Action Spécifique STIC-CNRS
“Modélisation et Algorithmique des Structures d’ARN”
2002-2003
Rapport final

Responsables de l’AS :
Alain Denise et Serge Dulucq

15 décembre 2003

Préliminaires

L’Action Spécifique “Modélisation et Algorithmique des Structures d’ARN” émane du RTP 41 “Bioinformatique”. Elle a été lancée à l’automne 2002 et financée à hauteur de 30500 euros. Les 5 équipes STIC participantes d’origine ont été rejointes quelques mois après par deux autres équipes et quelques chercheurs isolés.

Le but affiché de cette action était de fédérer les équipes STIC travaillant sur les problèmes de modélisation combinatoire et d’algorithmique liés à l’étude bioinformatique des molécules d’ARN afin de créer des synergies et aboutir à des recherches communes. Pour cela, deux réunions plénières ont été tenues et des visites inter-laboratoires ont été effectuées. Elles ont favorisé l’émergence de nouvelles collaborations au sein du groupe, et la consolidation des collaborations déjà existantes.

Nous présentons ici le rapport des activités scientifiques de l’action (pages 1 à 9), suivi du bilan financier (page 10) puis d’une annexe technique comprenant notamment les références des publications du groupe (pages 11 à 14). Un certain nombre de ces publications sont accessibles sur la page web de l’action : http ://www.lri.fr/∼denise/ArnStic.

1 Rapport scientifique

1.1 Introduction

Selon le dogme central de la biologie moléculaire, l’information génétique contenue dans l’ADN est transcrite en ARN messagers, puis traduite en protéines. Celles-ci sont les acteurs majoritaires du fonctionnement de la cellule. Dans ce schéma, l’ARN ne joue qu’un rôle d’intérmédiaire. Toutefois, on sait que certains ARN ne sont pas traduits en protéines :

ils ont un rôle fonctionnel en eux-mêmes. C'est le cas par exemple des ARN ribosomiques et des ARN de transfert qui jouent un rôle essentiel dans le processus de traduction. Depuis quelques années, on découvre que beaucoup plus d'ARN qu'on ne le croyait possèdent des propriétés enzymatiques, et donc jouent un rôle en tant que tels dans l'activité cellulaire [7]. Par ailleurs, on a montré que certains ARN messagers eux-mêmes ont, du fait de leur structure, un effet sur la régulation de l'expression des protéines qu'ils codent [27]. L'étude des molécules d'ARN, dans leurs aspects structurels, fonctionnels et dynamiques fait maintenant partie des problèmes phares dans le domaine de la bioinformatique.

Les équipes engagées dans l'action spécifique interviennent et interagissent sur la majorité des problèmes combinatoires et algorithmiques liés à l'étude des molécules d'ARN :

- Prédiction de structure
- Comparaison de structures
- Détection de motifs structurels
- Modélisation et génération aléatoire de structures
- Visualisation de structures

Chacun de ces sujets est développé dans la suite, après une section consacrée aux notions de base sur la structuration des molécules d'ARN

1.2 Structures d'ARN

Une molécule d'ARN est constituée d'une succession de nucléotides A, C, G et U. Cependant, contrairement à l'ADN qui est généralement dans une configuration en double brin, la plupart des molécules d'ARN ont un brin unique. Des liaisons se forment donc entre les nucléotides selon leurs affinités physico-chimiques. Les liaisons les plus courantes sont les appariements A–U et C–G, dites “Watson-Crick”. Il existe cependant plusieurs dizaines d'autres types de liaisons possibles entre les différents nucléotides [18, 17]. La molécule d'ARN adopte dans l'espace une conformation complexe qui dépend de ces interactions et qui détermine, en grande partie, la ou les fonctions de la molécule dans la cellule. Cette structure spatiale est la *structure tertiaire* de la molécule. Déterminer cette structure à partir de la séquence est un problème que l'on est encore loin de savoir résoudre à l'heure actuelle. D'un autre côté, on a déterminé expérimentalement, par cristallographie, la structure tertiaire de quelques dizaines de molécules d'ARN, et de là on sait déduire les liaisons existantes entre les nucléotides [17]. On modélise la topologie d'une structure par un graphe dont les sommets sont les nucléotides. Les liaisons sont représentées par des arcs ou des arêtes, selon qu'elles sont orientées ou non. Les liaisons formant la séquence primaire sont orientées, elles définissent un chemin hamiltonien dans le graphe. Les liaisons Watson-Crick ne sont pas orientées.

Une *structure secondaire* d'une séquence d'ARN est une représentation partielle de sa structure tertiaire. Généralement on convient qu'une structure secondaire ne contient que les liaisons de la séquence primaire et les liaisons Watson-Crick, et que son graphe est planaire. On dit qu'une structure secondaire ne contient pas de *pseudo-noeud* si elle peut être dessinée sur un plan en positionnant toutes les arêtes d'un même côté du chemin hamiltonien sans que deux arêtes ne se croisent. Une structure secondaire a l'avantage d'être bien plus aisément manipulable que la structure tertiaire, tout en contenant une quantité d'informations suffisante pour un certain nombre de traitements. Actuellement, les algorithmes de prédiction de structure se bornent à essayer de déterminer la structure secondaire.

1.3 Prédiction de structures

Le premier algorithme de prédiction d'une structure secondaire *ab initio*, c'est-à-dire sans autre donnée que la séquence, a été développé par Nussinov *et al.* [22] en 1978. Il s'agissait d'un algorithme de programmation dynamique visant à sélectionner la structure comportant le maximum d'appariements. La plupart des algorithmes qui ont suivi sont basés sur le même principe de programmation dynamique ; cependant la fonction à optimiser, plus réaliste, est basée sur un calcul d'énergie de la structure, approximée à partir de données expérimentales. L'algorithme le plus utilisé aujourd'hui est probablement celui de Zuker [38, 36] ; il est notamment implémenté dans les logiciels *mfold* [37] et *The Vienna RNA package* [14]. Fabrice Lefebvre a montré que le concept de grammaire formelle donne une vision unificatrice des algorithmes de prédiction de structure [16]. Tous ces algorithmes prédisent des structures secondaires sans pseudo-nœuds. D'un autre côté, qulques auteurs ont présenté des algorithmes de prédiction de structures avec pseudo-nœuds, comme Akutsu [1] et Rivas et Eddy [24, 25], ces derniers s'inspirant du formalisme grammatical de Lefebvre. Bien que ces algorithmes témoignent d'une avancée théorique, ils sont en pratique inapplicables du fait de leur complexité très forte (bien que polynomiale) et du peu de données expérimentales concernant l'énergie des structures complexes (comportant des pseudo-nœuds). Les algorithmes qui ne prennent pas les pseudo-nœuds en compte ont une meilleure complexité, mais leurs résultats sont encore loin d'être réellement satisfaisants. Ainsi, une étude effectuée en 1999 montre que l'algorithme de Zuker retrouve en moyenne seulement 73% des appariements avérés sur les ARN de moins de 700 bases, et vraisemblablement encore moins au-delà [20]. Les recherches visant à améliorer les prédictions sont très actives. L'équipe du LIX a récemment développé un algorithme de prédiction ayant pour principe la recherche de structures *super-optimales* [33]. Il existe d'autres types d'approches pour ce même problème ; notamment, en France, les travaux de Christine Gaspin [11] et ceux d'Hervé Isambert [15].

Une problématique un peu différente consiste à déterminer la structure commune à plusieurs séquences dont on suppose qu'elles ont des conformations similaires parce qu'elles sont homologues. La prise en compte de ces informations supplémentaires permet bien sûr de faire des prédictions plus fiables que dans l'approche précédente. D'un autre côté, on ne dispose pas toujours d'ensembles de séquences homologues ; et celles-ci doivent être "bien choisies", de façon à posséder effectivement une structure commune tout en étant assez différentes pour que l'inférence de structure soit possible. L'équipe du LaMI travaille sur une telle approche pour la prédiction de structures avec pseudo-nœuds [29, 28] et développe les logiciels DCFold et P-DCFold. On développe au LIFL des algorithmes de prédiction qui donnent des résultats satisfaisants avec un petit nombre de séquences seulement, à partir de deux séquences [23]. Ces travaux sont implémentés dans le logiciel CARNAC.

1.4 Comparaison de structures

La problématique de la comparaison est liée à celle de la prédiction, dans le contexte de la deuxième problématique ci-dessus : savoir comparer finement des structures peut permettre de les prédire plus précisément. Cependant l'application principale de la comparaison de structures réside dans le domaine de la phylogénie moléculaire, où l'on reconstitue "l'histoire du vivant" à partir de distances évolutives calculées entre des séquences homologues.

Dans ce cadre, les structures secondaires sont généralement modélisées par des arbres étiquetés [26, 35], et les distances entre deux ou plusieurs arbres sont calculées par des algorithmes de programmation dynamique. Au sein du groupe, cinq équipes travaillent sur ce thème. D'un point de vue strictement algorithmique, les équipes du LaBRI et du LIFL ont étudié de manière très précise la complexité moyenne des algorithmes classiquement utilisés pour la comparaison d'arbres [9, 10]. D'autre part, l'équipe de l'IRIN a démontré la NP-complétude d'un problème de comparaison [5], répondant ainsi à un problème ouvert posé par Wang et Zhang en 2001 [34].

Le groupe travaille aussi dans le but de rendre les algorithmes et les représentations des données plus proches de la réalité biologique. Les équipes du LaBRI et du LIFL ont proposé des variantes des algorithmes classiques, conçues de façon à tenir compte de plus de contraintes biologiques que les algorithmes classiques [8, 30]. Les équipes de Lyon/Marne-la-Vallée et de Bordeaux modélisent les structures simultanément selon plusieurs "niveaux de granularité", pour améliorer à la fois la vitesse et la précision des algorithmes. Un logiciel, MiGAL, est en cours de développement à Marne-la-Vallée. Parallèlement, un travail sur la conception de matrices de substitution spécifiques aux structures d'ARN a été entamé dans une collaboration entre Bordeaux et les deux laboratoires d'Orsay (IGM et LRI). Ce travail permettra de pondérer plus finement les opérations d'édition de structures utilisées dans les algorithmes.

1.5 Détection de motifs structurels

Cette thématique regroupe en fait deux types de travaux différents, selon que l'on cherche des motifs dans des séquences ou dans des structures.

Dans le premier cas, il s'agit de rechercher, dans une séquence d'ARN, un motif qui est susceptible de se replier similairement à une structure d'ARN donnée. Ce problème a pour application la recherche de structures liées à une fonction particulière. Des heuristiques générales ont été développées et implémentées, par exemple dans PALINGOL [4] et dans RNAMot [19]. Le groupe travaille sur cette problématique à la fois d'un point de vue théorique et d'un point de vue applicatif. Pendant sa thèse au LIAFA, Stéphane Vialette (LRI) a démontré précisément que, selon les contraintes que l'on se donne sur les motifs, le problème de recherche est polynomial ou NP-complet [32]. Des prolongements à ce travail sont en cours d'étude, en collaboration entre le LRI et l'IRIN. D'un autre côté, on développe au LRI et à l'IGM d'Orsay une méthode de prédiction de gènes soumis au "'décalage de traduction en -1" [3]. Ce travail fait appel à la recherche de motifs structurels particuliers dans les séquences d'ARN messagers.

Dans le second cas, la recherche de motifs s'effectue non pas dans une séquence mais dans une structure d'ARN déjà connue. Celle-ci est modélisée par un graphe. Ce type de problème est étudié depuis quelques années à Montréal dans l'équipe de François Major, en collaboration avec Daniel Gautheret [12, 13]. Dans la lignée de ces travaux, le LRI a entamé en 2003 une collaboration avec ces deux équipes sur la recherche de motifs *signifiants* dans une structure d'ARN, dans un esprit similaire à la problématique très féconde de la recherche de motifs sur- ou sous-représentés dans les séquences d'ADN. Dans notre cas, les motifs sont des sous graphes de graphes d'ARN, ce qui rend le problème bien plus complexe que dans les séquences.

1.6 Modélisation et génération aléatoire de structures

Les structures secondaires sans pseudo-noeuds sont des structures assez simples pour pouvoir être modélisées par les mots d'un langage algébrique proche du langage de Dyck. Cette représentation fut utilisée dans les années 1980 pour résoudre des problèmes de dénombrement par Vauchaussade et Viennot [31] et a été récemment reprise par Nebel [21] Naturellement, ces mêmes structures secondaires peuvent aussi être représentées par des arbres. D'autre part, les structures tertiaires sont modélisées, on l'a vu, par des graphes hamiltoniens. Ces divers modèles sont utilisés et examinés dans le cadre de l'algorithmique des structures. Ainsi par exemple, diverses variantes de modélisation par des arbres, à plusieurs niveaux de précision, sont étudiées dans le cadre de la comparaison des structures secondaires.

Nous nous intéressons aussi à la conception de modèles de structures aléatoires, généralisant ainsi la démarche couramment adoptée dans le cadre de l'analyse des séquences. Les séquences aléatoires sont couramment utilisées pour aider à la découverte de nouvelles propriétés, en comparant les propriétés de séquences aléatoires avec des propriétés observées sur des séquences réelles. Des différences observées, on déduit des faits biologiques. En quelque sorte, les séquences aléatoires représentent le “bruit de fond” à partir duquel, par contraste, on cherche à déterminer quels sont les éléments biologiquement pertinents dans les séquences réelles. Pour concevoir des structures aléatoires adéquates, il faut ajouter des paramètres structurels aux paramètres statistiques couramment utilisés pour les séquences. Le concept de grammaire algébrique pondérée a été développé dans ce but au LRI et à l'IGM Orsay [6], et le logiciel GenRGenS est fondé en partie sur ce travail. Des structures d'ARN aléatoires sont destinées à être employées au sein du groupe d'une part pour étalonner et évaluer les algorithmes de comparaison (section 1.4), d'autre part pour déterminer quels sont les motifs structurels signifiants dans les structures d'ARN (section 1.5). Dans ce dernier cas, la modélisation doit aller au-delà du cadre des langages algébriques.

1.7 Visualisation de structures

Il existe un certain nombre de logiciels de visualisation de structures d'ARN. Ceux qui concernent les structures tertiaires prennent en entrée les coordonnées atomiques des structures et les reproduisent fidèlement ; ils n'ont pas de difficulté particulière à résoudre d'un point de vue strictement algorithmique. En revanche, la représentation des structures secondaires (avec ou sans pseudo-noeuds) de façon claire et esthétique est un problème délicat, comme la plupart des problèmes de dessin de graphes. Il ne s'agit pas nécessairement de représenter une structure de façon "réaliste", mais plutôt de façon intelligible et manipulable. A ce jour, aucun programme existant ne donne pleinement satisfaction. De plus, les développements du groupe créent des besoins spécifiques dans le domaine. On aimerait en effet visualiser non seulement une structure d'ARN, mais aussi, par exemple, le résultat de la comparaison de deux ou plusieurs structures.

Des travaux dans ce sens sont réalisés au LaBRI dans le cadre du projet Tulip qui vise généralement à fournir des outils de visualisation de graphes [2]. Depuis plus récemment, le LRI se penche sur le problème de la représentation de structures avec pseudo-noeuds, en utilisant une approche "partiellement tridimensionnelle".

1.8 Conclusion et perspectives

Cette action spécifique a rassemblé des équipes qui travaillaient sur l'ARN, généralement depuis peu de temps. Le volume des publications et réalisations logicielles témoigne du dynamisme de chacune des équipes dans le domaine. Elles bénéficient de contacts et collaborations sur le sujet avec d'autres groupes en France, en informatique, en biologie et en mathématiques, et à l'étranger : universités de Montréal (Canada), d'Uppsala (Suède), de Trente (Italie) et le groupe de recherche de Celera Genomics aux (Etats-Unis).

L'action a eu un effet très positif en ce qui concerne les synergies qui se sont créées au sein du groupe. On note ainsi que des liens se sont créés ou ont été renforcés grâce à l'action. Les collaborations entre LaBRI, LIFL, BBE, IGM, IRIN et LRI autour de la comparaison de structures en sont une belle illustration. Elles se concrétisent notamment par une thèse en cotutelle qui vient de commencer entre le LaBRI et le LRI.

Une communauté active et soudée s'est créée dans le département STIC et tous les partenaires de l'action sont extrêmement favorables à ce que le groupe soit pérennisé d'une façon ou d'une autre. Nous pensons qu'il pourrait certainement bénéficier d'un élargissement à d'autres laboratoires, et nous prenons des contacts dans ce sens. Nous pensons à des partenaires en informatique et en bioinformatique (INRIA et INRA), en mathématiques et en physique, mais aussi et surtout en direction des biologistes, particulièrement des biologistes expérimentalistes spécialistes de l'ARN et des phénomènes qui le touchent. Il est en effet nécessaire d'une part de confronter nos modèles et résultats à la réalité biologique, d'autre part de porter nos concepts et outils à la disposition de la communauté des biologistes. Nous visons un élargissement peut-être national dans un premier temps, mais envisageons de le généraliser, à terme, à l'Europe voire plus.

Références

[1] T. Akutsu. Dynamic programming algorithms for rna secondary structure prediction with pseudoknots. *Discrete Applied Mathematics*, 104, 2000.

[2] D. Auber. *Springer Book on Graph Drawing Software*, chapter Tulip - A huge graph visualization software, page To appear. Mathematics and visualization series. Springer, 2003.

[3] M. Bekaert, L. Bidou, A. Denise, G. Duchateau-Nguyen, J.-P. Forest, C. Froidevaux, I. Hatin, J.-P. Rousset, and M. Termier. Towards a computational model for eukaryotic -1 frameshifting sites. *Bioinformatics*, 19(327–335), 2003.

[4] B. Billoud, M. Kontic, and A. Viari. Palingol, a descriptive programming language to describe nucleic acid's secondary structure and scan sequence databases. *Nucleic Acids Research*, 8 :1395–1404, 24.

[5] G. Blin, G. Fertin, I. Rusu, and C. Sinoquet. RNA sequences and the EDIT(NESTED,NESTED) problem. Rapport de recherche RR-IRIN-03.07, IRIN, 2003.

[6] A. Denise, O. Roques, and M. Termier. Random generation of words of context-free languages according to the frequencies of letters. In D. Gardy and A. Mokkadem, editors, *Mathematics and Computer Science : Algorithms, Trees, Combinatorics and probabilities*, Trends in Mathematics, pages 113–125. Birkhaüser, 2000.

[7] C. Dennis. Small RNAs, the genome's guiding hand ? *Nature*, 420 :732, 2002.

[8] S. Dulucq and L. Tichit. A propos de la comparaison de structures secondaires d'ARN. In *JOBIM 2001*, pages 125–134, 2001.

[9] S. Dulucq and L. Tichit. RNA secondary structure comparison : exact analysis of the zhang-shasha tree edit algorithm. *Theoretical Computer Science*, 306 :471–484, 2003.

[10] S. Dulucq and H. Touzet. Analysis of tree edit distance algorithms. In *14th Annual Symposium on Combinatorial Pattern Matching (CPM 2003)*, volume 2676, pages 83–95. Lecture Notes in Computer Science, 2003.

[11] C. Gaspin. RNA secondary structure determination and representation based on constraints satisfaction. *Constraints*, 6 :201–221, 2001.

[12] P. Gendron, D. Gautheret, and F. Major. Structural ribonucleic acid motif identification and classification. In J. Schaeffer, editor, *High Performance Computing Systems and Applications*, pages 323–331. Kluwer Academic Press, 1998.

[13] P. Gendron, S. Lemieux, and F. Major. Quantitative analysis of nucleic acid three-dimensional structures. *Journal of Molecular Biology*, 308(5) :919–936, 2001.

[14] I.L. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Research*, 31(13) :3429–3431, 2003.

[15] H. Isambert and E.D. Siggia. Modeling rna folding paths with pseudoknots : Application to hepatitis delta virus ribozyme. *Proc. Natl. Acad. Sci. USA*, 97(12) :6515, 2000.

[16] F. Lefebvre. A grammar-based unification of several alignment and folding algorithms. In D.J. States, P. Agarwal, T. Gaasterland, L. Hunter, and R.F. Smith, editors, *ISMB'96*, pages 143–154. AAAI press, 1996.

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

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

[19] T.J. Macke, D.J. Ecker, R.R. Gutell, D. Gautheret, D.A. Case, and R. Sampath. Rnamotif-an rna secondary structure definition and search algorithm. *Nucleic Acids Research*, 2001.

[20] D.H. Mathews, J. Sabina, M. Zuker, and D.H. Turner. Expanded sequence dependence of thermodynamic parameters improves prdiction of RNA secondary structure. *Journal of Molecular Biology*, 288 :911–940, 1999.

[21] M.E. Nebel. Combinatorial properties of RNA secondary structures. *Journal of Computational Biology*, 9(3) :541–574, 2002.

[22] R. Nussinov, G. Pieczenic, J.R. Griggs, and D.J. Kleitman. Algorithms for loop matchings. *SIAM Journal of Applied Mathematics*, 35 :68–82, 1978.

[23] O. Perriquet, H. Touzet, and M. Dauchet. Finding the common structure shared by two homologous rnas. *Bioinformatics*, 19 :108–116, 2003.

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

[25] E. Rivas and S.R. Eddy. The language of RNA : a formal grammar that includes pseudoknots. *Bioinformatics*, 16(4) :334–340, 2000.

[26] B.A. Shapiro. An algorithm for comparing multiple RNA secondary structures. *CABIOS*, 4 :387–393, 1988.

[27] G. Stahl and J.-P. Rousset. Les surprises du décodage de l'information génétique. *Médecine/Sciences*, 15 :1118–1125, 1999.

[28] F. Tahi, S. Engelen, and M. Régnier. A fast algorithm for RNA secondary structure prediction including pseudoknots. In *3rd IEEE Symposium on Bioinformatics and Bioengineering (BIBE 2003)*, pages 11–17, March 2003.

[29] F. Tahi, M. Gouy, and M. Régnier. Automatic rna secondary structure prediction with a comparative approach. *Computers and Chemistry*, 26(5) :521–530, 2002.

[30] H. Touzet. Tree edit distances with gaps. *Information Processing Letters*, 85(3) :123–129, 2003.

[31] M. Vauchaussade de Chaumont and X.G. Viennot. Enumeration of RNA's secondary structures by complexity. In V. Capasso, E. Grosso, and S.L. Paven-Fontana, editors, *Mathematics in Medecine and Biology*, volume 57 of *Lecture Notes in Biomathematics*, pages 360–365, 1985.

[32] S. Vialette. Pattern matching problems over 2-interval sets. In *Combinatorial Pattern Matching, 13th Annual Symposium, CPM 2002, Fukuoka, Japan, July 3-5, 2002, Proceedings*, volume 2373 of *Lecture Notes in Computer Science*, pages 53–63. Springer, 2002.

[33] J. Waldispühl, B. Behzadi, and J.-M. Steyaert. An approximate matching algorithm for finding (sub-)optimal sequences in s-attributed grammars. *Bioinformatics*, 18 :250–259, 2002.

[34] Z. Wang and K. Zhang. Alignment between two rna structures. In *MFCS 2001*, pages 690–702, 2001.

[35] K. Zhang and D. Shasha. Simple fast algorithms for the editing distance between trees and related problems. *SIAM Journal on Computing*, 1989.

[36] M. Zuker. Computer prediction of RNA structure. *Methods in Enzymology*, 180 :262–288, 1989.

[37] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13) :3406–3415, 2003.

[38] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9 :133–148, 1989.

2 Rapport financier

La somme allouée à l'AS, d'un montant de 30500 euros, a été partagée par le CNRS entre les cinq équipes présentes lors de la création du groupe (BBE, LaBRI, LaMI, LIFL, LRI) à raison de 6100 euros par équipe. Elle a été utilisée pour les raisons suivantes (le pourcentage indique la fraction du budget utilisée pour chaque rubrique).

- Frais lors des réunions et visites inter-laboratoires : 17%
- Déplacements lors de colloques internationaux et nationaux : 55%
- Invitations de chercheurs : 8%
- Gratification de stagiaires : 10%
- Livres et fournitures : 10%

3 Annexe technique

3.1 Site internet

http ://www.lri.fr/∼denise/ArnStic

3.2 Participants

– BBE, UMR 5558 (SdV et STIC) , Lyon : Julien Allali (Doct, IGM Marne-la-Vallée), Marie-France Sagot (DR INRIA), Marina Zelwer (Doct).
– IRIN, UPRESA 2157, Nantes : Guillaume Blin (Doct.), Guillaume Fertin (MdC), Irena Rusu (Pr).
– LaBRI, UMR 5800 (STIC), Bordeaux : David Auber (MdC), Serge Dulucq (Pr), Isabelle Dutour (MdC), Pascal Ferraro (MdC), Claire Herrbach (Doct, et LRI), Laurent Tichit (ATER).
– LaMI, UMR 8042 (STIC), Evry : Stefan Engelen (Doct), Fariza Tahi (MdC).
– LIFL, UMR 8022 (STIC), Lille : Olivier Perriquet (Doct), Hélène Touzet (MdC).
– LIX, UMR 7650 (STIC), Palaiseau : Behshad Behzadi (Doct), Jérôme Waldispühl (Doct), Jean-Marc Steyaert (Pr).
– LRI, UMR 8623 (STIC), Orsay : Patrick Amar (MdC), Alain Denise (Pr), Jean-Paul Forest (Doct), Christine Froidevaux (Pr), Claire Herrbach (Doct, et LaBRI), Yann Ponty (Doct), Romain Rivière Doct), Stéphane Vialette (MdC).
– Autres participants : Dominique Barth (Pr, PRISM Versailles), Michel Termier (MdC, IGM Orsay).

3.3 Programme des réunions plénières du groupe

Les transparents des exposés sont visibles sur le site de l'action.

Première réunion : mercredi 19 et jeudi 20 mars 2003 au LRI Orsay.
– Guillaume Blin (IRIN). Calcul de la distance d'édition entre deux structures secondaires d'ARN : complexité du problème Nested/Nested.
– Alain Denise et Yann Ponty (LRI). Génération aléatoire de structures secondaires d'ARN.
– Stefan Engelen et Fariza Tahi (LAMI). DCFold, un logiciel pour la prédiction de structures secondaires des ARN incluant les pseudo-nœuds.
– Pascal Ferraro et Laurent Tichit (LaBRI). Comparaison d'arborescences *multi-échelles pour l'évaluation de la ressemblance entre structures secondaires d'ARN.
– Jean-Paul Forest (LRI). Recherche de structures d'ARN caractéristiques des sites de frameshift.
– Claire Herrbach (IGM/LRI) et Michel Termier (IGM). Vers la construction de matrices de substitution pour la comparaison de structures secondaires d'ARN.
– Olivier Perriquet (LIFL). Prédiction de la structure secondaire à l'aide d'un petit nombre d'ARNs homologues.

– Hélène Touzet (LIFL). Distances d'arbres pour la comparaison de structures d'ARN.
– Jérôme Waldispühl (LIX). Prédiction de structures secondaires d'ARN avec des grammaires S-attribuées.

Seconde réunion : jeudi 9 novembre 2003 au LaBRI, Bordeaux.

– Julien Allali (IGM Marne-La-Vallée). MiGaL : modélisation et comparaison de structures secondaires d'ARN.
– Patrick Amar (LRI). Exposé prospectif sur la visualisation et la manipulation de structures d'ARN.
– Romain Rivière (LRI). Recherche de motifs signifiants dans une structure d'ARN.
– Stéphane Vialette (LRI). Modélisation des structures secondaires d'ARN par 2-intervalles.
– David Auber (LaBRI). Visualisation de structures secondaires d'ARN.

3.4 Diffusion des travaux (en 2002 et 2003)

3.4.1 Revues internationales à comité de lecture

Michaël Bekaert, Laure Bidou, Alain Denise, Guillemette Duchateau-Nguyen, Jean-Paul Forest, Christine Froidevaux, Isabelle Hatin, Jean-Pierre Rousset and Michel Termier. Towards a computational model for eukaryotic -1 frameshifting sites. Bioinformatics 19 (2003) 327-335.

Serge Dulucq and Laurent Tichit. RNA secondary structure comparison : exact analysis of the Zhang-Shasha tree edit algorithm. Theoretical Computer Science 306 no 1-3, pp 471–484, 2003

Olivier Perriquet, Hélène Touzet and Max Dauchet. Finding the common structure shared by two homologous RNAs. Bioinformatics 19, pp 108-116, 2003.

Fariza Tahi, Manolo Gouy and Mireille Régnier. Automatic RNA secondary structure prediction with a comparative approach. Computers and Chemistry, vol. 26, no 5, 2002. pages 521-530.

Hélène Touzet. Tree edit distance with gaps. Information Processing Letters 85 (3), pp 123 - 129, 2003.

Jérôme Waldispühl, Behshad Behzadi and Jean-Marc Steyaert. An Approximate Matching Algorithm for Finding (Sub-)Optimal Sequences in S-attributed Grammars. Bioinformatics 18 (2002) 250-259.

3.4.2 Actes de colloques internationaux à comité de sélection

M. Bekaert, J.P. Forest, L. Bidou, A. Denise, G. Duchateau Nguyen, C. Fabret, C. Froidevaux, I. Hatin, J.-P. Rousset and M. Termier. Searching the *Saccharomyces cerevisiae* genome for -1 frameshifting sites. Proc. ECCB 2003, (European Conference on Computational Biology), September 2003, Paris, pp 4-6.

Serge Dulucq and Hélène Touzet. Analysis of tree edit distance algorithms. Fourteenth Annual Symposium on Combinatorial Pattern Matching (CPM 2003), Mexico, 2003, Lecture Notes in Computer Science, 2676, pp 83–95.

Fariza Tahi, Stefan Engelen and Mireille Régnier. A fast algorithm for RNA secondary structure prediction including pseudoknots. 3rd IEEE Symposium on Bioinformatics and Bioengineering (BIBE 2003), March 10-12 2003, Bethesda, Maryland. In IEEE Computer Society proceedings, pages 11-17.

Stéphane Vialette. Pattern Matching Problems over 2-Interval Sets. In proceedings of Combinatorial Pattern Matching, 13th Annual Symposium, CPM 2002, Fukuoka, Japan, July 3-5, 2002. Lecture Notes in Computer Science 2373, 53-63

3.4.3 Rapport de recherche

Guillaume Blin, Guillaume Fertin, Irena Rusu et Christine Sinoquet. RNA Sequences and the EDIT(NESTED,NESTED) Problem. Rapport de Recherche RR-IRIN-03.07 de l'IRIN, soumis à publication (2003).

3.4.4 Posters

Guillaume Blin, Guillaume Fertin et Stéphane Vialette. A polynomial algorithm for the 2-interval pattern problem. Poster présenté au 3rd Workshop on Algorithms in Bioinformatics (WABI 2003), Septembre 2003.

Alain Denise, Yann Ponty and Michel Termier. Random generation of structured genomic sequences. Poster présenté à RECOMB 2003, April 2003, Berlin.

Yann Ponty. GenRGenS : génération aléatoire de séquences génomiques. Poster presenté à JOBIM 2002, Juin 2002, Sait-Malo.

F. Tahi, S. Engelen, M. Régnier. P-DCFold : an algorithm for RNA secondary structure prediction including all kinds of pseudoknots. Poster présenté à ECCB 2003, European Conference on Computational Biology, 27-30 septembre 2003, Paris, France.

F. Tahi, S. Engelen, M. Régnier. P-DCFold, an algorithm for RNA secondary structure prediction including all kinds of pseudoknots. Poster présenté à RECOMB 2003, International Conference on Computational Biolology, 10-13 avril 2003, Berlin, Allemagne. Currents in Computational Molecular Biology 2003, pages 363-364, 2003.

3.4.5 Thèses soutenues

Olivier Perriquet. Algorithmes pour la prediction de structures d'ARN. Thèse de l'université de Lille, soutenue en décembre 2003. Jury : Max Dauchet, Serge Dulucq (rapporteur), Daniel Gautheret (rapporteur), Rémi Gilleron, Marie-France Sagot, Hélène Touzet (directeur).

Laurent Tichit. Algorithmique des structures biologiques : l'édition d'arborescences pour la comparaison de structures secondaires d'ARN. thèse de l'université Bordeaux I, soutenue le 19 septembre 2003. Jury : Maylis Delest, Alain Denise (rapporteur), Serge Dulucq (directeur), Jean-Marc Fédou (rapporteur), Laurent Ferraro, Michel Termier.

3.5 Logiciels

CARNAC (Computer Alignment of RNA by Cofolding). CARNAC est une nouvelle méthode pour la prédiction de la structure secondaire d'une famille d'ARN homologues. CARNAC fonctionne à partir de deux séquences. Il prend en entrée les séquences non alignées et produit une structure consensus par séquence. L'algorithme s'applique aussi bien à de longues séquences bien conservées (16S ssu rRNA), qu'à des familles plus hétérogènes (RNase P).

GenRGenS (Generation of Random Genomic Sequences and Structures). Un outil de modélisation et de génération aléatoire de séquences et de structures génomiques, selon des formalismes statistiques (chaînes de Markov) et grammaticaux (grammaires algébriques pondérées). Permet d'engendrer notamment des structures secondaires d'ARN aléatoires.

MiGAL. Comparaison de structures secondaires d'ARN.

DCFold (Divid and Conquer Fold) et *P-DCFold* (Pseudoknots-DCfold). DCfold est un logiciel de prédiction de structure secondaire des ARN implémentant la méthode comparative. A partir d'alignement de séquences d'ARN homologues aux formats Fasta ou ClustalW, on obtient en sortie un fichier au format Rnaviz. Une extension appelée P-DCfold permet en plus de prédire des structures contenant des pseudonoeuds.