

ARNack : Une base de données de structures d'ARN

Yann Ponty

20 janvier 2005

1 Introduction

Les ARNs sont des polymères linéaires, c'est à dire grossièrement des séquences de bases A, C, G et U, d'un intérêt fondamental en génomique. Leurs rôles sont multiples : parfois intermédiaires entre l'ADN et les protéines, véritables *chevilles ouvrières* de la cellule, elles peuvent aussi intervenir plus directement, comme au sein du ribosome lors de la traduction de l'ARN en protéine. Ces composés chimiques sont structurés autour d'une *colonne vertébrale* simple, contrairement à l'ADN et sa double hélice. C'est pourquoi les ARNs, lâchés dans un milieu quelconque, ont une forte propension à se replier, formant ainsi des liaisons hydrogènes entre certaines bases, qualifiées de complémentaires. Une structure tridimensionnelle apparaît alors, qui joue un rôle déterminant dans la fonction de l'ARN. Pour cette raison, il est important d'étudier la structure de l'ARN, d'en proposer des modèles, de comparer ces structures grâce à des algorithmes d'alignements d'arbres... C'est ce à quoi se consacrent les participants au thème *Analyse de séquences et structures biologiques* de l'équipe Bioinformatique du LRI. Pour ce faire, ils doivent baser, affiner puis valider leurs travaux sur des données tant purement biologiques, c'est à dire issues d'expériences en biologie *humide*, que mixtes, comme par exemple les structures issues de l'application d'algorithmes de repliements à des données issues du séquençage.

2 Une grande diversité de formats

Historiquement, les premières données de structures secondaires d'ARN furent obtenues par cristallographie, puis apparurent des structures issues d'algorithmes de repliements. Ceux ci prennent en entrée la séquence de bases, et renvoient les structures associées les plus probables dans un modèle propre à chaque algorithme.

Les premiers algorithmes relatifs à l'ARN ayant été conçus par des biologistes et des physiciens, publics non alors sensibilisés aux mérites d'une organisation rationnelle des données, chaque logiciel et chaque laboratoire pratiquant la cristallographie ou presque développa son propre format de sortie pour les données.

Récemment, un type de fichier XML : RNAML, est apparu afin d'unifier les différents formats, mais les données disponibles sur les différents sites consacrés à la structure de l'ARN sont encore présentées dans une grande variété de formats.

3 But du stage

Le but de ce stage est la réalisation et l'implantation d'une base de donnée de structures secondaires d'ARN, à usage interne. La base devra être consultable par le biais d'une interface PHP. De plus, des outils de conversion de et vers chacun des formats pris en charge (liste à définir ultérieurement) seront codés. L'idée serait de se ramener à une représentation interne canonique, RNAML, puis d'exporter le plus grand sous ensemble des données reconnues vers le format destination. De plus, ces outils de conversion devront être utilisés dans un programme interfaçant la base en y injectant le contenu, converti au préalable en XML, d'un nombre arbitraire de fichiers.

4 Outils et techniques utilisés

- HTML
- PHP
- Parsing de fichiers
- Interfaçage de base MySQL(PHP et C++)