cerevisiae* by comparative genomics, Nucleic Acids Res. 31 (2003) 4119–4128.
- [12] R. Backofen, S.H. Bernhart, C. Flamm, R.G. Hackermu, C. Fried, G. Fritzsche, et al., RNAs everywhere: genome-wide annotation of structured RNAs, J. Exp. Biol. 308B (2007) 1–25.
- [13] S. He, C. Liu, G. Skogerbo, H. Zhao, J. Wang, T. Liu, et al., NONCODE v2.0: decoding the non-coding, Nucleic Acids Res. 36 (2008) D170–D172.
- [14] C.-L. Chen, H. Zhou, J.-Y. Liao, L.-H. Qu, L. Amar, Genome-wide evolutionary analysis of the noncoding RNA genes and noncoding DNA of *Paramecium tetraurelia*, Rna 15 (2009) 503–514.
- [15] J. Hertel, D. de Jong, M. Marz, D. Rose, H. Tafer, A. Tanzer, et al., Non-coding RNA annotation of the genome of *Trichoplax adhaerens*, Nucleic Acids Res. 37 (2009) 1602–1615.
- [16] L.A. Kavanaugh, F.S. Dietrich, Non-coding RNA prediction and verification in *Saccharomyces cerevisiae*, PLoS Genet. 5 (2009) e1000321.
- [17] C. Noirot, C. Gaspin, T. Schiex, J. Gouzy, LeARN: a platform for detecting, clustering and annotating, BMC Bioinformatics 11 (2008) 1–11.
- [18] P. Menzel, J. Gorodkin, P.F. Stadler, The tedious task of finding homologous noncoding RNA genes, RNA (New York N. Y.) 15 (2009) 2075–2082.
- [19] Jia, H., Osak, M., Bogu, G.K., Stanton, L.W., Johnson, R., Lipovich, L., Genome-wide computational identification and manual annotation of human long noncoding RNA genes, RNA 16 (2010) 1478–1487.
- [20] X. Xu, Y. Ji, G.D. Stormo, Discovering cis-regulatory RNAs in *Shewanella* genomes by support vector machines, PLoS Comput. Biol. 5 (2009) e1000338.
- [21] Y. Zhang, J. Wang, S. Huang, X. Zhu, J. Liu, N. Yang, et al., Systematic identification and characterization of chicken (*Gallus gallus*) ncRNAs, Nucleic Acids Res. 37 (2009) 6562–6574.
- [22] D. Sherman, P. Durrrens, E. Beyne, M. Nikolski, Génolevures: comparative genomics and molecular evolution of hemiascomycetous yeasts, Nucleic Acids Res. 32 (2004) D315–D318.
- [23] A. Gattiker, R. Rischatsch, P. Demouglin, S. Voegeli, F.S. Dietrich, P. Philippsen, et al., Ashbya Genome Database 3.0: a cross-species genome and transcriptome browser for yeast biologists, BMC Genomics 8 (2007) 9.
- [24] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
- [25] J.A. Doudna, T.R. Cech, The chemical repertoire of natural ribozymes, Nature 418 (2002) 222–228.
- [26] S.R. Eddy, R. Durbin, RNA analysis using covariance models, Nucleic Acids Res. 22 (1994) 2079–2088.
- [27] S. Lindgreen, P.P. Gardner, A. Krogh, Measuring covariation in RNA alignments: physical realism improves information measures, Bioinformatics (Oxford, England) 22 (2006) 2988–2995.
- [28] E.P. Nawrocki, D.L. Kolbe, S.R. Eddy, Infernal 1.0: inference of RNA alignments, Bioinformatics (Oxford, England) 25 (2009) 1335–1337.
- [29] P.P. Gardner, J. Daub, J.G. Tate, E.P. Nawrocki, D.L. Kolbe, S. Lindgreen, et al., Rfam: updates to the RNA families database, Nucleic Acids Res. 37 (2009) D136–D140.
- [30] C. Brown, H. Davis, Receiver operating characteristics curves and related decision measures: A tutorial, Chemom. and Intell. Lab. Syst. 80 (2006) 24–38.
- [31] T. Kubodera, M. Watanabe, K. Yoshiuchi, N. Yamashita, A. Nishimura, S. Nakai, et al., Thiamine-regulated gene expression of *Aspergillus oryzae* thiA requires splicing of the intron containing a riboswitch-like domain in the 5'-UTR, FEBS Lett. 555 (2003) 516–520.
- [32] M.T. Cheah, A. Wachter, N. Sudarsan, R.R. Breaker, Control of alternative RNA splicing and gene expression by eukaryotic riboswitches, Nature. 447 (2007) 497–500.
- [33] T.E. Edwards, A.R. Ferré-D'Amaré, Crystal structures of the thi-box riboswitch bound to thiamine pyrophosphate analogs reveal adaptive RNA-small molecule recognition, Structure (London, England: 1993) 14 (2006) 1459–1468.
- [34] S. Thore, M. Leibundgut, N. Ban, Structure of the eukaryotic thiamine pyrophosphate riboswitch with its regulatory ligand, Science 312 (2006) 1208–1211.
- [35] D.C. Zappulla, T.R. Cech, Yeast telomerase RNA: a flexible scaffold for protein subunits, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 10024–10029.
- [36] S. Gunisova, E. Elboher, J. Nosek, V. Gorkovoy, Y. Brown, J.F. Lucier, et al., Identification and comparative analysis of telomerase RNAs from *Candida* species reveal conservation of functional elements, RNA (New York, N.Y.) 15 (2009) 546–559.
- [37] S. Washietl, I.L. Hofacker, P.F. Stadler, Fast and reliable prediction of noncoding RNAs, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 2454–2459.
- [38] Z. Yao, Z. Weinberg, W.L. Ruzzo, CMfinder—a covariance model based RNA motif finding algorithm, Bioinformatics (Oxford, England) 22 (2006) 445–452.
- [39] E. Rivas, S.R. Eddy, Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs, Bioinformatics 16 (2000) 583–605.
- [40] P. Clote, F. Ferré, E. Kranakis, D. Krizanc, Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency, RNA (New York N. Y.) 11 (2005) 578–591.
- [41] S.M. Hedtke, D.M. Hillis, Taxon sampling and the accuracy of phylogenetic analyses, Evolution 46 (2008) 239–257.
- [42] D. Wu, P. Hugenholtz, K. Mavromatis, R. Pukall, E. Dalin, N.N. Ivanova, et al., A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea, Nature 462 (2009) 1056–1060.
- [43] J.S. Mattick, The functional genomics of noncoding RNA, Science (New York N. Y.) 309 (2005) 1527–1528.
- [44] P. Kapranov, J. Cheng, S. Dike, D.A. Nix, R. Duttagupta, A.T. Willingham, et al., RNA maps reveal new RNA classes and a possible function for pervasive transcription, Science (New York N. Y.) 316 (2007) 1484–1488.
- [45] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, et al., The transcriptional landscape of the yeast genome defined by RNA sequencing, Science (New York N. Y.) 320 (2008) 1344–1349.
- [46] Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics, Nat. Rev. Genet. 10 (2009) 57–63.
- [47] J.M. Cherry, C. Adler, C. Ball, S.A. Chervitz, S.S. Dwight, E.T. Hester, et al., SGD: *Saccharomyces* Genome Database, Nucleic Acids Res. 26 (1998) 73–79.
- [48] M. Blanchette, W.J. Kent, C. Riemer, L. Elnitski, A.F.A. Smit, K.M. Roskin, et al., Aligning multiple genomic sequences with the threaded blockset aligner, Genome Res. 14 (2004) 708–715.

## Chapter 6

# Detection of Structural Modules

An important characteristic of structured ncRNAs is the existence of identifiable three dimensional modules recurrently found in many evolutionary and functionally distinct molecules. Apart from the helices and hairpins that constitute the basic framework of an RNA architecture and are, by far, the most common of these recurrent elements, a number of other, more specific, recurrent modules occur in key structural and functional regions. They play important architectural and functional roles as they are often protein binding sites and structural organizers of the molecules. These modules are usually known as structural modules (or structural motifs) and their importance is evidenced by the number of tools and approaches recently proposed for module discovery on atomic resolution structures (Duarte et al., 2003; Hershkovitz, 2003; Wadley and Pyle, 2004; Wang et al., 2007; Djelloul and Denise, 2008; Sarver et al., 2008; Apostolico et al., 2009; Zhong et al., 2010). Remarkably, despite all the efforts put into module discovery in three dimensional structures, the discovery of modules in sequences was until now an almost unexplored subject.

In the present chapter I describe a novel approach for structural modules discovery in RNA sequences based on bayesian networks, joint base pair probability estimation and positional candidates clustering. I will also present several promising results obtained by applying this approach to publicly available ncRNA alignments.

### 6.1 Introduction to Structural Modules

Non coding RNA structural modules can be defined as:

“(...)directed and ordered stacked arrays of non-Watson-Crick base pairs forming distinctive foldings of the phospho-

diester backbones of the interacting RNA strands. They correspond to the 'loops' – hairpin, internal and junction – that intersperse the Watson-Crick two-dimensional helices as seen in two-dimensional representations of RNA structure(...)" (Leontis and Westhof, 2003b).

In other words, RNA structural modules are recurrent RNA elements that occur in phylogenetically and functionally unrelated molecules and present similar three-dimensional shapes independently of the surrounding structural context.

RNA structural modules are also known as RNA motifs, especially when reference is made to sequence. Here I prefer the term "structural modules" to distinguish from the different, although closely related, concepts such as sequence motifs – sequential patterns of nucleotides – and RNA motifs – sets of secondary structure elements. In the following I will refer to RNA structural modules simply as modules.

Although no complete catalog of structural modules is available at present, many modules can be found in the literature and some of them are consensual among the RNA structure community. The following list mentions the most commonly referred structural modules.

- **A-minor**: Formed by two consecutive adenines that form sugar-sugar base pairs with bases at the codon and anticodon positions during the translation process (Lescoute and Westhof, 2006).
- **AA-platform**: Occurs in three different positions at group I introns (Cate et al., 1996).
- **Bulge-helix-bulge**: Occurs in archaea introns in protein binding sites. It is recognized by spliceosomal protein complex (Diener and Moore, 1998).
- **C-loop**: Structural motif occurring in internal loops at several different positions of the ribosome (Lescoute et al., 2005).
- **G-bulged**: Found in many RNAs, specially the ribosomal subunits and some riboswitches. Stabilizes the structure and occurs at protein binding sites (Szewczak et al., 1993).
- **Kink-turn**: Found in many RNAs, U4 snRNA, some riboswitches and probably in snoRNAs. It provokes a sharp bend in the backbone and, in the ribosome, commonly occurs at protein binding sites (Klein et al., 2001).
- **Loop E**: Presents some similarities with the G-bulged. It is present in the bacterial 5S ribosomal RNA (Wimberly et al., 1993; Correll et al., 1997; Leontis and Westhof, 2003a).



- **Loop-helix interaction:** Is a tertiary interaction between a tetraloop and an helical region from distant regions of the secondary structure (Michel and Westhof, 1990).
- **Reverse Kink-turn:** Strikingly similar to the kink-turn but it bends to the opposite direction. It is only known to occur in group I introns (Antonioli et al., 2010).
- **Tandem GA:** Tandem GA-AG sugar-Hoogsteen base pairs that stack between regular WC base pairs (Gautheret et al., 1994).
- **Tetraloop:** Common four nucleotides terminal loops presenting the common sequence motifs **UNCG** or **GNRA** (Tuerk et al., 1988; Woese et al., 1990).
- **U Turn:** Terminal loop motif found in the anticodon loop of tRNAs (Schweisguth and Moore, 1997).

Most of the listed motifs are described in depth in the following papers (Moore, 1999; Leontis and Westhof, 2003b; Hendrix et al., 2005; Lescoute et al., 2005).

In this work I selected four modules that are among the most common of the observed motifs: G-bulged, kink-turn, C-loop and Tandem GA.

## 6.2 Search for Structural Modules

As it has been seen in the introduction, a number of tools is available for structural modules search within RNA atomic structures. No tool, however, is available to search for modules in sequences alone. The functional importance of modules and the fact that the number of ncRNAs for which atomic structures are available is still small in comparison with the number of available sequences was our main motivation for the development of such an approach.

### 6.2.1 Interaction Networks

Atomic structure models provide us with information about the three dimensional shape of modules. This structural information can be summarized with enough fidelity<sup>1</sup> by the set of base-base interactions of the module. This set of interactions is commonly known as the Interaction Network (IN) of the module. The IN derived from a crystal structure (see Figure 6.1) represents a single instance of the all the possible module instances and, even if one collects all the available crystal structures of the module, one would still obtain a very limited sample of the possible variations of those modules.

---

<sup>1</sup>By “enough fidelity” I mean that the set of base-base interactions provide sufficient information to rebuild a meaningful three dimensional representation of the module.

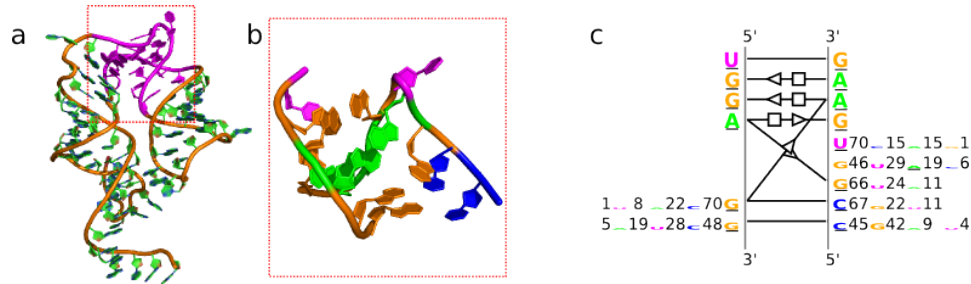


Figure 6.1: Atomic structure of a kink-turn and the respective Interaction network.

A way to increase the information contained on an IN is to improve it with the nucleotide frequencies for each position of the network by collecting data from high quality sequence alignments in which the module region can be well identified and aligned. If one compiles all the available INs in a “consensus” network one can obtain a reasonable representation of the possible sequence and structural diversity of a module and a good starting point to build a search model for structural modules (see Figure 6.2).

### 6.2.2 Position Weight Matrices

It is obvious from the displayed INs (see Figures 6.1 and 6.2) that each strand of the motif can be characterized by a sequence motif. A classic way to represent sequence motifs is to build a position weight matrix (PWM).

In a PWM, each row corresponds to a symbol of the sequence alphabet (in our case, the four nucleotides A, C, U and G), each column corresponds to a position of the motif and each individual position of the PWM corresponds to the log likelihood of observing the symbol  $i$  in position  $j$ :

$$PWM_{i,j} = \log_2 \left( \frac{P(a_j = r_i | Model)}{P(a_j = r_i | Null)} \right),$$

in which  $P(a_j = r_i | Model)$  represents the probability of observing the nucleotide  $r_i$  in the  $j^{th}$  position of the motif and  $P(a_j = r_i | Null)$  the probability of observing  $r_i$  in that same position just by chance<sup>2</sup>

The score of a given sequence  $s$  can be computed as:

$$score(s) = \sum_{i=1}^{|s|} PWM_{s_j, j}.$$

Notice that the positions of the PWM can be summed due to the assumption of independence between the positions of the sequence motif. This

<sup>2</sup>Usually this corresponds to the probability of observing  $r_i$  in a random sequence with the same nucleotide distribution as the original sequences, i.e., with the same GC content.



independence assumption is important to simplify the training and the use of the PWM. However, usually it does not apply, either in typical DNA motifs (Ben-Gal et al., 2005) or in structured RNAs where the WC base pairs present a strong dependency between bases normally.

To test the result of the use of a position independent model to search for a kink-turn model in a ribosomal sequence, I built the most stringent kink-turn model and went through the large ribosomal subunit of *Aciduliprofundum boonei*. Figure 6.3 shows the result of the search. Notice that 3 out of 5 kink-turns are found by the search. However, more than 30 false positive hits are also found making it particularly difficult to select the real from the false candidates.

### 6.3 RMDetect Approach

From the previous example it is clear that position independent models are not enough to discriminate the modules from the background noise within a long sequence. To do that one needs to obtain more information on the structural context of the model. Thus, I developed a module search strategy based on the following assumptions:

- The nucleotides present in the different positions of a module are not independent of each other and sometimes (as in the WC base pair case) are strongly correlated;
- Structural modules are usually flanked by secondary structure elements such as helices and loops;
- Modules are conserved across species.

Then I combined the information obtained by applying each of these assumptions to module search in a single search approach.

#### 6.3.1 Modeling with Bayesian Networks

To model the dependency between the positions of a module I resorted to the Bayesian Network formalism. Bayesian Networks have been applied to transcription factor binding site identification and proved a simple yet useful model to describe dependency between sequence sites (Barash et al., 2003).

A Bayesian Network (BN) is a probabilistic graphical model represented as a directed acyclic graph in which the nodes represent variables and the edges represent the dependency between them. A probability distribution is associated to each node of the BN and this probability is conditioned by the parent nodes.

Formally, a BN can be defined as:

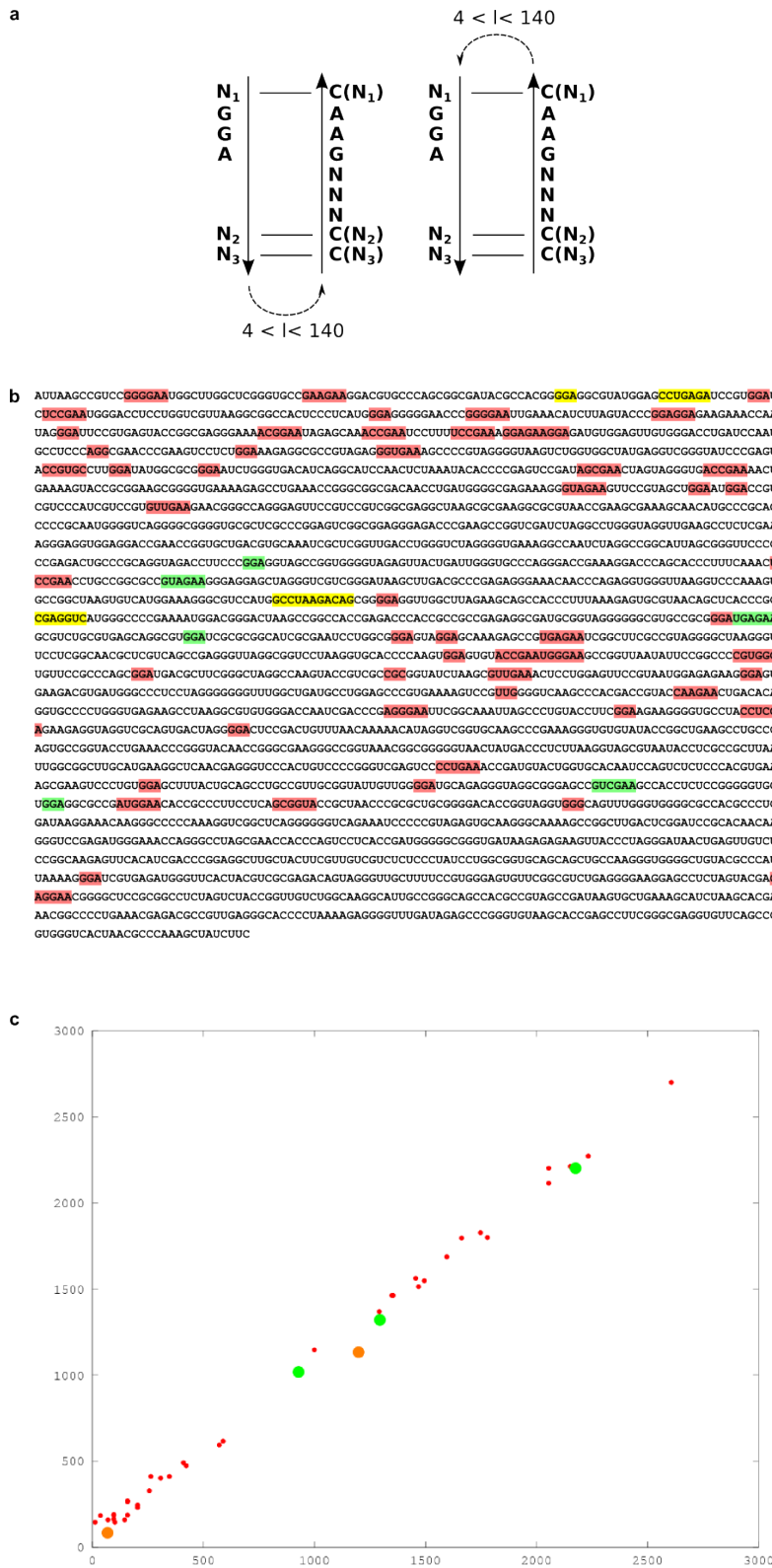
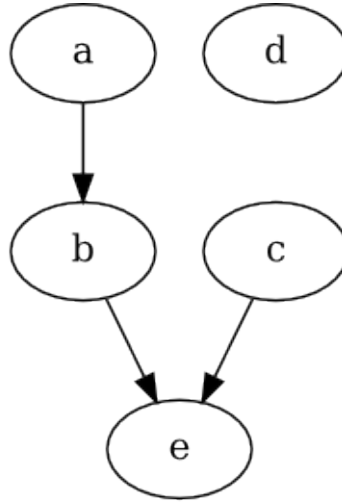


Figure 6.3: Kink-turn search using a simple position independent model. From the 5 existing kink turns 3 were correctly predicted (large green dots) and 2 were missed (large orange dots). False positives are represented as small red dots.



$$P(a,b,c,d,e) = P(a) P(c) P(d) P(b|a) P(e|b,c)$$

Figure 6.4: Simple Bayesian Network

$$BN = (N, E),$$

in which  $N$  is the set of nodes and  $E$  the set of edges. An edge ( $e \in E$ ) in its turn is defined as an ordered pair:

$$e = (a, b) : a, b \in N.$$

The set  $Pa_x$  of the parent nodes of  $x$  can be defined as:

$$Pa_x = y : \exists e \in E, e(y, x).$$

The joint probability density function of a BN is given by:

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^{i=n} P(X_i = x_i | X_j = x_j, \forall X_j \in Pa_{x_i})$$

Figure 6.4 exemplifies a simple BN and the joint probability of all its nodes.

To represent a module as a BN, first one has to establish the topology of the network and then to compute the parameters describing the nucleotide frequencies for each node. Inferring the topology of a BN is a NP-Hard problem (Chickering et al., 1994) and several approaches and tools exist to infer BN topologies (Hartemink et al., 2005; Shah and Woolf, 2009). Here I adopted a conservative approach in which the edges of the network correspond to the physical interactions between the bases such as base-base

or stacking interactions. To each node I associate a multinomial distribution corresponding to the relative frequency of each nucleotide in each given position. Figure 6.5 depicts the derived BNs for the studied modules.

### 6.3.2 Adding Base Pair Probability Information

Due to the short sequence signatures of modules – generally less than 30 nts – searching with a BN in a long sequence will still produce many false positive candidates occur just by chance. Real modules, however, are flanked by WC base pairs that belong to secondary structure elements of the molecule. False positive candidates, in general, are not forced to comply with the secondary structure of the molecule, thus I can use the predicted lower base pair probabilities to filter those false positive candidates. This prediction can be made using publicly available secondary structure prediction tools (Hofacker et al., 1994). Figure 6.6 shows the location of three modules predicted in a lysine riboswitch sequence. Notice that the Tandem GA candidate although very similar to a possible module is incompatible with the predicted secondary structure.

### 6.3.3 Adding Alignment Information

If several homologous sequences are available – which is more and more the case given the increasing availability of sequence databases – one can use the multiple sequence alignment data to improve module predictions. Assuming that modules are conserved across the homolog species – which is supported by observations – one would expect to find them in close columns of the alignment even if the alignment is not perfect for that region. Additionally, if sequences are sufficiently divergent, the false positive candidates should be distributed randomly across the alignment.

In the proposed approach I use this information to cluster the found candidates and select the clusters that most likely correspond to real modules. Figure 6.7 shows this search strategy applied to an alignment of the lysine riboswitch.

### 6.3.4 Automatic Construction of Models

To allow the search of other structural modules, beyond the four modules initially considered in our work we developed *RMBuild*, a tool for automatically building search models given training data. *RMBuild* requires as input two pieces of information:

- The atom coordinates of the structural module to consider.
- A multiple sequence alignment containing a representative sample of the module sequences.

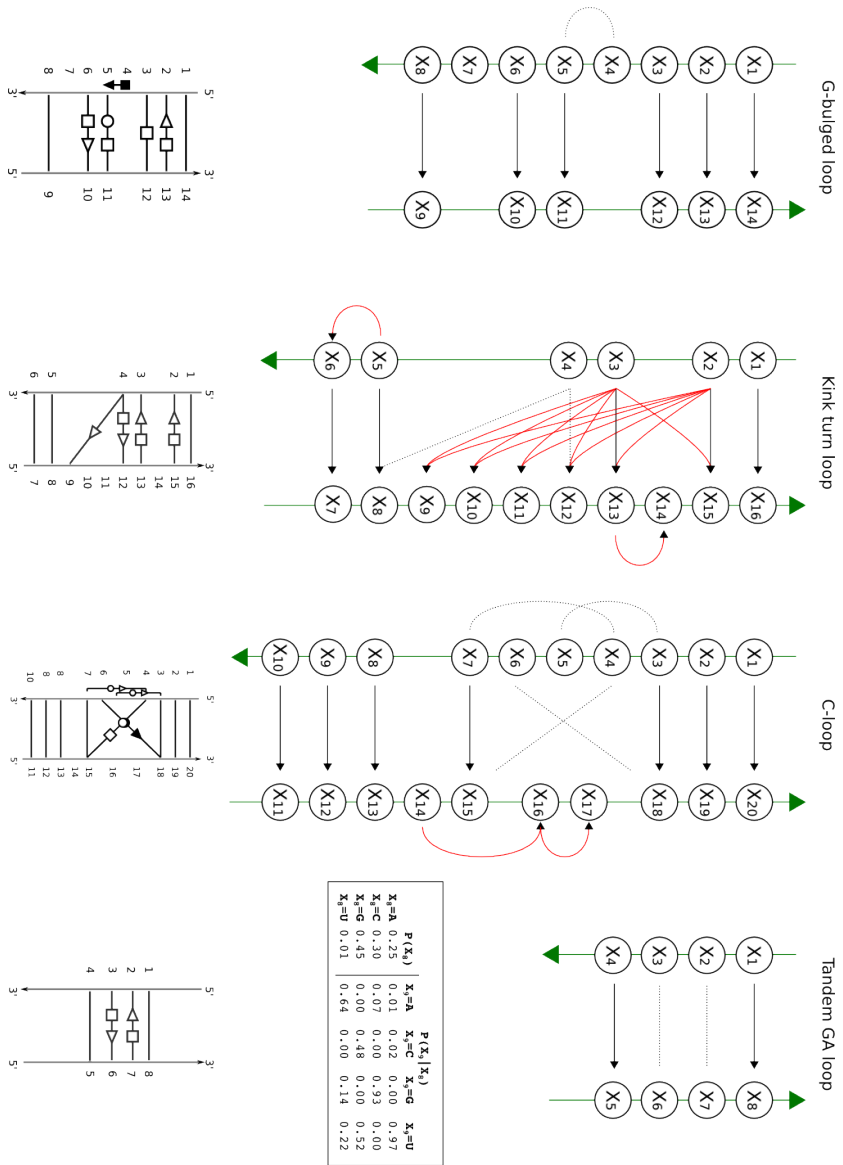


Figure 6.5: Bayesian Networks for the four studied modules. Figure extracted from (Cruz and Westhof, 2011b).



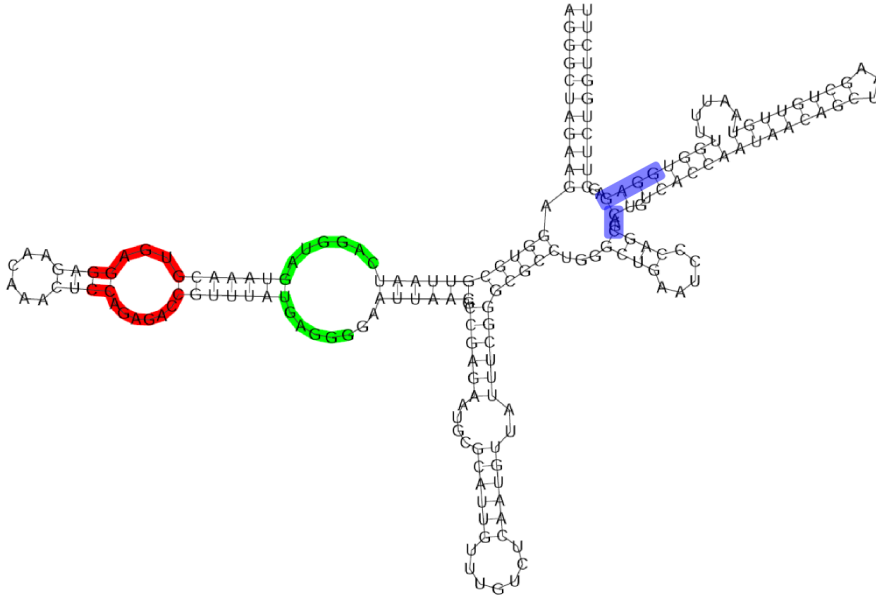


Figure 6.6: Predicted modules in the Lysine riboswitch secondary structure.

Based on this information `RMBuild` automatically generates a new BN model for the module according to the following steps:

1. Identify all the base-base interactions occurring between the bases of the module using automatic 3D annotation tools like `MC-Annotate` (Gendron et al., 2001) or `RNAView` (Yang et al., 2003).
2. Infer a possible structure for the BN assuming as the edges of the network the stacking, WC and non-WC base-base interactions, taking care to avoid building a cycle graph, i.e. avoiding circular references between nodes, (see Section 6.3.1).
3. Compute the multinomial distribution of the nodes probabilities from the frequency of the nucleotides in the alignment. The conditional probabilities are computed according to the BN defined in the previous step.

The result of this protocol is a model file, describing the BN and the probabilities of each node, that can be used by `RMDetect` to search for the new module. This definition file is a simple text file and users can change it in order to fine tune the inferred BN. In this case `RMBuild` can be used to recompute the nodes probabilities based on the new structure. It corresponds to executing the steps 1 and 2 of the protocol manually letting `RMBuild` to execute step 3.

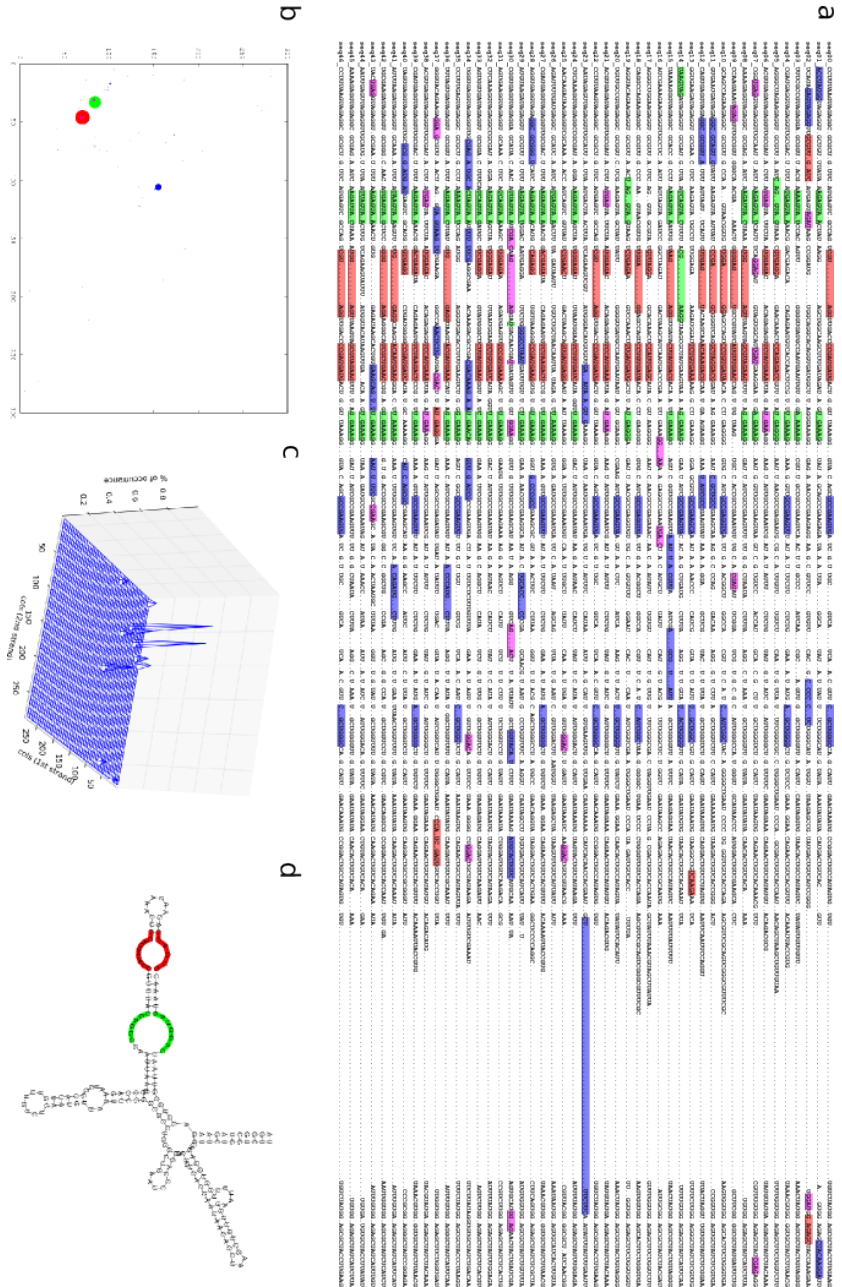


Figure 6.7: Module clustering on a lysine riboswitch alignment.

## 6.4 Search Algorithms

In this section I present the formal definition of each of the algorithms informally described above.

### 6.4.1 Single Sequence Search Algorithm

For a formal definition of the algorithm for module search in single sequences let:

- $M$  be a structural RNA module;
- $S$  be a nucleotide sequence to be searched for  $M$ ;
- $M_{BN}$  be a Bayesian Network model of  $M$ ;
- $M_{GC}$  be a null model in which all the positions are independent and have the same nucleotide distribution of  $S$ ;
- $sp_{ij} = (seq_i, seq_j)$  be a pair of non overlapping sub sequences of  $S$  starting from positions  $i$  and  $j$ , corresponding to the strands of the module;
- $WC_M$  be the set of all WC base pairs of  $M$ ;
- $FE_{all}$  be the free energy of a folding ensemble corresponding to the folding of the unconstrained original sequence;
- $FE_{ij}$  be the free energy of a folding ensemble corresponding to the folding of the original sequence constrained by the base pairs of  $WC_M$  in the positions determined by  $sp_{ij}$ .

For simplicity I will only describe modules formed by pairs of sub-sequences, i.e., modules with two strands. The extension to modules formed by more than two strands, like in n-way junctions, would simply require redefinition of  $sp$  as a tuple  $sp_{i_1, \dots, i_n} = (seq_{i_1}, \dots, seq_{i_n})$  and to include the respective additional nested for loops in Algorithm 2.

### 6.4.2 Multiple Sequence Clustering Algorithm

As seen in the single sequence search algorithm each module candidate can be defined as an ordered pair of alignment coordinates  $sp_{ij} = (seq_i, seq_j)$ . A hierarchical clustering algorithm can be applied to the selected candidates of the different sequences according to their distance. The algorithm goes as follows:

---

**Algorithm 2** Single sequence search

---

```

selected =  $\emptyset$ 
for i IN len(S) do
  for j IN len(S) do
    if i  $\neq$  j then
       $score_{ij} = \log_2 \left( \frac{P(sp_{ij}|M_{BN})}{P(sp_{ij}|M_{GC})} \right)$ 
       $bpp_{ij} = e^{\left( \frac{FE_{all} - FE_{ij}}{kT} \right)}$ 
      if  $score_{ij} > limit_{score}$  AND  $bpp_{ij} > limit_{bpp}$  then
        append  $sp_{ij}$  to selected
      end if
    end if
  end for
end for

```

---



---

**Algorithm 3** Multiple Sequence Clustering

---

```

# Remove from selected all overlapping candidates on the same sequence
# retaining only the one with the higher score.
for candij IN selected do
  append candij to clusters
end for
repeat
  exit = TRUE
  for clusterij IN clusters do
    for clusterkl IN clusters do
      if i  $\neq$  k AND j  $\neq$  l AND  $max(|i - k|, |j - l|) \leq DIST$  then
        clusterij = MERGE (clusterij, clusterkl)
        exit = FALSE
      end if
    end for
  end for
until exit == TRUE

```

---

### 6.4.3 Candidate Evaluation

At the end of the algorithm each candidate will be characterized by five variables:

1. *occur* – the absolute number of the aligned sequences in which the candidate occurs;
2. *perc* – the percentage of the aligned sequences in which the candidate occurs (occurrence);
3. *score* – the mean score of all candidates (see 3);
4. *bpp* – the mean bpp of all candidates (see 3);
5. *MI* – the mutual information between the bases of each *WC* base pair from  $WC_M$ , measured along all candidates (see A).

For a candidate to be selected it must be sufficiently represented in the alignment (*occur* and *perc*) must have a significant *score* and *bpp* and should present some covariance between WC base pairs supporting the evolutionary pressure on conservation of the secondary structure of the module (*MI*). The actual values used as threshold for each decision are described in (Cruz and Westhof, 2011b)

## 6.5 Results

I applied the described algorithms to 1444 alignments from public databases: 1309 alignments from Rfam (Gardner et al., 2009), 14 alignments from group I intron database (Zhou et al., 2008) and 121 bacterial ncRNA alignments from recent metagenomic studies (Weinberg et al., 2009; Weinberg et al., 2010). *RMDetect* found 141 of known modules and predicted 21 yet unreported modules. Table 6.5 summarizes the results obtained in the search. For a complete description of the results see (Cruz and Westhof, 2011b).

## 6.6 Conclusions

The present approach to detect RNA structural modules in sequences was able to detect a number of already known modules and to propose several yet undetected module instances. I hope that this approach and the tools provided with it could be useful as is for biologists working with specific RNA molecules for which no structural information is available. Additionally I

<b>Alignments</b>	<b>Results</b>	<b>G-bulged</b>	<b>Kink-turn</b>	<b>C-loop</b>	<b>Tandem GA</b>
Rfam (1309)	total candidates	13	119	22	68
	known modules	6	105 <sup>†</sup>	0	21 <sup>‡</sup>
	new candidates	1	1	3	8
	not confirmed	6	13	19	39
Group I Introns (14)	total candidates	1	1	0	1
	known modules	1	0	0	0
	new candidates	0	0	0	1
	not confirmed	0	1	0	0
Rfam (121)	total candidates	4	4	1	16
	known modules	3	0	0	5*
	new candidates	1	1	0	5
	not confirmed	0	3	1	6

Table 6.1: Results of RMDetect on public database alignments. a)Number of alignments searched in the database. b)Number of selected candidates. c)Number of selected candidates corresponding to known modules. d)Number of selected candidates corresponding to new putative modules. e)Number of false positive candidates or candidates for which no confirmation was possible. <sup>†</sup>99 snoRNAs. <sup>‡</sup>20 kink-turns. \*1 kink-turn.

expect to extend `RMDETECT` in order to include the remaining well known modules and several other less known for which just a few examples are available.

## 6.7 Article – Sequence-based Identification of 3D Structural Modules in RNA with RMDetect

This chapter is an extended summary of the following article:

Cruz, J. A. and Westhof, E. (2011). *Sequence-based Identification of 3D Structural Modules in RNA with RMDetect*. *Nature Methods* (8)6:513-519.



# Sequence-based identification of 3D structural modules in RNA with RMDetect

José Almeida Cruz & Eric Westhof

**Structural RNA modules, sets of ordered non-Watson-Crick base pairs embedded between Watson-Crick pairs, have central roles as architectural organizers and sites of ligand binding in RNA molecules, and are recurrently observed in RNA families throughout the phylogeny. Here we describe a computational tool, RNA three-dimensional (3D) modules detection, or RMDetect, for identifying known 3D structural modules in single and multiple RNA sequences in the absence of any other information. Currently, four modules can be searched for: G-bulge loop, kink-turn, C-loop and tandem-GA loop. In control test sequences we found all of the known modules with a false discovery rate of 0.23. Scanning through 1,444 publicly available alignments, we identified 21 yet unreported modules and 141 known modules. RMDetect can be used to refine RNA 2D structure, assemble RNA 3D models, and search and annotate structured RNAs in genomic data.**

Structured RNAs present hierarchical architectures in which double-stranded helices and single-stranded loops are organized in three-dimensional (3D) space by tertiary interactions. The helices are formed by stacks of Watson-Crick base pairs and the tertiary interactions consist mainly of non-Watson-Crick base pairs<sup>1</sup>. Tertiary interactions occur either between nucleotides in the same domain (for example, internal loops and junctions) or between nucleotides from distant domains (for example, loop-loop and loop-helix interactions, and pseudoknots). The tertiary interactions, by establishing local and specific contacts, build up 3D structural modules that are characterized by sets of non-Watson-Crick base pairs organized in a precise order. Modules occur recurrently in different RNAs stemming from any phylogenetic branch and display similar 3D shapes independently of the surrounding structural context<sup>2</sup>. They have important functional roles in RNA molecules as protein and RNA binding sites<sup>3</sup> and as local structural organizers in junctions or internal loops<sup>4</sup>.

Structural RNA modules are often referred to as RNA motifs. We favor the term ‘module’ to distinguish between closely related, although distinct, concepts: sequence motifs, which are patterns of nucleotides; RNA motifs, which are sets of secondary structure elements (helices, single strands, hairpins, loops and others)<sup>5,6</sup>; and structural RNA modules, which are ensembles of stacked

arrays of ordered non-Watson-Crick base pairs<sup>3</sup>. This distinction separates ‘objects’ that exist in tertiary structure from those that exist only in sequence.

The identification of a module in a RNA sequence can provide key information about the secondary structure and the resulting tertiary fold<sup>7–9</sup>. Therefore, the identification of a structural RNA module lends support to the identification of a transcript as a structured RNA<sup>10</sup>, presents clues for the local function of the molecule<sup>11,12</sup> and explains chemical probing data because modules present defined chemical probing signatures and mutational data<sup>4</sup>. Recent tools for module searching in structures<sup>13–16</sup> illustrate the importance of module discovery. However, none of these tools have been designed to find modules in sequence alone.

Several RNA motif search tools are currently available. Some (RNAMotif<sup>5</sup> or MilPat<sup>17</sup>) rely on user-defined descriptors of sequence and secondary structure. Others (CMFinder<sup>6</sup>) infer assemblies of secondary-structure elements from homologous sequences. These tools search for specific secondary structure elements that can span up to hundreds of nucleotides with extensive helical regions and perform poorly when searching small sequence motifs with less than 20 nucleotides (**Supplementary Note 1**). The 3D structure prediction tools can, in theory, provide information about structural modules, but they require considerable amount of computer resources and expertise.

Here we present a computational tool for structural RNA module searching based solely on sequence information, which we called RNA 3D modules detection (RMDetect). To capture all the possible variations of the allowed tertiary interactions and base pairs, RMDetect relies on Bayesian network models, base-pair probability prediction and positional clustering of candidates. We tested the performance of RMDetect on 1,444 noncoding (nc)RNA alignments for finding four recurrent modules: G-bulge loop (referred to as G-bulge)<sup>4</sup>, kink-turn<sup>2,12</sup>, C-loop<sup>2</sup> and tandem GA/AG loop (referred to as tandem GA)<sup>18</sup>. From the 1,444 alignments, we identified 141 cases of known instances of the modules and 21 new candidates. RMDetect can be used on single sequences or on multiple sequence alignments and can be applied to any newly discovered module irrespective of the complexity or number of strands involved. The use of RMDetect with 2D structure algorithms can improve accuracy of predictions.

Architecture et Réactivité de l'ARN, Institut de Biologie Moléculaire et Cellulaire du Centre National de la Recherche Scientifique, Université de Strasbourg, Strasbourg, France. Correspondence should be addressed to E.W. (e.westhof@ibmc-cnrs.unistra.fr).

RECEIVED 8 NOVEMBER 2010; ACCEPTED 11 APRIL 2011; PUBLISHED ONLINE 8 MAY 2011; DOI:10.1038/NMETH.1603

Together with presently available modeling tools<sup>7–9</sup>, RMDetect can be used to build relevant RNA models and also to search and annotate ncRNAs in genomic data.

Other modules not covered by the current implementation of RMDetect exist and new modules are likely yet to be discovered. Some structured RNA may not contain any of the modules discussed here. Therefore we also provide a tool to build Bayesian network models corresponding to new modules based simply on 3D coordinates of the new module and sequence alignments representative of the module, called RNA 3D modules builder or RMBuild.

## RESULTS

### Structural RNA modules and interaction networks

The most accurate way to characterize a module and its interaction network is to analyze crystal structures. The comparison of many instances of a given module conveys essential information about its structural regularity and variation. However, typically only a few of the possible sequences compatible with the given module are found in existing crystal structures. To obtain a larger sample of the range of possible sequence variation one must resort to carefully curated alignments of homologous sequences. Such alignments indicate the nucleotides that can occur at each position of a module.

Interaction networks represent both the sequential regularity and the variation present in structural modules without atomic details. They depict the nucleotide frequencies and base-base interactions for each module instance. After merging the interaction networks of all instances, one obtains an integrated interaction network that captures the full sequential regularity and variation of that module, irrespective to the specific molecule in which it is embedded and of the module location. We selected four known recurrent modules because they have key roles in many types of RNAs (**Fig. 1**).

### Descriptions of modules

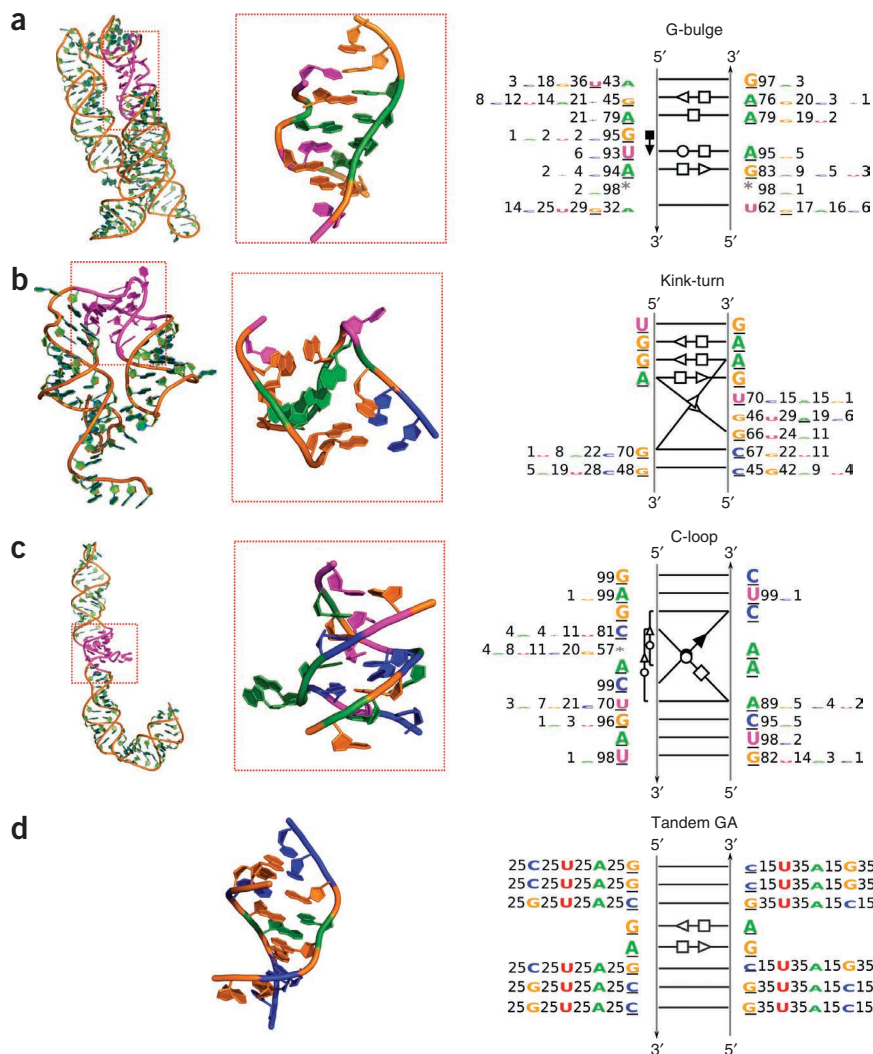
The G-bulge module is observed in the three rRNAs<sup>19</sup>, in the lysine riboswitch<sup>20</sup>, in the group I intron P7.1/P7.2 domain<sup>21</sup> and in the T-box leader<sup>22</sup>. G-bulge modules are formed by four stacked non-Watson-Crick base pairs (**Fig. 1a**) with a characteristic bulging G that participates in a triple interaction with the flanking base pair. The G-bulge module organizes internal loops and junctions, and often forms binding platforms for proteins<sup>4,19</sup>.

The kink-turn module, an asymmetric internal loop, leads to a sharp bend between two helical regions<sup>12</sup> (**Fig. 1b**). One of the helices contains exclusively Watson-Crick base pairs, and the

three base pairs of second helical region, closest to the internal loop, usually form a GAA/GGA Hoogsteen-Sugar edge platform<sup>2</sup>. The kink-turn modules bind several ribosomal proteins. The *U4* small nuclear (sn)RNA and small nucleolar (sno)RNAs bind the 15.5 kDa protein in eukaryotes and the homologous archaeal protein L7 (ref. 23).

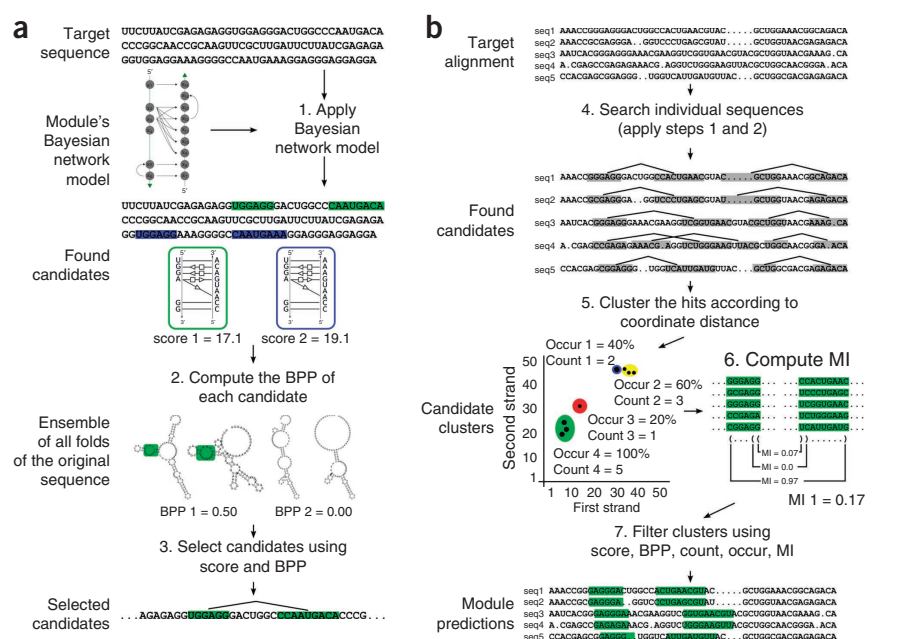
The C-loop module is an asymmetrical internal loop between two canonical helices. C-loops increase the helical twist between the helices<sup>2</sup> (**Fig. 1c**). C-loop modules have been observed in rRNAs and in a synthetase mRNA regulatory element<sup>24</sup>.

The tandem-GA module is a small module formed by two consecutive Hoogsteen/sugar edge base pairs, A-G and G-A<sup>18</sup>. They are frequently observed within regular helices. We considered tandem GAs with four stacked base



**Figure 1** | Details of the analyzed RNA structural modules. (a) The G-bulge from the lysine riboswitch (Protein Data Bank (PDB) code 3DIG)<sup>20</sup>. (b) The kink-turn from the helix 46 of the bacterial large subunit (PDB code 2WRJ)<sup>40</sup>. (c) The C-loop from the helix 38 of the bacterial large subunit (PDB code 2WRJ)<sup>40</sup>. (d) A tandem GA from a synthetic RNA octamer (PDB code 1SA9)<sup>41</sup>. For each module, a detailed structure (center), the position in the original molecule (left) and the interaction network (right) are shown. The underlined bases in the interaction network correspond to the nucleotides present in the crystal structure. Numbers next to the bases indicate the observed percentages of each nucleotide in the alignment.

**Figure 2** | Steps of single- and multiple-sequence search algorithms. **(a)** In step 1 of the single-sequence search algorithm, first the Bayesian network model is applied to the target sequence to obtain potential candidates and their respective scores. Step 2 is to fold the target sequence and compute the proportion of the ensemble (set of all possible folds for that sequence) compatible with the candidates found in the previous step. This proportion is referred to as base-pair probability (BPP). In step 3, candidates are filtered using predefined score and BPP thresholds. **(b)** In the multiple sequence search algorithm used with multiple homologous sequences, step 4 is to apply the previous algorithm to each individual sequence of the target alignment to obtain the candidates for all sequences (seq1–seq5). Step 5 is to represent each candidate, in a matrix, using the starting column of the candidate strands in the alignment as coordinates. Cluster the candidates according to their location in the matrix and compute the frequency of each cluster in the alignment (occur). Overlapping candidates are discarded. Step 6 is to compute the average mutual information (MI) of each cluster as a measure of variation between positions. The MI of the cluster is the mean of the individual MI of expected Watson-Crick base pair positions, and it is normalized by the maximum possible MI (2 bits per base pair). In step 7, heuristic rules are used to filter candidates based on score, BPP, occur, count and MI values.



pairs: Watson-Crick, G-A, A-G and Watson-Crick. The module contains two sequences of four nucleotides, NGAN (in which N is any nucleotide), which can occur, by chance, once in each 16 random bases. This short sequence makes it difficult to distinguish tandem GAs from background. However, the conservation of the GA nucleotides across homologous sequences is usually distinguishable in sequence alignments (Fig. 1d).

### Structural modules as Bayesian networks

The direct use of nucleotide distributions, observed in the interaction networks, to search for modules in sequences, presents the limitation of assuming statistical independence between the positions of the module. This independence is generally not verified. For example, Watson-Crick base pairs present a strong correlation between the bases. Sometimes, the same base pair can adopt more than one interaction type depending on the particular instance of a module, which imposes a dependency between bases even if, in some of the instances, one of the bases is fixed. Such a situation occurs in the kink-turn module in which the first base pair of the noncanonical stem usually adopts a Hoogsteen/Sugar edge interaction with an invariant A, but it can also adopt a Watson-Crick interaction, which imposes the corresponding isostericity constraints. Other dependencies can also occur between edge-interacting or stacking-interacting nucleotides. A way to overcome this limitation is to interpret an interaction network as a Bayesian network and explicitly model all the dependencies between the bases of the module observed in systematic structural alignments.

Bayesian networks are probabilistic models in which random variables and the dependency between them are represented as an acyclic directed graph. The nodes of the graph correspond to the random variables and the edges to the dependencies. Bayesian networks have been applied to sequence-analysis problems, for

example, for detection of transcription factors<sup>25</sup>. For modeling RNA modules as Bayesian networks, the nodes represent individual bases occupying a defined structural position, and the edges represent the dependencies between them.

### Single sequence search

When searching for structural modules in single sequences, RMDetect computes, for all subsequences, the log-likelihood score corresponding to the likelihood that the given subsequence was generated by the Bayesian network of the module. Owing to the small size of Bayesian networks and the short (four-letter) alphabet of nucleotides, this scan will normally produce a large number of medium to high score hits, many of them false positives. To reduce the number of false positives RMDetect uses the predicted joint base-pair probability of the module's Watson-Crick base pairs to select the subsequences for which a compatible secondary structure is likely to be observed (Fig. 2a).

To evaluate RMDetect for single sequence search we built 15 test cases, corresponding to the molecules in which the modules had been identified in crystal structures and for which we obtained reliable sequence alignments (Supplementary Table 1 and Online Methods). We obtained Matthews correlation coefficient values of individual test cases<sup>26</sup>, with fixed parameters, which varied between 0.93 for the kink-turn model and 0.13 for the tandem-GA model. We calculated the true positive rates to be above 0.5 for all but the tandem-GA module, indicating that RMDetect consistently found more than half of the positive candidates. However, false discovery rates higher than 0.5 for the three tandem-GAs confirmed the difficulty in discarding false positive candidates for small modules with few non-Watson-Crick interactions using single sequence information (Table 1). The diversity of the training set should be as complete as possible to obtain a representative model of a module. For example, a G-bulge model trained only

**Table 1** | RMDetect analysis of the single-sequence test set

Searched module	Molecule <sup>a</sup>	Instances <sup>b</sup>	Best MCC <sup>c</sup>					Fixed parameters <sup>d</sup>				
			MCC	TPR <sup>e</sup>	FDR <sup>f</sup>	Score <sup>g</sup>	BPP <sup>h</sup>	MCC	TPR <sup>e</sup>	FDR <sup>f</sup>	Score <sup>g</sup>	BPP <sup>h</sup>
G-bulge	16S rRNA bacteria	2 × 250	0.73	0.70	0.22	12.3	0.001	0.58	0.71	0.53	8.0	0.001
G-bulge	23S rRNA archaea	6 × 100	0.68	0.55	0.15	13.8	0.001	0.61	0.64	0.42		
G-bulge	23S rRNA bacteria	5 × 250	0.71	0.67	0.24	11.5	0.001	0.63	0.71	0.44		
G-bulge	Lysine riboswitch	1 × 150	0.66	0.48	0.09	8.3	0.010	0.64	0.51	0.20		
Kink-turn	16S rRNA bacteria	1 × 250	0.97	0.96	0.01	17.2	0.041	0.67	0.97	0.53	11.0	0.001
Kink-turn	23S rRNA archaea	5 × 100	0.70	0.61	0.20	14.5	0.001	0.64	0.67	0.40		
Kink-turn	23S rRNA bacteria	4 × 250	0.67	0.52	0.16	15.7	0.001	0.59	0.65	0.48		
Kink-turn	SAM riboswitch <sup>i</sup>	1 × 150	0.93	0.93	0.07	8.7	0.001	0.93	0.91	0.06		
Kink-turn	U4 snRNA	1 × 500	0.71	0.54	0.06	12.3	0.001	0.70	0.55	0.10		
C-loop	16S rRNA bacteria	1 × 250	0.84	0.85	0.16	18.5	0.011	0.80	0.91	0.29	16.0	0.010
C-loop	23S rRNA archaea	3 × 100	0.66	0.50	0.11	22.4	0.001	0.48	0.54	0.57		
C-loop	23S rRNA bacteria	3 × 250	0.62	0.56	0.32	15.9	0.021	0.60	0.58	0.38		
Tandem GA	16S rRNA bacteria	1 × 250	0.42	0.66	0.74	9.6	0.161	0.36	0.67	0.81	9.0	0.100
Tandem GA	23S rRNA archaea	1 × 100	0.41	0.21	0.19	9.9	0.990	0.13	0.33	0.95		
Tandem GA	23S rRNA bacteria	2 × 250	0.53	0.67	0.58	9.5	0.530	0.39	0.82	0.82		

<sup>a</sup>Sequence alignments searched. <sup>b</sup>Number of (module) instances present in the alignment; module instances present in each sequence times the number of sequences. <sup>c</sup>Sensitivity and specificity analysis for the parameter that maximize the Matthews correlation coefficient (MCC). <sup>d</sup>Sensitivity and specificity analysis for fixed score and bpp for all test sets of the same module. <sup>e</sup>True positive rate (TPR) = TP / (TP + FN). <sup>f</sup>False discovery rate (FDR) = FP / (TP + FP). <sup>g</sup>Threshold score. <sup>h</sup>Threshold BPP values used to discriminate candidates. <sup>i</sup>SAM, S-adenosylmethionine.

with 16S rRNA and 23S rRNA did not identify most of the lysine riboswitch G-bulge instances (**Supplementary Note 2**).

### Multiple sequence alignment search

The increasing availability of databases of homologous, or related, sequences for many RNA molecules<sup>27</sup> and the existence of effective RNA sequence alignment tools for close sequences<sup>28</sup> provides powerful sources of information for module discovery. When searching for modules in aligned RNA sequences, even if the positions where the modules occur are misaligned, we expect that the true positive candidates would be located in columns relatively close to each other. When sequences are sufficiently divergent, which is the case of many RNA sequences, false positive candidates should be distributed across the alignment. Based on these assumptions, we devised a clustering strategy to exploit multiple sequence alignment information for module searching. We clustered candidates according to the distance on the column space of the alignment and selected the most represented clusters, with higher score candidates and covariation signals between bases of Watson-Crick base pairs, as potential hits (**Fig. 2b**).

To test RMDetect on multiple sequence alignments, we applied it to the same 15 datasets of the single sequence search. RMDetect correctly found all of the 37 known module instances

(true positive rate of 1) with 11 false positive candidates (false discovery rate of 0.23), five of them falsely identified as tandem-GA modules. These results show that RMDetect is effectively improved by adding alignment information (**Table 2** and **Supplementary Data 1**).

### Search in public databases

We applied RMDetect to multiple sequence alignments from the RFam database, the group I intron database<sup>29</sup> and new bacterial ncRNAs reported in references 30 and 31 (**Supplementary Data 2**). Using the same selection conditions as in previous tests, we selected 250 candidates. From those, 21 predictions correspond to presently unreported modules, 141 correspond to previously predicted or observed modules, and the remaining 88 were unconfirmed candidates (**Table 3** and **Supplementary Data 3**).

### Rfam results

Searching 1,309 Rfam alignments resulted in 222 module candidates, 132 of which were known modules and 77 of which corresponded to unconfirmed candidates. Not surprisingly, 99 of the known candidates corresponded to kink-turns in the snoRNAs C/D or C/D' boxes. We found 13 previously undetected modules, including one kink-turn, one G-bulge, three C-loop and eight tandem GA modules (**Supplementary Fig. 1**).

**Table 2** | RMDetect analysis of the multiple-sequence test set

Alignment searched	G-bulge		Kink-turn		C-loop		Tandem GA	
	TP (TPR) <sup>a</sup>	FP (FDR) <sup>b</sup>	TP (TPR)	FP (FDR)	TP (TPR)	FP (FDR)	TP (TPR)	FP (FDR)
16S_P	2 (1.00)	0 (0.00)	1 (1.00)	2 (0.66)	1 (1.00)	0 (0%)	1 (1.00)	2 (0.66)
23S_A	6 (1.00)	0 (0.00)	5 (1.00)	1 (0.17)	3 (1.00)	3 (50%)	1 (1.00)	0 (0.00)
23S_P	5 (1.00)	0 (0.00)	4 (1.00)	0 (0.00)	3 (1.00)	0 (0%)	2 (1.00)	3 (0.60)
SamRS	–	–	1 (1.00)	0 (0.00)	–	–	–	–
LysRS	1 (1.00)	0 (0.00)	–	–	–	–	–	–
U4 snRNA	–	–	1 (1.00)	0 (0.00)	–	–	–	–
Total	14 (1.00)	0 (0.00)	12 (1.00)	3 (0.20)	7 (1.00)	3 (0.30)	4 (1.00)	5 (0.55)

–, not applicable.

<sup>a</sup>Number of true positives (TP) and true positive rate (TPR = TP / (TP + FN)). <sup>b</sup>Number of false positives (FP) and false discovery rate (FDR = FP / (TP + FP)).

**Table 3** | RMDetect analysis of public database alignments

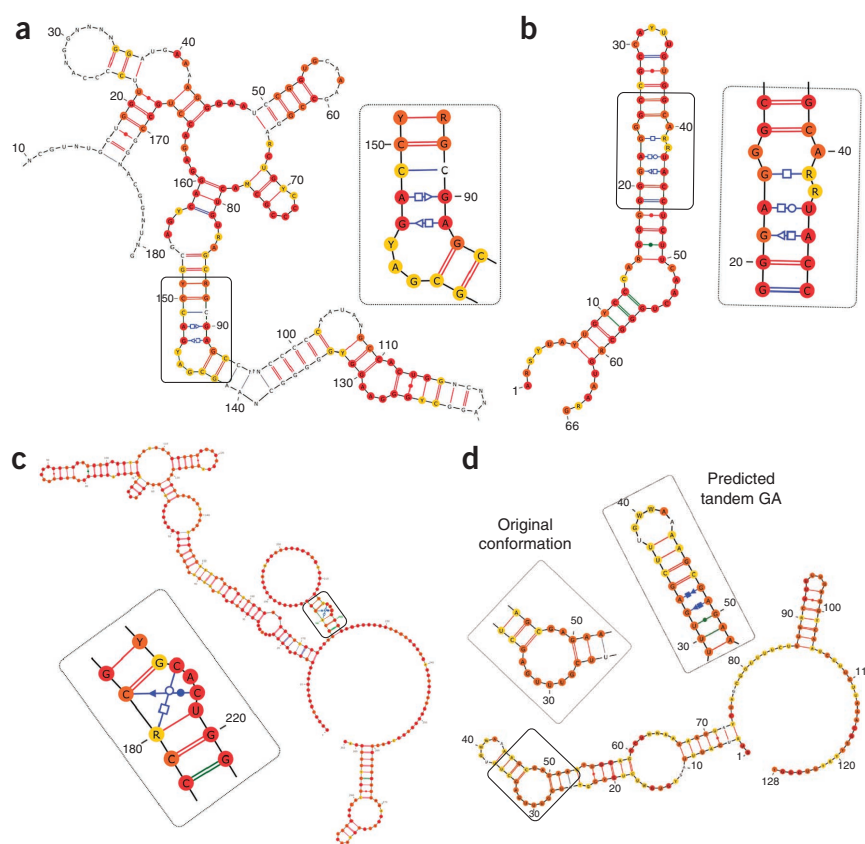
Database (alignments searched)		G-bulge	Kink-turn	C-loop	Tandem GA
Rfam (1,309 alignments)	Total selected candidates	13	119	22	68
	Known modules <sup>a</sup>	6	105 (99 snoRNAs)	0	21 (20 kink-turns)
	New candidates <sup>b</sup>	1	1	3	8
	Not confirmed <sup>c</sup>	6	13	19	39
Group I introns (14 alignments)	Total selected candidates	1	1	0	1
	Known modules <sup>a</sup>	1	0	0	0
	New candidates <sup>b</sup>	0	0	0	1
	Not confirmed <sup>c</sup>	0	1	0	0
Bacterial ncRNAs (121 alignments)	Total selected candidates	4	4	1	16
	Known modules <sup>a</sup>	3	0	0	5 (1 kink-turn)
	New candidates <sup>b</sup>	1	1	0	5
	Not confirmed <sup>c</sup>	0	3	1	6

<sup>a</sup>Number of selected candidates corresponding to known modules. <sup>b</sup>Number of selected candidates corresponding to new putative modules. <sup>c</sup>Number of false positive candidates or candidates for which no confirmation was possible.

We detected the newly predicted kink-turn in 353 (16%) sequences in the variable region of the cobalamin riboswitch alignment<sup>32</sup>. With realignment, using the predicted kink-turn as an anchor, we established the full conservation of the tandem-GA sequences as well as the perfect pairing of at least two Watson-Crick base pairs in both helical stems with strong covariation (**Fig. 3a**). Another strong candidate was a G-bulge found in 109 (11%) sequences of the Hepatitis C virus stem-loop SL-VII<sup>33</sup>. Unlike the cobalamin riboswitch candidate, this G-bulge is conserved, correctly aligned and stands out in the secondary structure derived from the full alignment (**Fig. 3b**). Although alternative folding is possible, in which the G-bulge region participates in a helix interrupted by two bulged adenines, the conservation of the AGUA-GA sequences and the covariation of the base pairs in the hairpin support the prediction of a G-bulge. We detected three potential C-loops in the *c-myc* internal ribosome entry site (IRES)<sup>34</sup> (**Fig. 3c**), enterovirus *cis*-acting replication element (CRE)<sup>35</sup> and QUAD bacterial ncRNA<sup>36</sup> in 37 (45%), 112 (54%) and 174 (49%) sequences respectively. In the first case, we found the candidate in a region flanking a pseudoknot in the structure of the IRES. The covariation of the helices and the conservation of the characteristic 'CAC' motif support the prediction. In the cases of the enterovirus CRE and QUAD RNA the candidates stand out from the originally proposed secondary structure with no rearrangement needed.

We detected a tandem GA in 157 sequences (40%) of the *rtT* alignment, a bacterial ncRNA observed as a transcription product of the *tyrT* operon of *Escherichia coli*<sup>37</sup>. The detected module suggests a rearrangement of one internal loop of the originally proposed structure. It is possible that the module is not present in all sequences. We detected a second tandem GA in 16 sequences (47%) of the 5' untranslated region of the voltage-gated potassium channel mRNA where the proposed secondary structure

suggests the detected module. A tandem GA, predicted in 20 (71%) sequences in the *purD* alignment<sup>38</sup>, stands at an internal loop compatible with a rare type of kink-turn with four nucleotides in the bulge (**Fig. 3d**). One can rearrange the secondary structure to obtain the minimal tandem GA maintaining the covariation, but we cannot discard the possibility of a more complex module. Notably, 21 of the identified tandem GAs correspond to kink-turns. This is not surprising because kink-turns contain a tandem GA and that the first base of the bulge can often be predicted as forming a base pair with the base in the opposite strand.



**Figure 3** | Examples of the newly predicted modules. (a) Kink-turn in the Cobalamin riboswitch. (b) G-bulge in the SL VII domain of the Hepatitis C virus. (c) C-loop in the internal ribosomal entry site from the C-myc mRNA. (d) Tandem GA in the *purD* bacterial RNA (original conformation as in ref. 38).

RMDetect allowed the correct detection of the three modules (two G-bulges and one kink-turn) in the T-box riboswitch<sup>22</sup>, the two modules (one G-bulge and one kink-turn) in the lysine riboswitch<sup>20</sup> and the G-bulge module in the IRES of the Hepatitis C virus<sup>39</sup>.

### Group I intron results

Searching the 14 alignments of the group I intron database, we detected the known G-bulge module present in the P7 domain of type IA2 introns, confirmed by the crystal structure of the phage Twort intron<sup>21</sup>. Additionally we detected a tandem GA in 12 (38%) sequences of type IC2 intron. This candidate was predicted in P5d domain (**Supplementary Fig. 2**).

### Bacterial ncRNA results

Several modules had been originally identified on 121 alignments of structured ncRNAs from recently published metagenomic data<sup>30,31</sup>. We applied our algorithm to all of these alignments. We found a new kink-turn in the GEMM-II alignment, a new G-bulge in group-II-D1D4-1 molecule and five new tandem-GA modules (**Supplementary Fig. 2**). The twoAYGGAY motif<sup>31</sup> bacterial RNA alignment is an interesting case: we detected two tandem GAs in the same hairpin stem, a distal one (10 base pairs (bp) from the loop) in 80 sequences (39%), and a proximal one (2 bp from the loop) in 28 sequences (14%). All combinations of one and both tandem GAs can be found in the alignment (**Supplementary Fig. 3**). Homologous sequences that do not contain the module instead have Watson-Crick base pairs at the positions corresponding to the module. This alignment raises interesting questions about how structural modules evolve and interchange and how this will affect the final 3D geometry of the molecule. However, RMDetect missed four of seven previously reported G-bulge in the dataset. One was that of GOLLD that differs slightly from the defined G-bulge model, putting it outside the scope of the Bayesian network model. Two others correspond to Dictyoglomi-1 G-bulges that spanned more than the sliding window length (150 nucleotides). We detected the missing modules by reapplying the algorithm with no window length limitation. We discarded the final Dictyoglomi-1 G-bulge despite a high score (17.0) and occurrence (75%) owing to the small alignment (4 sequences) and total conservation in the module region (mutual information score of 0.0). This observation highlights the fact that the RMDetect parameters, although necessary owing to the large number of searched alignments, will not guarantee the exhaustive search of sequence space. When searching a small number of alignments, different window lengths and steps together with more relaxed selection criteria should be applied.

### DISCUSSION

In the single-sequence test set we detected more than half of the searched modules in molecules as complex as the ribosome. In multiple sequence alignment test sets, we identified all known modules with an overall false discovery rate of 0.23. We extended the search to 1,444 publicly available alignments used without realignment. We found most of the known modules in all major classes of structured ncRNAs and identified 21 new candidates. With the RMBuild tool (**Supplementary Note 3**), our approach can be extended to additional modules and newly discovered ones. The Bayesian network models can be further improved with new instances of known modules.

RMDetect is available as **Supplementary Software** and at <http://sourceforge.net/projects/rmdetect/>.

### METHODS

Methods and any associated references are available in the online version of the paper at <http://www.nature.com/naturemethods/>.

*Note: Supplementary information is available on the Nature Methods website.*

### ACKNOWLEDGMENTS

We thank R. Backofen for useful suggestions. J.A.C. is supported by the Ph.D. Program in Computational Biology of the Instituto Gulbenkian de Ciência, Portugal (sponsored by Fundação Calouste Gulbenkian, Siemens SA and Fundação para a Ciência e Tecnologia; SFRH/BD/33528/2008).

### AUTHOR CONTRIBUTIONS

J.A.C. conceived the algorithms, performed the computations and wrote the manuscript. E.W. conceived the research and wrote the manuscript.

### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at <http://www.nature.com/naturemethods/>.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

- Leontis, N.B., Stombaugh, J. & Westhof, E. The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.* **30**, 3497–3531 (2002).
- Lescoute, A. *et al.* Recurrent structural RNA motifs, Isostericity Matrices and sequence alignments. *Nucleic Acids Res.* **33**, 2395–2409 (2005).
- Leontis, N.B. & Westhof, E. Analysis of RNA motifs. *Curr. Opin. Struct. Biol.* **13**, 300–308 (2003).
- Leontis, N.B. & Westhof, E. A common motif organizes the structure of multi-helix loops in 16 S and 23 S ribosomal RNAs. *J. Mol. Biol.* **283**, 571–583 (1998).
- Macke, T.J. *et al.* RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.* **29**, 4724–4735 (2001).
- Yao, Z., Weinberg, Z. & Ruzzo, W.L. CMfinder—a covariance model based RNA motif finding algorithm. *Bioinformatics* **22**, 445–452 (2006).
- Parisien, M. & Major, F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **452**, 51–55 (2008).
- Jossinet, F., Ludwig, T.E. & Westhof, E. Assemble: an interactive graphical tool to analyze and build RNA architectures at the 2D and 3D levels. *Bioinformatics* **26**, 2057–2059 (2010).
- Das, R., Karanicolas, J. & Baker, D. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Methods* **7**, 291–294 (2010).
- Westhof, E. The amazing world of bacterial structured RNAs. *Genome Biol.* **11**, 108 (2010).
- Moore, P.B. Structural Motifs in RNA. *Annu. Rev. Biochem.* **68**, 287–300 (1999).
- Klein, D.J. *et al.* The kink-turn: a new RNA secondary structure motif. *EMBO J.* **20**, 4214–4221 (2001).
- Djelloul, M. & Denise, A. Automated motif extraction and classification in RNA tertiary structures. *RNA* **14**, 2489–2497 (2008).
- Sarver, M. *et al.* FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *J. Math. Biol.* **56**, 215–252 (2008).
- Apostolico, A. *et al.* Finding 3D motifs in ribosomal RNA structures. *Nucleic Acids Res.* **37**, e29 (2009).
- Zhong, C., Tang, H. & Zhang, S. RNAMotifScan: automatic identification of RNA structural motifs using secondary structural alignment. *Nucleic Acids Res.* **38**, e176 (2010).
- Thébault, P. *et al.* Searching RNA motifs and their intermolecular contacts with constraint networks. *Bioinformatics* **22**, 2074–2080 (2006).
- Gautheret, D., Konings, D. & Gutell, R.R. A major family of motifs involving G? A mismatches in ribosomal RNA. *J. Mol. Biol.* **242**, 1–8 (1994).
- Leontis, N.B., Stombaugh, J. & Westhof, E. Motif prediction in ribosomal RNAs Lessons and prospects for automated motif prediction in homologous RNA molecules. *Biochimie* **84**, 961–973 (2002).

20. Serganov, A., Huang, L. & Patel, D.J. Structural insights into amino acid binding and gene control by a lysine riboswitch. *Nature* **455**, 1263–1267 (2008).
21. Golden, B.L., Kim, H. & Chase, E. Crystal structure of a phage Twort group I ribozyme-product complex. *Nat. Struct. Mol. Biol.* **12**, 82–89 (2005).
22. Wang, J., Henkin, T.M. & Nikonowicz, E.P. NMR structure and dynamics of the specifier loop domain from the *Bacillus subtilis* tyrS T box leader RNA. *Nucleic Acids Res.* **38**, 3388–3398 (2010).
23. Kuhn, J.F., Tran, E.J. & Maxwell, E.S. Archaeal ribosomal protein L7 is a functional homolog of the eukaryotic 15.5kD/Snu13p snoRNP core protein. *Nucleic Acids Res.* **30**, 931–941 (2002).
24. Sankaranarayanan, R. *et al.* The structure of threonyl-tRNA synthetase-tRNA(Thr) complex enlightens its repressor activity and reveals an essential zinc ion in the active site. *Cell* **97**, 371–381 (1999).
25. Ben-Gal, I. *et al.* Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics* **21**, 2657–2666 (2005).
26. Matthews, B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405**, 442–451 (1975).
27. Gardner, P.P. *et al.* Rfam: updates to the RNA families database. *Nucleic Acids Res.* **37**, D136–D140 (2009).
28. Wilm, A., Mainz, I. & Steger, G. An enhanced RNA alignment benchmark for sequence alignment programs. *Algorithms Mol. Biol.* **1**, 19 (2006).
29. Zhou, Y. *et al.* GISSD: Group I Intron Sequence and Structure Database. *Nucleic Acids Res.* **36**, D31–D37 (2008).
30. Weinberg, Z. *et al.* Exceptional structured noncoding RNAs revealed by bacterial metagenome analysis. *Nature* **462**, 656–659 (2009).
31. Weinberg, Z. *et al.* Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biol.* **11**, R31 (2010).
32. Vitreschak, A.G. Regulation of the vitamin B12 metabolism and transport in bacteria by a conserved RNA structural element. *RNA* **9**, 1084–1097 (2003).
33. Lee, H. *et al.* cis-acting RNA signals in the NS5B C-terminal coding sequence of the Hepatitis C virus genome. *J. Virol.* **78**, 10865–10877 (2004).
34. Le Quesne, J.P. *et al.* Derivation of a structural model for the c-myc IRES. *J. Mol. Biol.* **310**, 111–126 (2001).
35. Paul, A.V. *et al.* Identification of an RNA hairpin in poliovirus RNA that serves as the primary template in the *in vitro* uridylylation of VPg. *J. Virology* **74**, 10359–10370 (2000).
36. Wassarman, K.M. *et al.* Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev.* 1637–1651 (2001).
37. Bösl, M. & Kersten, H. A novel RNA product of the tyrT operon of *Escherichia coli*. *Nucleic Acids Res.* **19**, 5863–5870 (1991).
38. Weinberg, Z. *et al.* Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. *Nucleic Acids Res.* **35**, 4809–4819 (2007).
39. Klinck, R. *et al.* A potential RNA drug target in the hepatitis C virus internal ribosomal entry site. *RNA* **6**, 1423–1431 (2000).
40. Gao, Y.-G. *et al.* The structure of the ribosome with elongation factor G trapped in the posttranslocational state. *Science* **326**, 694–699 (2009).
41. Jang, S.B. *et al.* Structures of two RNA octamers containing tandem G.A base pairs. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 829–835 (2004).



## ONLINE METHODS

**Data sources.** All crystal structures were obtained from the Protein Data Bank (PDB)<sup>42</sup>. The alignments of the bacterial ribosome (both subunits) are from ref. 2; archaeal ribosomal large subunit data were obtained both from ref. 2 and the comprehensive ribosomal RNA databases (Silva) version 102 (ref. 43), the later was manually corrected in the regions corresponding to the studied modules; Rfam alignments correspond to version Rfam 9.1 and were downloaded from <http://rfam.sanger.ac.uk/><sup>27</sup>, group I intron alignments were downloaded from group I intron sequence and structure database (GISSD)<sup>29</sup>; and the new bacterial RNA alignments are from the supplementary information in references 30,31. All 2D diagrams were produced using the visualization applet for RNA secondary structure (VARNA)<sup>44</sup>.

**Design of the Bayesian networks.** The design of a Bayesian network is a two-step process. First, the network topology is established, that is, the set of dependencies between all variables of the model is determined. Second, the parameters describing the probability distribution of each node based on the observed data and specified dependencies are computed.

In the present case the Bayesian network topology closely follows the established interaction networks. All Watson-Crick (WC) base pairs, most of the non-WC base pairs and some base stacking interactions will map to edges of the Bayesian network. Interactions involving a fully conserved nucleotide were not included because they would not add any information to the Bayesian network. Some additional edges were included that connect structurally important but less conserved bases to bulged bases (**Supplementary Fig. 4**).

A multinomial distribution, corresponding to the occurrence probability of the four nucleotides and a gap, is associated to each node of the Bayesian network. In the case of dependent nodes, the local distribution is conditioned by the parent nodes distributions. The parameters were estimated by maximum likelihood using the sequence alignments of each module as the observations (**Supplementary Fig. 5**). Given the high number of observations (5,735 observations for G-bulge, 7,677 observations for kink-turns and 3,545 observations for C-loops) the parameters correspond to the relative frequency of each nucleotide in the full sample<sup>45</sup>. As the different alignments have different sequence frequencies, counts were normalized so that all alignments would contribute equally to the final count. The tandem-GA module was an exception to the above method as the parameters were not computed from sequence alignment data but were defined based on an ideal tandem GA. For this module, the first base of each WC base pair has a distribution identical to the nucleotide content of the sequence, and the second base had a conditional probability of  $P(U|A) = P(G|C) = 1.0$ ;  $P(A|U) = P(C|G) = 0.6$ ; and  $P(G|U) = P(U|G) = 0.4$ . The two non-WC base pairs were invariant with probabilities  $P1(A) = P2(G) = 1.0$ .

Finally, each WC base pair was classified as mandatory or optional. This information is used to compute the base pair probabilities and mutual information when filtering candidates.

**Interaction networks analysis for parameter estimation.** We analyzed 11 G-bulge modules. Computed nucleotide frequencies confirmed previous predictions by isostericity analysis<sup>4</sup>. Four base pairs were invariant in all occurrences of the module: the bulged

G-U *cis* Hoogsteen/sugar edge, the A-G *trans* Hoogsteen/sugar edge, the U-A *trans* WC/Hoogsteen and the A-A *trans* Hoogsteen/Hoogsteen. In the remaining positions some variation was allowed (**Supplementary Fig. 6**).

We analyzed 14 kink-turn instances and grouped them into families based on their interaction network similarities (**Supplementary Fig. 7**). To obtain a consensus interaction network, we excluded the instances KT-16S-11-P, the interaction network of which is unique and too divergent from all other families; and KT-23S-15-A and KT-23S-58-A, which present atypical nucleotide insertions in the short strand (in the abbreviations, KT is kink turn; 23S or 16S indicate large or small ribosomal subunit; 11, 15 or 58 are helix 11, 15 or 58; and A or P indicate archaeal or bacterial rRNA alignment).

Seven analyzed C-loops revealed an invariant core formed by two crossing, noncanonical interactions pairing the first and last bases of the loop with the bases of the flanking base pairs in the opposite strand. Despite this interaction regularity the C-loop presents big sequence variation, except for the first base of the loop, invariably a C, and the third base of the loop, either a A or a C with the same frequency (**Supplementary Fig. 8**).

**Single sequence search algorithm.** A formal definition of the single sequence search algorithm used by RMDetect (**Fig. 2**) can be stated as: let  $M$  be a structural RNA module;  $S$  be a nucleotide sequence to be searched for  $M$ ;  $M_{BN}$  a Bayesian network model of  $M$ ;  $M_{GC}$  a null model in which all the bases are independent and have the same nucleotide distribution of  $S$ ;  $sp_{ij} = \{seq_p, seq_j\}$  a pair of non-overlapping subsequences of  $S$  starting from positions  $i$  and  $j$ , corresponding to the strands of the module; and  $WC_M$  be the set of all WC base pairs from  $M$ . For simplicity, we will describe only modules formed by pairs of subsequences, that is, modules with two strands. The extension to modules formed by more than two strands, as in  $n$ -way junctions<sup>2</sup>, would simply require redefinition of  $sp$  as a tuple  $sp_{i_1, \dots, i_n} = (seq_{i_1}, \dots, seq_{i_n})$ .

For each  $sp_{ij}$  compute the corresponding score <sub>$ij$</sub> :

$$\text{score}_{ij} = \log_2 \left( \frac{P(sp_{ij} | M_{BN})}{P(sp_{ij} | M_{GC})} \right);$$

For each  $sp_{ij}$  compute  $\text{bpp}_{ij}$ , the corresponding joint base pair probability of all WC base pairs:

$$\text{BPP}_{ij} = \frac{e^{-\frac{\text{Ens. FE}_{ij}}{kT}}}{e^{-\frac{\text{Ens. FE}_{\text{all}}}{kT}}},$$

in which Ens. FE stands for the free energy of a folding ensemble, Ens. FE<sub>all</sub> corresponds to the folding of the unconstrained original sequence, and Ens. FE <sub>$ij$</sub>  corresponds to the folding of the original sequence constrained by the base pairs of ( $WC_M$ ) in the positions determined by  $sp_{ij}$ .

Select all  $sp_{ij}$  with score <sub>$ij$</sub>  and  $\text{bpp}_{ij}$  higher than a given threshold. These will be considered the candidates for the module considered.

The single sequence search was performed with a window length of 150 nt and a window step of 75 nt. All candidates scoring less than the specified score and BPP values (**Table 1**) were



discarded, and the remaining ones were retained as candidates. We considered a candidate as true positive (TP) if it occurred in the same sequence positions as the known module instances (plus or minus two positions to account for unexpected gaps and alignment errors) all the other candidates are considered false positives (FP). The free energies of the ensembles were computed with 'RNAfold -p'<sup>46</sup> (-p to calculate the partition function) to obtain Ens. FE<sub>all</sub> (the unconstrained FE) and RNAfold -p -C' (-C to calculate structures subject to constraints) to obtain Ens. FE<sub>motif</sub> (the constrained FE). The parameters used to compute the joint base pair probabilities where  $T = 274.5\text{K}$  and  $k = 1.98717 \times 10^{-3} \text{ kcal mol}^{-1}$  (from Vienna package source code). The algorithm performance scaled linearly with sequence length (for a fixed window length) and scaled quadratically with window length (Supplementary Notes 4 and 5 and Supplementary Figs. 9 and 10).

**Multiple sequence search algorithm.** As seen in the single sequence search algorithm, each module candidate can be defined by an ordered pair of alignment coordinates  $\text{cand}_{ij} = (\text{seq}_i, \text{seq}_j)$ . A hierarchical clustering algorithm was applied to group the candidates of the different sequences according to their distance. The algorithm is as follows. (i) Remove all overlapping candidates on the same sequence retaining only the one with the higher score. (ii) Each candidate,  $\text{cand}_{ij}$ , will be assigned to the cluster,  $\text{cluster}_{ij}$ , centered at the position  $(i, j)$ . (iii) Merge all pairs of clusters for which  $\text{dist}(\text{cluster}_{ij}, \text{cluster}_{kl}) < \text{DLIMIT}$ , where  $\text{dist}(\text{cluster}_{ij}, \text{cluster}_{kl}) = \max(|i - k|, |j - l|)$ . Notice that  $i, j, k$  and  $l$  are columns of the alignment and DLIMIT is the maximum tolerated column distance between two candidates so that they can be considered to belong to the same cluster. (iv) Recompute the center of each cluster as the most represented position  $(i, j)$ . (v) Repeat from (iii) until no more clusters are merged.

At the end, each cluster will correspond to a module candidate characterized by five measures: (i) absolute number of the aligned sequences in which the candidate occurs (sequence count); (ii) percentage of aligned sequences in which the candidate occurs (occurrence); (iii) mean score of all candidates; (iv) mean BPP of all candidates; and (v) mutual information (MI) between the bases of each WC base pair from  $\text{WC}_{M^2}$ , measured as along all candidates<sup>47</sup> (Supplementary Note 6). Thus, for a cluster to be considered it must be sufficiently represented in the alignment, must have a score and BPP higher than the defined threshold and should have covariance between WC base pairs, supporting the evolutionary pressure on conservation of the secondary structure of the module (Fig. 2b).

The multiple sequence search algorithm, described above, produced a set of clusters that was filtered according to the following conditions: (i) (sequence count > 2) and (occurrence  $\geq$  10%); (ii) MI > 0 or (occurrence > 33% and sequence\_count > 10); (iii) score  $\geq$  limit\_score; and (iii) BPP  $\geq$  limit\_BPP.

Both limit\_score and limit\_BPP vary across the models. Limit\_score was 8.0 for G-bulge, 11.0 for kink-turn, 16.0 for C-loop and 9.0 for tandem-GA. limit\_bpp is 0.1 for the tandem GA, 0.01 for the C-loop and 0.001 for all other models. These values were chosen as they allowed the detection of at least half of the modules in all but one single sequence search test case (Table 2). The DLIMIT distance, discussed above, was set to five columns.

At the end of this process each selected cluster corresponds to a module prediction that was manually validated according to the compatibility with published structure, sequence alignment or co-variation information obtained from the alignment.

**Test cases for known modules.** Fifteen test cases were generated each corresponding to one module and one alignment (Supplementary Table 1). For each test case, the original alignment was randomly split in one training set and one test set. The training set was used to compute model parameters, and the test set was used for the search. The training set was then augmented with sequences from the other alignments containing the searched module. As a negative control, each sequence of the test set was duplicated and shuffled to preserve the nucleotide composition of the sequence. Single sequence and multiple sequence search algorithms were performed in each test set as described above. For example, when searching for the G-bulge module in the 16S rRNA sequences, the training set was composed by 523 randomly selected sequences of the 16S rRNA alignment plus all 6,956 sequences from 23S bacterial rRNA, 23S archaea rRNA and lysine-riboswitch alignments. The test set included the remaining 250 sequences of the 16S rRNA alignment plus 250 shuffled sequences. Both algorithms were applied as described above.

**Search in database alignments.** We systematically searched 1,309 Rfam families, 14 group I intron alignments and 121 alignments of structured ncRNAs from meta-genomic data<sup>30,31</sup>. The Rfam alignments with more than 7,000 sequences were reduced to shorter versions containing 500 randomly selected sequences from the original alignment. The group I intron alignments were converted to Stockholm format. All alignments were searched 'as-is' with no realignment or manual adjustments. The following alignments were excluded from the search: the U4 snRNA, all small and large subunit rRNAs (4 families) and the SAM riboswitch that were used for training; the group I intron alignment that were searched in specific databases; the tRNA family; And all the families with less than five sequences (56 families).

**Implementation and software availability.** The described algorithms were implemented as a set of python scripts publicly available as open source from: <http://sourceforge.net/projects/rmdetect/>. A user guide is provided (Supplementary Note 3).

42. Berman, H.M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
43. Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **35**, 7188–7196 (2007).
44. Darty, K., Denise, A. & Ponty, Y. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25**, 1974–1975 (2009).
45. Durbin, R. *et al.* *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge University Press, 1998).
46. Hofacker, I.L. *et al.* Fast folding and comparison of RNA secondary structures. *Monatshfte für Chemie* **125**, 167–188 (1994).
47. Lindgreen, S., Gardner, P.P. & Krogh, A. Measuring covariation in RNA alignments: physical realism improves information measures. *Bioinformatics* **22**, 2988–2995 (2006).

## Chapter 7

# Conclusions and Perspectives

In this thesis I applied computer science and bioinformatics techniques to approach three open problems on ncRNA studies: (i) how to meaningfully compare three-dimensional RNA models; (ii) how to annotate complete eukaryotic genomes for ncRNA in an accurate and as automatic as possible fashion; and (iii) how to detect three-dimensional structural RNA modules in sequences using no additional information.

(i) With the group of François Major in Montreal, we developed two new structural comparison metrics that take into account the structural specificities of RNAs molecules: The Deformation Index enriches the Root Mean Square Deviation (*RMSD*) with base pair prediction accuracy measurements; and the Deformation Profile aims to provide multi-scale information about the differences between target and reference models at local, intra-domain and inter-domain scales. These metrics can be used to evaluate predicted RNA models against the observed RNA structures and, I hope, will help the structure prediction community in accessing the quality of their models and improving their prediction models and tools.

An immediate continuation of this work should be to apply the *RMSD* based model significance analysis proposed by (Hajdin et al., 2010) to the Deformation Index and Deformation Profile metrics in order to understand how both metrics behave in the RNA conformational space. One limitation of the current RNA structure comparison approaches is the lack of a good random model for RNA structure, i.e., a model that would describe how “random” RNA structures could be distributed in the conformational space. This model would allow us to obtain representative random samples of RNA structures against which to compare the decoys produced by automatic prediction tools. I would also like to explore the application of rotational invariant descriptors (such as 3D Zernike descriptors (Novotni and Klein, 2004)) to RNA structure comparison. This approach has been explored with relative success for protein-ligand docking by (Venkatraman

et al., 2009). A fundamental difference, however, is that for protein-ligand docking the comparison is done between the transforms of molecular surfaces while to compare RNA structures one would require a four dimensional transformation in order to compare the internal domains of the RNA models.

Following our work on structural comparison metrics, together with several members of the RNA community, we developed the first RNA prediction assessment experiment: **RNAPuzzles**. We performed the three first rounds of prediction evaluation with the participation of seven research groups, representative of the RNA structure prediction community. New structure prediction challenges should be published in the near future and I believe that the effort to develop the comparison pipeline will greatly simplify the future rounds of the experiment.

The improvement of the **RNAPuzzles** mechanism, namely with better and more complete evaluation tools, is in our road map. A useful addition to the **RNAPuzzles** would be the setup of a benchmark structure database with a set of controlled and representative examples of different sizes and complexities to allow a systematic comparison of new methods and a training ground to newcomers to the RNA structure prediction field.

(ii) I implemented a ncRNA annotation pipeline, not only to answer the immediate need for ncRNA annotation in the context of the Génolevures consortium but also, and more importantly, to provide a fast and reliable annotation protocol for the next sequencing projects of the Génolevures consortium, as well as of its successor, the Dikaryome project.

The full integration, in the development pipeline of new tools, namely the *de novo* ncRNA gene prediction tool **CMFinder** (Yao et al., 2007), as well as the possibility to process data from other sources, such as deep sequencing experiments, are the next short term challenges in this line of work.

By applying the developed annotation pipeline to the available fully sequenced yeast genomes from the Génolevures consortium and the Dikaryome project I was able to annotate an important proportion of the known ncRNA families in yeast. Besides the practical utility of annotating and publishing ncRNA gene information on new genomes, I also obtained a set of observations: New potential ncRNA genes; New synteny relationships between ncRNA loci; New examples of extra domains in well known essential ncRNAs.

The experimental confirmation of these observations, which is beyond the bioinformatics approach, should be the natural continuation of the annotation project. In the strict bioinformatics domain, I would like to pursue the development of a specific approach to detect elusive ncRNA genes such

as the telomerase RNA component taking advantage of the known structural features and the recently obtained synteny data.

(iii) Finally, I developed a new algorithm, **RMDetect**, to identify structural RNA modules from sequence information alone. This algorithm resorts on a Bayesian network for module description and on joint base pair probability estimation to candidate selection. Applying **RMDetect** to a set of available ncRNA alignments I found a number of potential motifs not yet reported. We believe the **RMDetect** is a useful step to bridge the gap between pure sequence analysis and 3D RNA studies.

I hope to improve the current approach by adding new models for structural modules collected from the current and the yet to come crystallographic structures and structural alignments. The performance improvement of our algorithm will be key to allow the effective search for structural modules in full genomes. Until now the definition of the Bayesian network topology was done either empirically or using a set of simple heuristic rules. Some preliminary experiments, using automatic tools for Bayesian network prediction (Hartemink et al., 2005), generated overly complex topologies with no simple structural interpretation and it is difficult to evaluate if those complex topologies really improve on simpler ones or just correspond to the over fit of the training data. Thus, it would be important to thoroughly explore the correlations between the bases of the structural modules and, eventually, the effects of the structural and sequence neighborhood on the nucleotides frequencies. Finally, knowing that structural modules frequently correspond to protein binding sites we would like to explore the relationship between the protein and RNA sequences on those specific binding regions.

# Appendix A

## Useful Concepts

### A.1 Score

In sequence analysis the *score* of a given alignment – also known as matching *score* – is a value that quantifies the number (and type) of matches and mismatches of the alignment. The *score* is computed according to an empirical scoring matrix that establishes correspondence values between nucleotides or amino acids based on their evolutionary or chemical properties (Durbin et al., 1998).

In decision theory a *score* – also known as log *score* – is a function defined by a ratio of probabilities:

$$score = \log \left( \frac{P(X|M)}{P(X|M_{null})} \right),$$

where  $X$  is an observation,  $M$  is the model being tested and  $M_{null}$  is the null model against which we compare  $M$ . Informally the *score* tells us how much the occurrence of  $X$  is better explained by the model  $M$  or just by chance (i.e. by the model  $M_{null}$ ).

### A.2 E-value

The expect value (*E-value*) of a hit with score  $S$  is the expected number of hits with scores equal or higher than  $S$  that one would expect to find by chance in the same “experimental” conditions (e.g. sequence length, nucleotide composition, ...). In some cases it is possible to obtain an analytical formula for the E-value. For example, the *E-value* of the hits obtained by the BLAST (Altschul et al., 1990) tool are of the form:

$$E = Ke^{-\lambda S},$$

where  $K$  and  $\lambda$  are scale parameters for, respectively, the search space size and the scoring system and  $S$  is the hit’s score (Karin and Altschul,

1990).

For many models it is not possible (or it is too complex) to derive an analytical expression for the  $E$  - value. For those model the  $E$  - value can be computed numerically using sampling strategies.

### A.3 Quantity of Information

The quantity of information, or self information, as defined in Shannon's information theory, is a measure of the information provided by the outcome of a random variable. It is formally defined as:

$$I(X = x_i) = \log_2 \left( \frac{1}{P(X = x_i)} \right).$$

The information unit depends on the base of logarithm used. In this case the this unit is the *bit* given that we use a logarithm of base 2.

The quantity of information will be important to the subsequent definitions of entropy and mutual information.

### A.4 Entropy

Entropy is the expected value of the quantity of information of a random variable. It is formally defined as:

$$H(X) = \sum_{i=1}^n P(X = x_i) \times I(X = x_i),$$

where  $X$  is a discrete random variable that can assume  $n$  values.

In sequence analysis entropy can be used as a rough measure of conservation of the individual columns of an alignment. For example, each column of an alignment can be interpreted as a random variable  $C$  and the corresponding rows as the outcomes of  $C$ . Thus, the relative frequency of occurrence of  $nt$  can be described as  $P(C = nt)$  and the entropy of  $C$  as:

$$H(C) = \sum_{nt \in [A,C,G,U]} P(C = nt) \times I(C = nt).$$

It is then easy to see that if  $C$  is conserved in all rows of the alignment we will have  $H(C) = 0$ . Conversely in the extreme case where  $C$  can be any of the four nucleotides with equal probability  $H(C) = 2$ . This way  $H(C)$  can be used as a simple measure of conservation as it measures the variability of the columns of an alignment.

## A.5 Mutual Information

The mutual information is a measure of correlation between two random variables. It is given by:

$$MI(X, Y) = \sum_{i=1}^n \sum_{j=1}^m P(X = x_i, Y = y_j) \times \log_2 \left( \frac{P(X = x_i, Y = y_j)}{P(X = x_i)P(Y = y_j)} \right).$$

Informally one can assume that the mutual information between  $X$  and  $Y$  expresses the knowledge one obtains about the outcome of  $Y$  when the outcome of  $X$  is known (and vice versa).

RNA helices consist on stacks of Watson-Crick base pairs formed by the following combinations of the four bases: CG, GC, AU, UA, GU and UG. When comparing homologous sequences corresponding to the same helical regions one can often observe compensatory mutations between the base pair position (e.g. a CG base pair becomes a CG or an AU base pair). In multiple sequences alignments the columns corresponding to a base pair will be statistically correlated if compensatory mutations occurred. Thus, computing the mutual information of all pairs of columns in an alignment will help us to detect potential paired bases. We can rewrite the mutual information definition to deal with alignment columns  $C_i$  and  $C_j$ :

$$MI(C_i, C_j) = \sum_{(nt_i, nt_j) \in [A, C, G, U]} P(C_i = nt_i, C_j = nt_j) \times \log_2 \left( \frac{P(C_i = nt_i, C_j = nt_j)}{P(C_i = nt_i)P(C_j = nt_j)} \right)$$

A good discussion on the different applications of mutual information for measuring covariation between RNA alignment columns can be found in (Lindgreen et al., 2006).

## A.6 Radius of gyration

The radius of gyration  $R_g$  of a molecule  $M$  is a measure for the average dimension of a molecule and it is given by:

$$R_g(M) = \frac{\sum_{i=1}^N (r_i - \langle r \rangle)^2}{N}$$

where  $N$  is the number of “components” to consider (they can be the atoms or the residues of the molecule),  $r_i$  is the position of component  $i$  and  $\langle r \rangle$  is the mean position of all components.

Candidate	Score1	Score2	P/N
1	95	96	P
2	76	71	P
3	51	51	P
4	22	22	P
5	19	15	P
6	10	10	N
7	21	21	N
8	41	63	N
9	62	64	P
10	84	82	P

Table A.1: Example of a list of gene candidates produced by two gene discovery systems with the respective scores. The rightmost column indicates if the candidate is a real gene (“P”) or a wrong prediction (“N”).

## A.7 Receiver Operating Characteristic

The receiver operating characteristic (ROC) curve is a method for evaluating the performance of binary classifiers. The ROC curve is a plot of the False Positive Rate (FPR)<sup>1</sup> against the True Positive Rate (TPR)<sup>2</sup> for the full range of threshold values of a given discrimination quantity.

To exemplify the usage of the ROC curve in the context of gene discovery let assume that two gene discovery systems produce, each, a list of gene candidates. Table A.1 represents a fictitious list of candidates produced by those systems and the respective scores. The rightmost column indicates if the candidate is a real gene or a wrong prediction (remember that this information is unknown previously to the gene discovery process).

Now let order the candidates according to *Score1* (see Table A.2). Notice that below a given score value (51) all candidates correspond to true genes. If one assumes that all candidates below this value are true candidates and all other are false predictions (binary classifier) one obtains a perfect specificity (i.e., no false candidates are classified as true,  $FPR = 0.0$ ). However, the sensitivity will be less than perfect (i.e., several true candidates will be classified as false,  $TPR < 1.0$ ). It is easy to see from the values on the

<sup>1</sup>The FPR, also represented as 1-specificity, is defined as:  $FPR = \frac{FP}{FP+TN}$ , where  $FP$  and  $TN$  are, respectively, the number of False Positives and True Negatives for a given discriminator threshold.

<sup>2</sup>The TPR, also known as sensitivity, is defined as:  $TPR = \frac{TP}{TP+FN}$ , where  $TP$  and  $FN$  are, respectively, the number of True Positives and False Negatives for a given discriminator threshold.



<b>Cand.</b> ( <i>Score1</i> ) <b>T/F</b>	100	80	60	40	20	0
1 (95) P	FN	TP	TP	TP	TP	TP
10 (84) P	FN	TP	TP	TP	TP	TP
2 (76) P	FN	FN	TP	TP	TP	TP
9 (62) P	FN	FN	TP	TP	TP	TP
3 (51) P	FN	FN	FN	TP	TP	TP
8 (41) N	TN	TN	TN	FP	FP	FP
4 (22) P	FN	FN	FN	FN	TP	TP
7 (21) N	TN	TN	TN	TN	FP	FP
5 (19) P	FN	FN	FN	FN	FN	TP
6 (10) N	TN	TN	TN	TN	TN	FP
<b>TP</b>	0	2	4	5	6	7
<b>FN</b>	7	5	3	2	1	0
<b>TN</b>	3	3	3	2	1	0
<b>FP</b>	0	0	0	1	2	3
<b>TPR</b>	0.0	0.28	0.57	0.71	0.86	1.0
<b>FPR</b>	0.0	0.0	0.0	0.33	0.67	1.0

Table A.2: List of gene candidates ordered by *Score1* and respective *TPR* and *FPR* values for several thresholds.

table that no threshold will produce a perfect classifier (i.e.  $FPR = 0.0$  and  $TPR = 1.0$ ), which is almost always the case in real life applications.

Repeating the same exercise using the *Score2* as the discriminator quantity one obtains the values on table A.3.

A legitimate question would be: “Which of the scores, *Score1* or *Score2*, is the best discriminator?”. This is precisely the type of question that the ROC curve is supposed to help answering. Figure A.1 displays the ROC curves for both cases. The discriminator for which the area under the curve is largest is, usually, considered a better discriminator (in our toy example it would be *Score1*). Notice the ideal classifier would produce the dashed green curve in the example, while a non discriminative, random classifier would produce the green diagonal line).

For more information on ROC analysis please see (Brown and Davis, 2006).

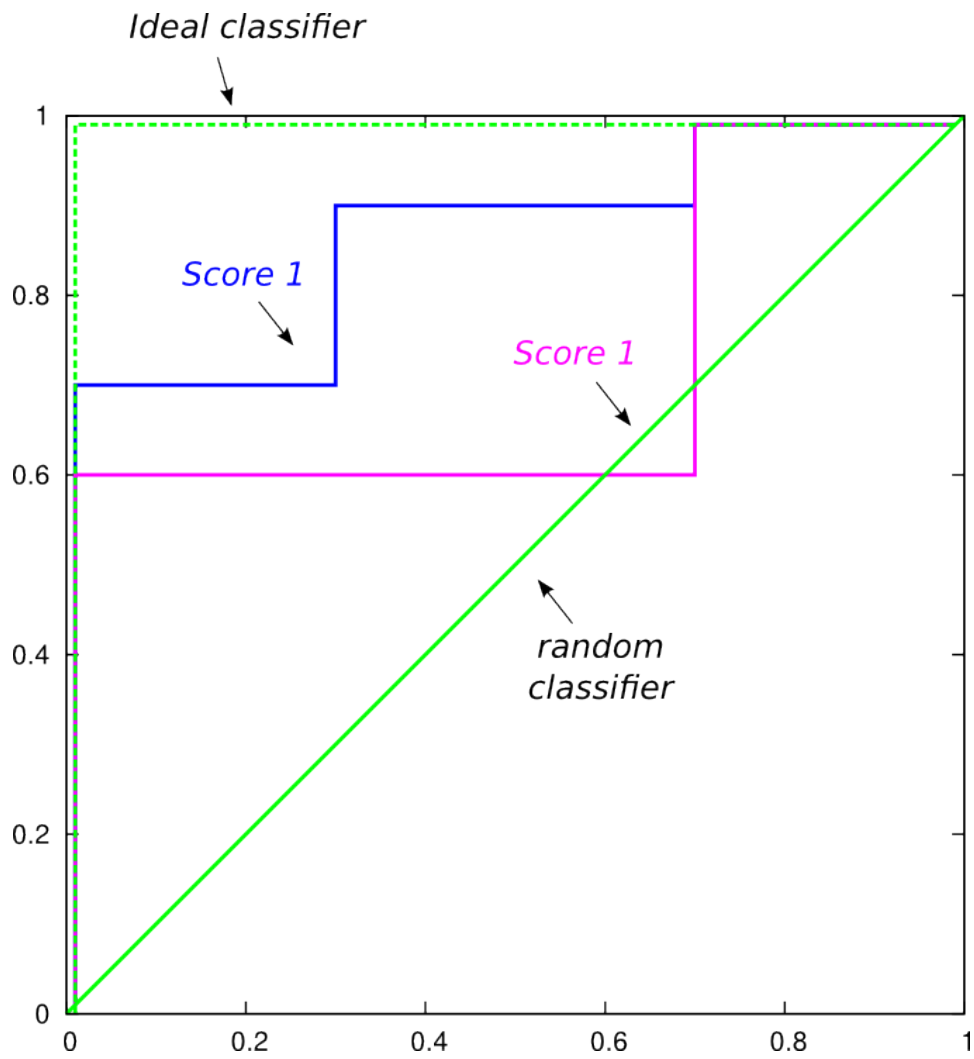


Figure A.1: Example of the ROC curve for *Score1* (blue) and *Score2* (magenta). The dashed green curve corresponds to an, ideal, perfect classifier and the diagonal green line to a random classifier.

<b>Cand. (<i>Score2</i>) P/N</b>	100	80	60	40	20	0
1 (96) P	FN	TP	TP	TP	TP	TP
10 (82) P	FN	TP	TP	TP	TP	TP
2 (71) P	FN	FN	TP	TP	TP	TP
9 (64) P	FN	FN	TP	TP	TP	TP
8 (63) N	TN	TN	FP	TP	FP	FP
7 (56) N	TN	TN	TN	FP	FP	FP
3 (51) P	FN	FN	FN	FP	TP	TP
4 (22) P	FN	FN	FN	FN	TP	TP
5 (21) P	FN	FN	FN	FN	TP	TP
6 (10) N	TN	TN	TN	TN	TN	FP
<b>TP</b>	0	2	4	5	7	7
<b>FN</b>	7	5	3	2	0	0
<b>TN</b>	3	3	2	1	1	0
<b>FP</b>	0	0	1	2	2	3
<b>TPR</b>	0.0	0.29	0.57	0.71	1.0	1.0
<b>FPR</b>	0.0	0.0	0.33	0.67	0.67	1.0

Table A.3: List of gene candidates ordered by *Score2* and respective *TPR* and *FPR* values for several thresholds.

## A.8 Z-Score

The Z-score is a dimensionless quantity that indicates how many standard deviations a certain measured value diverges from the mean of the population. The formal definition is given by:

$$Z_{score} = \frac{x - \mu}{\sigma}$$

Where  $x$  is the measured quantity,  $\mu$  is the population mean and  $\sigma$  the population variance.

## Appendix B

# Base Pairs

Nucleotide bases interact with each other through hydrogen bonds forming “base pairs” which are the most ubiquitous structural features of RNA molecules: More than three quarters of the bases in structured RNA molecules form a base pair with at least one other base. Base pairs are the building blocks of the helices – stacks of Watson-Crick (WC) base pairs – and of structural modules – stacks of non-Watson-Crick (non-WC) base pairs. Base pairs (either WC or non-WC) stabilize the tertiary structure of the molecule by establishing long range interactions between secondary structure elements such as loop-loop and loop-helix interactions or between structural modules. The study and classification of base pairs is a useful tool for systematizing and understanding RNA structures.

RNA molecules are long chains of nucleotides each of which contains one of four bases: Adenine, Cytosine, Guanine and Uracil. Each base presents three edges: the Watson-Crick edge, the Sugar edge and the Hoogsteen edge (see Figure B.1). A base pair is formed when the exposed atoms of one base edge establish hydrogen bonds with the exposed atoms of another base edge. Base pairs can, thus, be classified according to the edges involved in the interaction.

Twelve types of base pairs can be observed (see Figure B.2): Each one of the six edge-edge combinations can occur in either *cis* or *trans* conformations depending on the relative positions of the respective sugar groups (see Figure B.3). In order to represent the several base pair types in secondary structure diagrams the Leontis-Westhof notation (Leontis and Westhof, 2001) establishes one unique symbol for each edge as depicted in figure B.2.

### B.1 Isostericity

Two base pairs of the same type are said to be isosteric if they have the same distance between the C1' atoms of the respective nucleotides, irrespective of the nature of the bases involved (see Figure B.4). Isostericity is an important

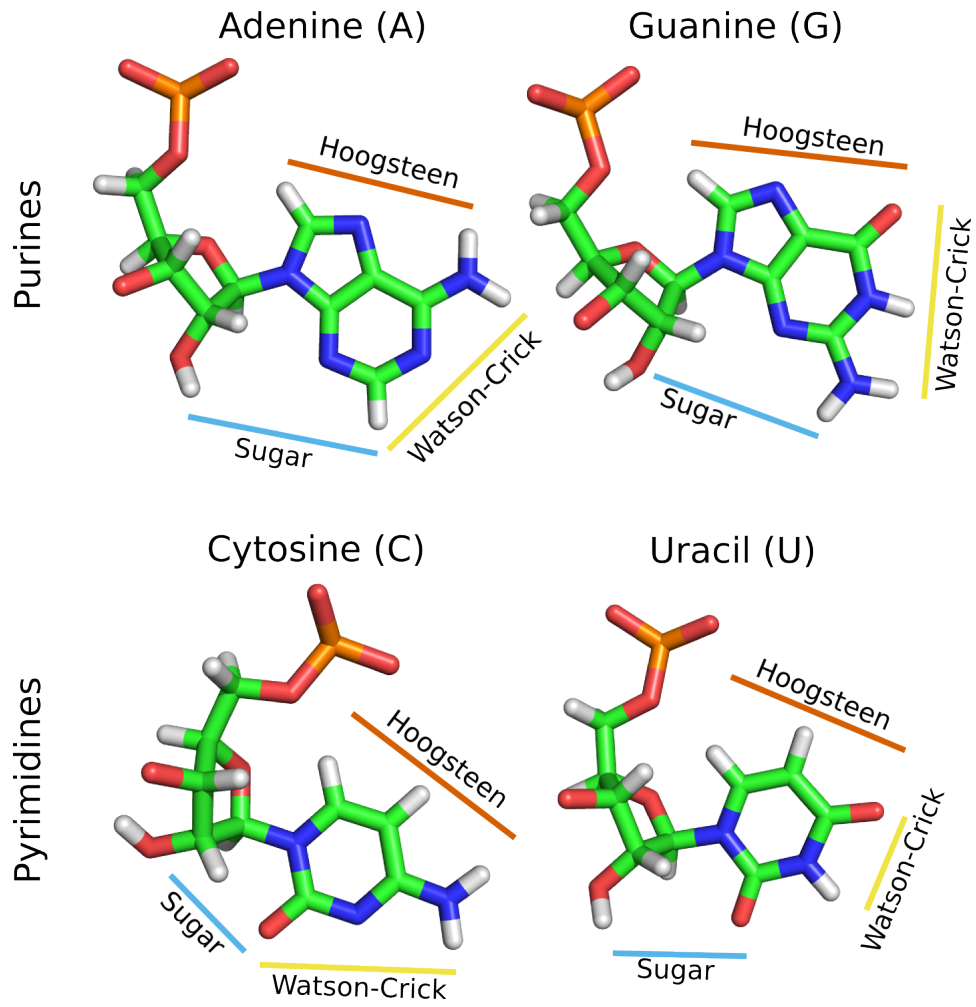


Figure B.1: The four bases and their respective edges.

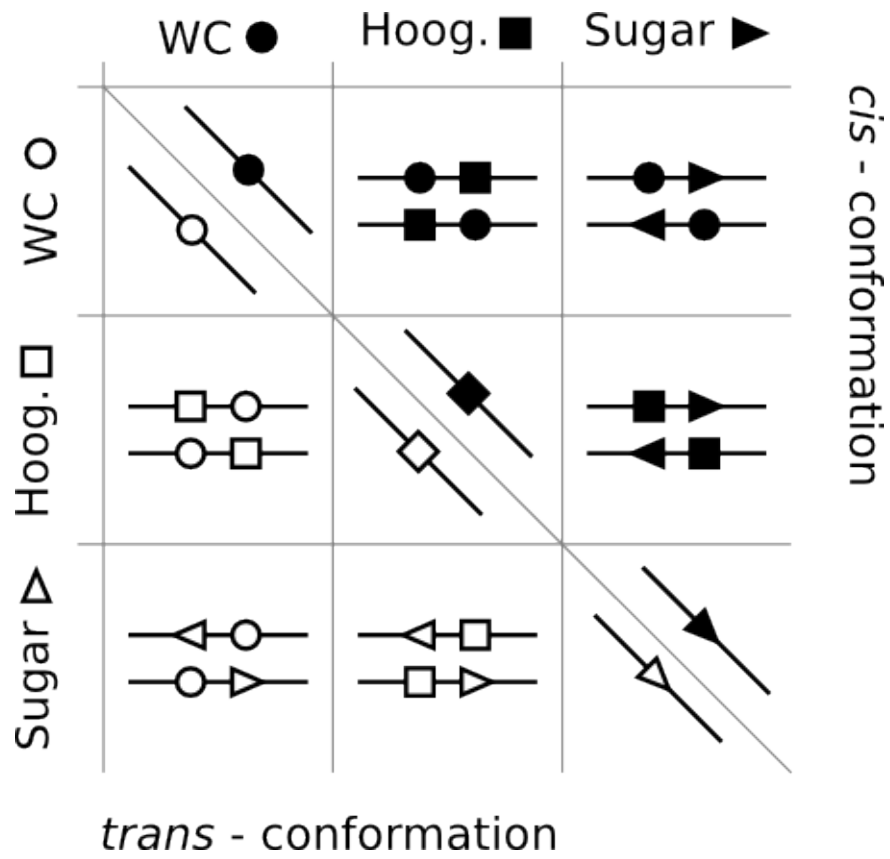


Figure B.2: Leontis-Westhof classification of the twelve base pair types.

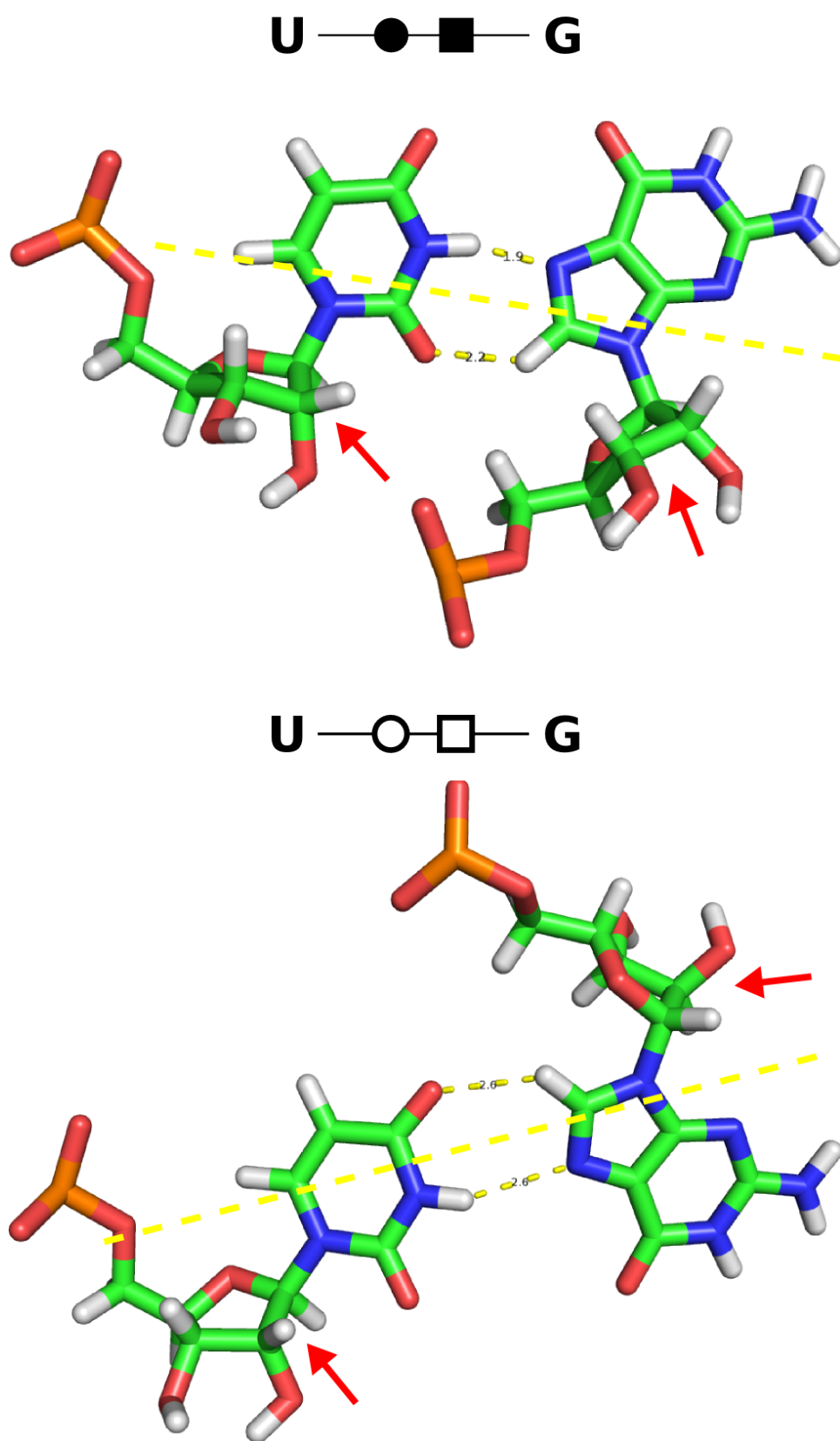


Figure B.3: Example of a *cis* and *trans* base pairs. In the *cis* case (top), the sugar moieties (red arrows) stand at the same side of the H-bonds (dashed yellow lines). In the *trans* case (bottom) the sugar moieties (red arrows) stand at opposite sides of the H-bonds (dashed yellow lines). Base pairs from the PDB file 1JJ2 (Klein et al., 2001).

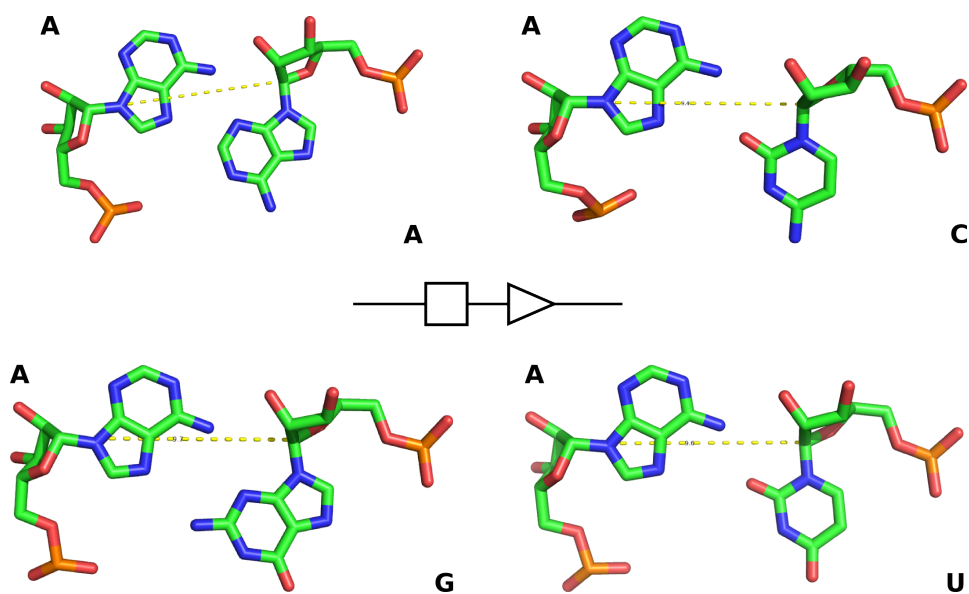


Figure B.4: Examples of four isosteric base pairs. Four Hoogsteen-Sugar base pair presenting similar C1'-C1' distances: AA: 9.3 Å (top left, PDB:1GID) (Cate et al., 1996), AC: 9.4 Å (top right, PDB 1JJ2) (Klein et al., 2001), AG: 9.7 Å (bottom left, PDB 354D) (Correll et al., 1997) and AU 9.6 Å (bottom right, PDB:1JJ2) (Klein et al., 2001).

property as isosteric base pairs can replace each other without changing the local conformation of the molecule where they occur, thus allowing for structurally synonymous substitutions. A complete list of isosteric base pairs can be found in (Leontis et al., 2002).



# Bibliography

- Adilakshmi, T., Bellur, D. L., and Woodson, S. A. (2008). Concurrent nucleation of 16S folding and induced fit in 30S ribosome assembly. *October*, 455(October):18–22.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215:403–410.
- Amaral, P. P., Neyt, C., Wilkins, S. J., Askarian-Amiri, M. E., Sunkin, S. M., Perkins, A. C., and Mattick, J. S. (2009). Complex architecture and regulated expression of the Sox2ot locus during vertebrate development. *RNA (New York, N.Y.)*, 15(11):2013–27.
- Antonioli, A. H., Cochrane, J. C., Lipchock, S. V., and Strobel, S. a. (2010). Plasticity of the RNA kink turn structural motif. *RNA (New York, N.Y.)*, 16(4):762–8.
- Apostolico, A., Ciriello, G., Guerra, C., Heitsch, C. E., Hsiao, C., and Williams, L. D. (2009). Finding 3D motifs in ribosomal RNA structures. *Nucleic acids research*, 37(4):e29.
- Barash, Y., Elidan, G., Friedman, N., and Kaplan, T. (2003). Modeling dependencies in protein-DNA binding sites. *Proceedings of the seventh annual international conference on Computational molecular biology - RECOMB '03*, pages 28–37.
- Barrick, J. E. and Breaker, R. R. (2007). The distributions, mechanisms, and structures of metabolite-binding riboswitches. *Genome biology*, 8(11):R239.
- Ben-Gal, I., Shani, a., Gohr, a., Grau, J., Arviv, S., Shmilovici, a., Posch, S., and Grosse, I. (2005). Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics (Oxford, England)*, 21(11):2657–66.
- Benson, D. a., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2008). GenBank. *Nucleic acids research*, 36(Database issue):D25–30.

- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic acids research*, 28(1):235–42.
- Betancourt, M. R. and Skolnick, J. (2001). Universal similarity measure for comparing protein structures. *Biopolymers*, 59(5):305–9.
- Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F. A., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., Hausler, D., and Miller, W. (2004). Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. *Genome Research*, 14:708–715.
- Bocobza, S., Adato, A., Mandel, T., Shapira, M., Nudler, E., and Aharoni, A. (2007). Riboswitch-dependent gene regulation and its evolution in the plant kingdom. *Genes & Development*, pages 2874–2879.
- Breaker, R. R. (2008). Complex riboswitches. *Science (New York, N.Y.)*, 319(5871):1795–7.
- Brown, C. and Davis, H. (2006). Receiver operating characteristics curves and related decision measures: A tutorial. *Chemometrics and Intelligent Laboratory Systems*, 80(1):24–38.
- Butler, G., Rasmussen, M. D., Lin, M. F., Santos, M. a. S., Sakthikumar, S., Munro, C. a., Rheinbay, E., Grabherr, M., Forche, A., Reedy, J. L., Agrafioti, I., Arnaud, M. B., Bates, S., Brown, A. J. P., Brunke, S., Costanzo, M. C., Fitzpatrick, D. a., de Groot, P. W. J., Harris, D., Hoyer, L. L., Hube, B., Klis, F. M., Kodira, C., Lennard, N., Logue, M. E., Martin, R., Neiman, A. M., Nikolaou, E., Quail, M. a., Quinn, J., Santos, M. C., Schmitzberger, F. F., Sherlock, G., Shah, P., Silverstein, K. a. T., Skrzypek, M. S., Soll, D., Staggs, R., Stansfield, I., Stumpf, M. P. H., Sudbery, P. E., Srikantha, T., Zeng, Q., Berman, J., Berriman, M., Heitman, J., Gow, N. a. R., Lorenz, M. C., Birren, B. W., Kellis, M., and Cuomo, C. a. (2009). Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature*, 459(7247):657–62.
- Cao, S. and Chen, S.-J. (2005). Predicting RNA folding thermodynamics with a reduced chain representation model. *RNA (New York, N.Y.)*, 11(12):1884–97.
- Cao, S. and Chen, S.-J. (2011). Physics-Based De Novo Prediction of RNA 3D Structures. *The journal of physical chemistry. B*, 115(14):4216–26.
- Carthew, R. W. and Sontheimer, E. J. (2009). Origins and Mechanisms of miRNAs and siRNAs. *Cell*, 136(4):642–55.

- Carugo, O. (2003). How root-mean-square distance (r.m.s.d.) values depend on the resolution of protein structures that are compared. *Journal of Applied Crystallography*, 36(1):125–128.
- Cate, J. H., Gooding, a. R., Podell, E., Zhou, K., Golden, B. L., Szewczak, a. a., Kundrot, C. E., Cech, T. R., and Doudna, J. a. (1996). RNA tertiary structure mediated by adenosine platforms. *Science (New York, N.Y.)*, 273(5282):1696–9.
- Chauhan, S. and Woodson, S. a. (2008). Tertiary interactions determine the accuracy of RNA folding. *Journal of the American Chemical Society*, 130(4):1296–303.
- Cheah, M. T., Wachter, A., Sudarsan, N., and Breaker, R. R. (2007). Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. *Nature*, 447(May):3–7.
- Chickering, D., Heckerman, D., Meek, C., and Madigan, D. (1994). Learning Bayesian networks is NP-hard. *Microsoft Research Tech report*, MSR-TR-94-.
- Clote, P., Ferré, F., Kranakis, E., and Krizanc, D. (2005). Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA (New York, N.Y.)*, 11(5):578–91.
- Correll, C. C., Freeborn, B., Moore, P. B., and Steitz, T. a. (1997). Metals, motifs, and recognition in the crystal structure of a 5S rRNA domain. *Cell*, 91(5):705–12.
- Cruz, J. A., Boniecki, M., Bujnicki, J. M., Chen, S.-J., Cao, S., Das, R., Ding, F., Dokholyan, N. V., Flores, S. C., Lavender, C. A., Major, F., Mikolajczak, K., Philips, A., Puton, T., Santalucia, J., Hermann, T., Rother, K., Rother, M., Serganov, A., Skorupski, M., Soltysinski, T., Sripakdeevong, P., Tuszyńska, I., Weeks, K. M., Waldsich, C., Wildauer, M., Leontis, N. B., and Westhof, E. A CASP-like evaluation of RNA 3D structure predictions. *in preparation*.
- Cruz, J. A. and Westhof, E. (2009). The dynamic landscapes of RNA architecture. *Cell*, 136(4):604–9.
- Cruz, J. A. and Westhof, E. (2011a). Identification and annotation of noncoding RNAs in Saccharomycotina. *Comptes Rendus Biologies*, 334:671–678.
- Cruz, J. A. and Westhof, E. (2011b). sequence-based identification of 3d structural modules in RNA with rmdetect. *Nature Methods*, 8(6):513–519.

- Das, R. and Baker, D. (2007). Automated de novo prediction of native-like RNA tertiary structures. *Proceedings of the National Academy of Sciences of the United States of America*, 104(37):14664–9.
- Das, R., Karanicolas, J., and Baker, D. (2010). Atomic accuracy in predicting and designing noncanonical RNA structure. *Nature methods*, 7(4):291–4.
- Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall, W. B., Snoeyink, J., Richardson, J. S., and Richardson, D. C. (2007). MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic acids research*, 35(Web Server issue):W375–83.
- Diener, J. L. and Moore, P. B. (1998). Solution structure of a substrate for the archaeal pre-tRNA splicing endonucleases: the bulge-helix-bulge motif. *Molecular cell*, 1(6):883–94.
- Ding, F., Sharma, S., Chalasani, P., Demidov, V. V., Broude, N. E., and Dokholyan, N. V. (2008). Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA (New York, N.Y.)*, 14(6):1164–73.
- Dinger, M. E., Amaral, P. P., Mercer, T. R., Pang, K. C., Bruce, S. J., Gardiner, B. B., Askarian-Amiri, M. E., Ru, K., Soldà, G., Simons, C., Sunkin, S. M., Crowe, M. L., Grimmond, S. M., Perkins, A. C., and Mattick, J. S. (2008). Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome research*, 18(9):1433–45.
- Djelloul, M. and Denise, A. (2008). Automated motif extraction and classification in RNA tertiary structures. *RNA (New York, N.Y.)*, 14(12):2489–97.
- Duarte, C. M., Wadley, L. M., and Pyle, A. M. (2003). RNA structure comparison, motif search and discovery using a reduced representation of RNA conformational space. *Nucleic Acids Research*, 31(16):4755–4761.
- Dujon, B. (2010). *Evolutionary Genomics of Yeasts*, chapter 6, pages 95–120. John Wiley & Sons, Inc.
- Dunham, C. M., Murray, J. B., and Scott, W. G. (2003). A Helical Twist-induced Conformational Switch Activates Cleavage in the Hammerhead Ribozyme. *Journal of Molecular Biology*, 332(2):327–336.
- Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

- Eddy, S. R. and Durbin, R. (1994). RNA analysis using covariance models. *22*(11).
- Edwards, T. E. and Ferré-D'Amaré, A. R. (2006). Crystal structures of the thi-box riboswitch bound to thiamine pyrophosphate analogs reveal adaptive RNA-small molecule recognition. *Structure (London, England : 1993)*, *14*(9):1459–68.
- Eidhammer, I., Jonassen, I., and Taylor, W. R. (2000). Structure comparison and structure patterns. *Journal of computational biology : a journal of computational molecular cell biology*, *7*(5):685–716.
- ENCODE (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, *447*(7146):799–816.
- Falicov, a. and Cohen, F. E. (1996). A surface of minimum area metric for the structural comparison of proteins. *Journal of molecular biology*, *258*(5):871–92.
- Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, *391*:806–811.
- Flores, S. C. and Altman, R. B. (2010). Turning limited experimental information into 3D models of RNA. *RNA (New York, N.Y.)*.
- Frith, M. C., Pheasant, M., and Mattick, J. S. (2005). The amazing complexity of the human transcriptome. *European journal of human genetics : EJHG*, *13*(8):894–7.
- Gardner, P. P., Daub, J., Tate, J. G., Nawrocki, E. P., Kolbe, D. L., Lindgreen, S., Wilkinson, A. C., Finn, R. D., Griffiths-Jones, S., Eddy, S. R., and Bateman, A. (2009). Rfam: updates to the RNA families database. *Nucleic acids research*, *37*(Database issue):D136–40.
- Gattiker, A., Rischatsch, R., Demougin, P., Voegeli, S., Dietrich, F. S., Philippsen, P., and Primig, M. (2007). Ashbya Genome Database 3.0: a cross-species genome and transcriptome browser for yeast biologists. *BMC genomics*, *8*:9.
- Gautheret, D., Konings, D., and Gutell, R. R. (1994). A major family of motifs involving G - A mismatches in ribosomal RNA. *Journal of Molecular Biology*, *242*(1):1–8.
- Gendron, P., Lemieux, S., and Major, F. (2001). Quantitative analysis of nucleic acid three-dimensional structures. *Journal of molecular biology*, *308*(5):919–36.

- Gerstein, M. and Levitt, M. (1998). Comprehensive Assessment of Automatic Structural Alignment against a Manual Standard , the Scop Classification of Proteins Keywords. *Protein science*, (7):445–456.
- Gilbert, W. (1986). Origin of life: The RNA world. *Nature*, 319(6055):618.
- Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S. G. (1996). Life with 6000 Genes. *Science*, 274:562–566.
- Greenleaf, W. J., Frieda, K. L., Foster, D. a. N., Woodside, M. T., and Block, S. M. (2008). Direct observation of hierarchical folding in single riboswitch aptamers. *Science (New York, N.Y.)*, 319(5863):630–3.
- Griffiths-Jones, S. (2005). RALEE–RNA ALignment editor in Emacs. *Bioinformatics (Oxford, England)*, 21(2):257–9.
- Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., and Altman, S. (1983). The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35(3, Part 2):849–857.
- Gunisova, S., Elboher, E., Nosek, J., Gorkovoy, V., Brown, Y., Lucier, J.-F., Laterreur, N., Wellinger, R. J., Tzfati, Y., and Tomaska, L. (2009). Identification and comparative analysis of telomerase RNAs from *Candida* species reveal conservation of functional elements. *RNA (New York, N.Y.)*, 15(4):546–59.
- Hajdin, C. E., Ding, F., Dokholyan, N. V., and Weeks, K. M. (2010). On the significance of an RNA tertiary structure prediction. *Spring*, pages 1–10.
- Harrow, J., Nagy, A., Reymond, A., Alioto, T., Patthy, L., Antonarakis, S. E., and Guigó, R. (2009). Identifying protein-coding genes in genomic sequences. *Genome biology*, 10(1):201.
- Hartemink, A. J., Sladeczek, J., and Robinson, J. (2005). BANJO (Bayesian Network Inference with Java Objects).
- Hendrix, D. K., Brenner, S. E., and Holbrook, S. R. (2005). RNA structural motifs: building blocks of a modular biomolecule. *Quarterly reviews of biophysics*, 38(3):221–43.
- Hershkovitz, E. (2003). Automated identification of RNA conformational motifs: theory and application to the HM LSU 23S rRNA. *Nucleic Acids Research*, 31(21):6249–6257.

- Hertel, J., Hofacker, I. L., and Stadler, P. F. (2008). SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics (Oxford, England)*, 24(2):158–64.
- Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte fur Chemie Chemical Monthly*, 125(2):167–188.
- Holm, L. and Sander, C. (1993). Protein Structure Comparison by Alignment of Distance Matrices. *Journal of molecular biology*, (233):123–138.
- Igel, A. H. and Ares, M. (1988). Internal sequences that distinguish yeast from metazoan U2 snRNA are unnecessary for pre-mRNA splicing. *Nature*, 334(6181):450–453.
- Jossinet, F., Ludwig, T. E., and Westhof, E. (2010). Assemble: an interactive graphical tool to analyze and build RNA architectures at the 2D and 3D levels. *Bioinformatics (Oxford, England)*, 26(16):2057–2059.
- Jung, C.-H., Hansen, M. a., Makunin, I. V., Korbie, D. J., and Mattick, J. S. (2010). Identification of novel non-coding RNAs using profiles of short sequence reads from next generation sequencing data. *BMC genomics*, 11:77.
- Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica*, 32:922–923.
- Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica*, (34):827–828.
- Kachouri, R., Stribinskis, V., Zhu, Y., Ramos, K. S., Westhof, E., and Li, Y. (2005). A surprisingly large RNase P RNA in *Candida glabrata*. 11:1064–1072.
- Kachouri-Lafond, R., Dujon, B., Gilson, E., Westhof, E., Fairhead, C., and Teixeira, M. T. (2009). Large telomerase RNA, telomere length heterogeneity and escape from senescence in *Candida glabrata*. *FEBS letters*, 583(22):3605–10.
- Karlin, S. and Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences*, 87:2264–2268.
- Klein, D. J., Schmeing, T. M., Moore, P. B., and Steitz, T. a. (2001). The kink-turn: a new RNA secondary structure motif. *The EMBO journal*, 20(15):4214–21.

- Kretzner, L., Krol, a., and Rosbash, M. (1990). *Saccharomyces cerevisiae* U1 small nuclear RNA secondary structure contains both universal and yeast-specific domains. *Proceedings of the National Academy of Sciences of the United States of America*, 87(2):851–5.
- Kretzner, L., Rymond, B. C., and Rosbash, M. (1987). *S. cerevisiae* U1 RNA is large and has limited primary sequence homology to metazoan U1 snRNA. *Cell*, 50(4):593–602.
- Kruger, K., Grabowski, P. J., Zaug, A. J., Sands, J., Gottschling, D. E., and Cech, T. R. (1982). Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. *Cell*, 31(1):147–157.
- Kubodera, T., Watanabe, M., Yoshiuchi, K., Yamashita, N., Nishimura, A., Nakai, S., Gomi, K., and Hanamoto, H. (2003). Thiamine-regulated gene expression of *Aspergillus oryzae* thiA requires splicing of the intron containing a riboswitch-like domain in the 5-UTR. *FEBS Letters*, 555(3):516–520.
- Kurtzman, C. P. and Fell, J. W. (2006). Yeast Systematics and Phylogeny Implications of Molecular Identification Methods for Studies in. In *The Yeast Handbook*, number C, chapter 2, pages 11–30.
- Lancia, G. and Istrail, S. (2003). Protein Structure Comparison : Algorithms and. In Guerra, C. and Istrail, S., editors, *Protein Structure Analysis and Design*, pages 1–33.
- Leontis, N. and Westhof, E. (2003a). Analysis of RNA motifs. *Current Opinion in Structural Biology*, 13(3):300–308.
- Leontis, N. B., Stombaugh, J., and Westhof, E. (2002). The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic acids research*, 30(16):3497–531.
- Leontis, N. B. and Westhof, E. (2001). Geometric nomenclature and classification of RNA base pairs. *RNA (New York, N.Y.)*, 7(4):499–512.
- Leontis, N. B. and Westhof, E. (2003b). Analysis of RNA motifs. *Current Opinion in Structural Biology*, 13(3):300–308.
- Lescoute, A., Leontis, N. B., Massire, C., and Westhof, E. (2005). Recurrent structural RNA motifs, Isostericity Matrices and sequence alignments. *Nucleic acids research*, 33(8):2395–409.
- Lescoute, a. and Westhof, E. (2006). The A-minor motifs in the decoding recognition process. *Biochimie*, 88(8):993–9.



- Liao, X. L., Kretzner, L., Seraphin, B., and Rosbash, M. (1990). Universally conserved and yeast-specific U1 snRNA sequences are important but not essential for U1 snRNP function. *Genes & Development*, 4:1766–1774.
- Lindgreen, S., Gardner, P. P., and Krogh, a. (2006). Measuring covariation in RNA alignments: physical realism improves information measures. *Bioinformatics (Oxford, England)*, 22(24):2988–95.
- Lowe, T. M. (1999). A Computational Screen for Methylation Guide snoRNAs in Yeast. *Science*, 283(5405):1168–1171.
- Lowe, T. M. and Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research*, 25(5):955–64.
- Maiorov, V. N. and Crippen, G. M. (1994). Significance of Root-Mean-Square Deviation in Comparing Three-dimensional Structures of Globular Proteins. *Journal of molecular biology*, 255(4):625–634.
- Mamidipally, C., Noronha, S. B., and Roy, S. D. (2009). Automated Identification of Protein Structural Features. In *Pattern Recognition and Machine Intelligence: Third International Conference, PReMI 2009*.
- Martick, M. and Scott, W. G. (2006). Tertiary contacts distant from the active site prime a ribozyme for catalysis. *Cell*, 126(2):309–20.
- Martinez, H. M., Maizel, J. V., and Shapiro, B. a. (2008). RNA2D3D: a program for generating, viewing, and comparing 3-dimensional models of RNA. *Journal of biomolecular structure & dynamics*, 25(6):669–83.
- Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451.
- McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–19.
- Mehler, M. F. and Mattick, J. S. (2006). Non-coding RNAs in the nervous system. *The Journal of physiology*, 575(Pt 2):333–41.
- Mercer, T. R., Dinger, M. E., Sunkin, S. M., Mehler, M. F., and Mattick, J. S. (2008). Specific expression of long noncoding RNAs in the mouse brain. *Proceedings of the National Academy of Sciences of the United States of America*, 105(2):716–21.
- Michel, F. and Westhof, E. (1990). Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *Journal of Molecular Biology*, 216(3):585–610.

- Mironov, A. S., Gusarov, I., Rafikov, R., Lopez, L. E., Shatalin, K., Kreneva, R. a., Perumov, D. a., and Nudler, E. (2002). Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell*, 111(5):747–56.
- Moore, P. B. (1999). Structural Motifs in RNA. *Annual Review of Biochemistry*, 68:287–300.
- Moult, J., Pedersen, J. T., Judson, R., and Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, 23(3):ii—iv.
- Nahvi, A., Sudarsan, N., Ebert, M. S., Zou, X., Brown, K. L., and Breaker, R. R. (2002). Genetic control by a metabolite binding mRNA. *Chemistry & biology*, 9(9):1043.
- Nawrocki, E. P., Kolbe, D. L., and Eddy, S. R. (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics (Oxford, England)*, 25(10):1335–7.
- Nechooshtan, G., Elgrably-Weiss, M., Sheaffer, A., Westhof, E., and Altuvia, S. (2009). A pH-responsive riboregulator. *Genes & development*, 23(22):2650–62.
- Noble, D. (2008). Genes and causation. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 366(1878):3001–15.
- Noeske, J., Buck, J., Furtig, B., Nasiri, H. R., Schwalbe, H., and Wohnert, J. (2007). Interplay of induced fit and preorganization in the ligand induced folding of the aptamer domain of the guanine binding riboswitch. *Structure*, 35(2):572–583.
- Noiro, C., Gaspin, C., Schiex, T., and Gouzy, J. (2008). LeARN : a platform for detecting , clustering and annotating. *BMC Bioinformatics*, 11:1–11.
- Novotni, M. and Klein, R. (2004). Shape retrieval using 3D Zernike descriptors. *Computer-Aided Design*, 36(11):1047–1062.
- Parisien, M., Cruz, J. A., Westhof, E., and Major, F. (2009). New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA (New York, N.Y.)*, 15(10):1875–85.
- Parisien, M. and Major, F. (2008). The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, 452(7183):51–5.
- Pfaller, M. a. and Diekema, D. J. (2007). Epidemiology of invasive candidiasis: a persistent public health problem. *Clinical microbiology reviews*, 20(1):133–63.

- Piekna-Przybylska, D., Decatur, W. A., and Fournier, M. J. (2007). New bioinformatic tools for analysis of nucleotide modifications in eukaryotic rRNA. *Spring*, pages 305–312.
- Podlevsky, J. D., Bley, C. J., Omana, R. V., Qi, X., and Chen, J. J.-L. (2008). The telomerase database. *Nucleic acids research*, 36(Database issue):D339–43.
- Ponting, C. P., Oliver, P. L., and Reik, W. (2009). Evolution and functions of long noncoding RNAs. *Cell*, 136(4):629–41.
- Regalia, M., Rosenblad, M. A., and Samuelsson, T. (2002). Prediction of signal recognition particle RNA genes. *Nucleic acids research*, 30(15):3368–77.
- Reuter, J. S. and Mathews, D. H. (2010). RNAstructure: software for RNA secondary structure prediction and analysis. *BMC bioinformatics*, 11:129.
- Richardson, J. S., Schneider, B., Murray, L. W., Kapral, G. J., Immormino, R. M., Headd, J. J., Richardson, D. C., Ham, D., Hershkovits, E. L. I., Williams, L. D., Keating, K. S., Pyle, A. M., Micallef, D., Westbrook, J., and Berman, H. M. (2008). RNA backbone : Consensus all-angle conformers and modular string nomenclature ( an RNA Ontology Consortium contribution ). *Society*, 14:465–481.
- Rivas, E. and Eddy, S. R. (2000). Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, 16(7):583–605.
- Rivas, E. and Eddy, S. R. (2001). Noncoding RNA gene detection using comparative sequence analysis. *BMC bioinformatics*, 2:8.
- Rother, M., Rother, K., Puton, T., and Bujnicki, J. M. (2011). ModeRNA: a tool for comparative modeling of RNA 3D structure. *Nucleic acids research*, pages 1–16.
- Saenger, W. (1984). *Principles of nucleic acid structure*. Springer advanced texts in chemistry. Springer-Verlag.
- Sarver, M., Zirbel, C. L., Stombaugh, J., Mokdad, A., and Leontis, N. B. (2008). FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *Journal of mathematical biology*, 56(1-2):215–52.
- Schattner, P., Decatur, W. a., Davis, C. a., Ares, M., Fournier, M. J., and Lowe, T. M. (2004). Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the *Saccharomyces cerevisiae* genome. *Nucleic acids research*, 32(14):4281–96.

- Schultes, E. A. and Bartel, D. P. (2000). One Sequence , Two Ribozymes : Implications for the Emergence of New Ribozyme Folds. *Science*, 289:448–452.
- Schweisguth, D. C. and Moore, P. B. (1997). On the conformation of the anticodon loops of initiator and elongator methionine tRNAs. *Journal of molecular biology*, 267(3):505–19.
- Shah, A. and Woolf, P. (2009). Python Environment for Bayesian Learning: Inferring the Structure of Bayesian Networks from Knowledge and Data. *Journal of machine learning research : JMLR*, 10:159–162.
- Shakhnovich, E. and Gutin, A. (1990). Enumeration of all compact conformation. of copolymers with random sequence of links. *Journal of Chemical Physics*, 93(8):5967–5971.
- Sharma, S., Ding, F., and Dokholyan, N. V. (2008). iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics (Oxford, England)*, 24(17):1951–2.
- Sherman, D., Durrens, P., Beyne, E., and Nikolski, M. (2004). Génolevures : comparative genomics and molecular evolution of hemiascomycetous yeasts. *Nucleic acids research*, 32(Database Issue):D315–D318.
- Siew, N., Elofsson, a., Rychlewski, L., and Fischer, D. (2000). MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics (Oxford, England)*, 16(9):776–85.
- Souciet, J., Aigle, M., Artiguenave, F., Blandin, G., Bolotin-Fukuhara, M., Bon, E., Brottier, P., Casaregola, S., de Montigny, J., Dujon, B., Durrens, P., Gaillardin, C., Léplinge, a., Llorente, B., Malpertuy, a., Neuvéglise, C., Ozier-Kalogéropoulos, O., Potier, S., Saurin, W., Tekaiia, F., Toffano-Nioche, C., Wésolowski-Louvel, M., Wincker, P., and Weissenbach, J. (2000). Genomic exploration of the hemiascomycetous yeasts: 1. A set of yeast species for molecular evolution studies. *FEBS letters*, 487(1):3–12.
- Souciet, J.-L., Dujon, B., Gaillardin, C., Johnston, M., Baret, P. V., Cliften, P., Sherman, D. J., Weissenbach, J., Westhof, E., Wincker, P., Jubin, C., Poulain, J., Barbe, V., Segurens, B., Artiguenave, F., Anthouard, V., Vacherie, B., Val, M.-E., Fulton, R. S., Minx, P., Wilson, R., Durrens, P., Jean, G., Marck, C., Martin, T., Nikolski, M., Rolland, T., Seret, M.-L., Casaregola, S., Despons, L., Fairhead, C., Fischer, G., Lafontaine, I., Leh, V., Lemaire, M., de Montigny, J., Neuvéglise, C., Thierry, a., Blanc-Lenfle, I., Bleykasten, C., Diffels, J., Fritsch, E., Frangeul, L., Goeffon, a., Jauniaux, N., Kachouri-Lafond, R., Payen, C., Potier, S., Pribylova, L., Ozanne, C., Richard, G.-F., Sacerdot, C.,

- Straub, M.-L., and Talla, E. (2009). Comparative genomics of protoploid Saccharomycetaceae. *Genome Research*, 19(10):1696–1709.
- Sussman, J. L., Holbrook, S. R., Warrant, R. W., Church, G. M., and Kim, S. H. (1978). Crystal structure of yeast phenylalanine transfer RNA. I. Crystallographic refinement. *Journal of molecular biology*, 123(4):607–30.
- Szewczak, a. a., Moore, P. B., Chang, Y. L., and Wool, I. G. (1993). The conformation of the sarcin/ricin loop from 28S ribosomal RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 90(20):9581–5.
- Teichmann, S. A. and Veitia, R. A. (2004). Genes encoding subunits of stable complexes are clustered on the yeast chromosomes: an interpretation from a dosage balance perspective. *Genetics*, 167(4):2121–5.
- Thore, S., Leibundgut, M., and Ban, N. (2006). Structure of the Eukaryotic Thiamine Pyrophosphate Riboswitch with Its Regulatory Ligand. *Science*, 312:1208–1211.
- Tinoco, I., Borer, P. N., Dengler, B., Levine, M. D., Uhlenbeck, O. C., Crothers, D. M., and Gralla, J. (1973). Improved Estimation of Secondary Structure in Ribonucleic Acids. *Nature New Biology*, 246:40–41.
- Tinoco, I. and Bustamante, C. (1999). How RNA folds. *Journal of molecular biology*, 293(2):271–81.
- Toor, N., Keating, K. S., Taylor, S. D., and Pyle, A. M. (2008). Crystal structure of a self-spliced group II intron. *Science (New York, N.Y.)*, 320(5872):77–82.
- Tuerk, C., Gauss, P., Thermes, C., Groebe, D. R., Gayle, M., Guild, N., Stormo, G., D'Aubenton-Carafa, Y., Uhlenbeck, O. C., and Tinoco, I. (1988). CUUCGG hairpins: extraordinarily stable RNA secondary structures associated with various biochemical processes. *Proceedings of the National Academy of Sciences of the United States of America*, 85(5):1364–8.
- Venkatraman, V., Chakravarthy, P. R., and Kihara, D. (2009). Application of 3D Zernike descriptors to shape-based ligand similarity searching. *Journal of cheminformatics*, 1:19.
- Wadley, L. M. and Pyle, A. M. (2004). The identification of novel RNA structural motifs using COMPADRES: an automated approach to structural discovery. *Nucleic acids research*, 32(22):6650–9.

- Wang, X., Huan, J., Snoeyink, J. S., and Wang, W. (2007). Mining RNA Tertiary Motifs with Structure Graphs. In *19th International Conference on Scientific and Statistical Database Management*, number Ssdbm.
- Washietl, S., Hofacker, I. L., and Stadler, P. F. (2005). Fast and reliable prediction of noncoding RNAs. *Proceedings of the National Academy of Sciences of the United States of America*, 102(7):2454–9.
- Waters, L. S. and Storz, G. (2009). Regulatory RNAs in bacteria. *Cell*, 136(4):615–28.
- Weinberg, Z., Perreault, J., Meyer, M. M., and Breaker, R. R. (2009). Exceptional structured noncoding RNAs revealed by bacterial metagenome analysis. *Nature*, 462(7273):656–9.
- Weinberg, Z., Wang, J. X., Bogue, J., Yang, J., Corbino, K., Moy, R. H., and Breaker, R. R. (2010). Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome biology*, 11(3):R31.
- Wimberly, B., Varani, G., and Tinoco, I. (1993). The conformation of loop E of eukaryotic 5S ribosomal RNA. *Biochemistry*, 32(4):1078–87.
- Woese, C. R., Winker, S., and Gutell, R. R. (1990). Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". *Proceedings of the National Academy of Sciences of the United States of America*, 87(21):8467–71.
- Wong, T. N., Sosnick, T. R., and Pan, T. (2007). Folding of noncoding RNAs during transcription facilitated by pausing-induced nonnative structures. *Proceedings of the National Academy of Sciences of the United States of America*, 104(46):17995–8000.
- Xayaphoummine, a., Bucher, T., and Isambert, H. (2005). Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic acids research*, 33(Web Server issue):W605–10.
- Yang, a. S. and Honig, B. (2000). An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *Journal of molecular biology*, 301(3):665–78.
- Yang, H., Jossinet, F., Leontis, N., Chen, L., Westbrook, J., Berman, H., and Westhof, E. (2003). Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Research*, 31(13):3450–3460.

- Yao, Z., Barrick, J., Weinberg, Z., Neph, S., Breaker, R., Tompa, M., and Ruzzo, W. L. (2007). A computational pipeline for high-throughput discovery of cis-regulatory noncoding RNA in prokaryotes. *PLoS computational biology*, 3(7):e126.
- Yathindra, N. and Sundaralingam, M. (1973). Correlation between the backbone and side chain conformations in 5-nucleotides. The concept of a rigid nucleotide conformation. *Biopolymers*, 12(2):297–314.
- Zhang, Y. and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4):702–10.
- Zhong, C., Tang, H., and Zhang, S. (2010). RNAMotifScan: automatic identification of RNA structural motifs using secondary structural alignment. *Nucleic acids research*, 38(18):e176.
- Zhou, Y., Lu, C., Wu, Q.-J., Wang, Y., Sun, Z.-T., Deng, J.-C., and Zhang, Y. (2008). GISSD: Group I Intron Sequence and Structure Database. *Nucleic acids research*, 36(Database issue):D31–7.
- Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13):3406–3415.