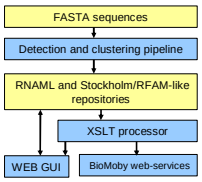
LeARN: a platform for detecting, clustering and annotating non-coding RNAs

Celine Noirot [1], Christine Gaspin[2], Thomas Schiex[2], Jérôme Gouzy [1]
[1] Laboratoire des Interactions Plantes-Microorganismes INRA/CNRS
[2] Laboratoire de Biometrie et d'Inteligence Artificielle INRA
Celine.Noirot@toulouse.inra.fr, Jerome.Gouzy@toulouse.inra.fr
http://symbiose.toulouse.inra.fr/LeARN/cgi-bin/learn.cgi

Les projets de séquençage génomiques de cette dernière décade ont conduit au développement d'outils et de bases de données nécessaires à la détection et à l'annotation fonctionnelle des gènes codant pour les protéines. Plus récemment, l'importance des gènes non codants pour des protéines (ncRNA) a été démontrée dans un très grand nombre de processus biologiques. Ces découvertes se sont accompagnées de développements bioinformatiques qui portent indépendamment sur l'amélioration des programmes de détection, la définition de stratégies d'analyses à l'échelle de génomes [1] et la mise en place de « repository » généralistes [2], spécifiques de certaines familles [3] ou dédiés à certain génomes [4].
LeARN est un logiciel qui vise à combler le manque d'outils et d'interfaces entre les logiciels de détection et les banques de données publiques consacrées aux ncRNAs. Pour atteindre cet objectif, LeARN intègre les résultats de logiciels de détection, définit automatiquement les différentes familles d'ARN et propose une interface utilisateur pour éditer et annoter les ARN et les familles d'ARN. Ainsi, le package LeARN permet de gérer la globalité du processus d'annotation semi-automatique des ncRNAs dans le cadre de projets de séquençage et d'annotation.

Architecture



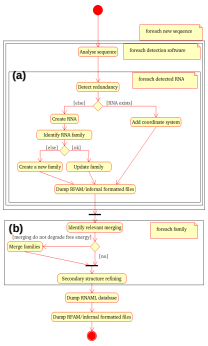
LeARN est une plateforme logicielle développée en Perl orienté objet composée de deux parties:
1) un pipeline qui analyse les fichiers de séquences, intègre les résultats et structure les connaissances dans une base de données au format RNAML [7] et Stockholm/Infernal [8]
2) une interface web développée en Perl-CGI pour la visualisation et l'annotation des ARN et des familles.
Les deux composantes de la plateforme intègre les outils standards d'analyse et de représentation des molécules d'ARNs tel `mfold` [9] et la suite logicielle Vienna [10].

Afin d'assurer la persistance, nous avons choisi de construire d'une part un « repository » de fichiers au format RNAML pour décrire les ARNs et les familles et d'autre part un « repository » de fichiers au format stockholm pour les familles. Ce choix a été guidé par la volonté d'assurer une interopérabilité native avec les formats et les systèmes existants. Ainsi RNAML est un format standard pour échanger des informations sur les ARNs et est utilisé aussi bien en entrée de programmes de visualisation (S2S [11]) qu'en sortie de programme de calcul (mfold, erpin [12]). Le choix du format stockholm et des modèles de covariance associés nous permet d'assurer la compatibilité avec la base de référence Rfam et de bénéficier des outils qui ont été développés pour l'interroger (rfam_scan).

Le moteur d'accès aux données est constitué du processeur XSLT qui permet d'interroger directement et efficacement des « repository » de fichiers XML. Le couple RNAML/XSLT compose donc notre système de gestion de base de données et nous permet de disposer d'un outil performant ne nécessitant pas d'installation lourde.

La visualisation des données est effectuée par transformation des documents RNAML grâce aux feuilles de style XSL et ce autant pour l'interface web que pour les web-services BioMoby.

Détection et « clustering »



Le pipeline d'intégration enchaîne les programmes de détection généralistes comme `rfam_scan` ou `blast` ou plus spécifiques comme `tRNAscan-SE`. Il est paramétrable et extensible via un fichier de configuration. Ainsi, si l'utilisateur dispose d'un programme d'analyse dont les sorties sont au format GFF, il peut l'intégrer au pipeline par simple modification du fichier de configuration.

(a) La première étape du pipeline contrôle l'exécution des programmes de détection, la création des ncRNA et le « clustering » en familles.
De plus, afin de générer une banque non redondante, le pipeline gère d'une part le fait que plusieurs programmes peuvent détecter le même ARN et d'autre part que la même molécule d'ARN puisse être présente sur plusieurs séquences chevauchantes (Fig 1). Le premier type de redondance est éliminé grâce au niveau de priorité de chaque programme de détection. Ainsi pour une région donnée de la séquence en cours de traitement, ne sera conservée que la région prédite par le programme considéré comme le plus fiable alors que les prédictions « chevauchantes » seront ignorées. La redondance induite par le chevauchement de séquences génomiques est gérée grâce à la notion de système de coordonnées qui permet de localiser un même ARN sur plusieurs séquences c'est-à-dire qu'à un ARN seront associés un ou plusieurs systèmes de coordonnées.
Lorsqu'un nouvel ARN est créé, nous testons son appartenance à une famille existante par l'intermédiaire du programme `rfam_scan` paramétré pour être exécuté contre l'instance en cours de la banque LeARN.

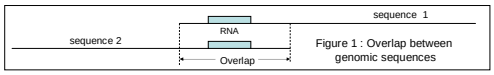


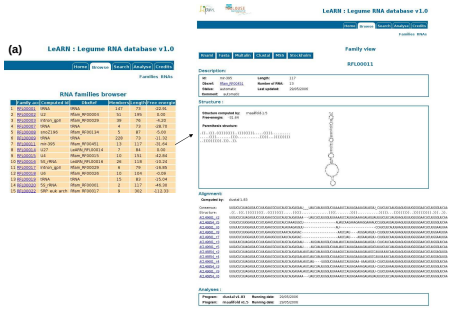
Figure 1 : Overlap between genomic sequences

(b) La deuxième partie du pipeline est un post traitement destiné à fusionner éventuellement des familles. Ce processus est nécessaire pour corriger le biais lié à la nature de l'algorithme d'intégration dont le résultat dépend de l'ordre de traitement des séquences. La fusion aura lieu lorsque l'énergie libre de la famille résultant de la fusion sera peu dégradée par rapport à celles des familles originelles.

✓ le processus d'intégration est incrémental puisque une étape de chargement en mémoire des fichiers RNAML, de la version précédente de la base, précède l'exécution du pipeline.
✓ la partie calcul peut être dissociée de la partie traitement pour être parallélisée. En effet, les programmes de détection peuvent être lancés indépendamment et en parallèle sur l'ensemble des séquences et les résultats stockés dans une arborescence qui jouera le rôle de cache. Ainsi, lorsque le processus glouton souhaitera « consommer » une analyse il cherchera d'abord dans le cache si elle est présente et ne lancera le calcul que dans le cas où le fichier attendu n'est pas disponible.

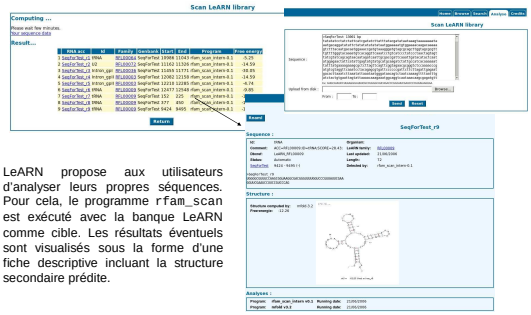
Visualisation et annotation

Navigation



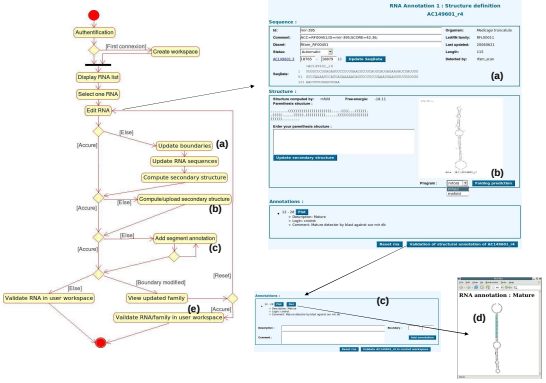
La navigation permet d'accéder aux fiches descriptives des ARNs et des familles. L'accès à ces fiches peut se faire via la liste récapitulative (a) ou via un formulaire de recherche (b).

Analyse d'une séquence



LeARN propose aux utilisateurs d'analyser leurs propres séquences. Pour cela, le programme `rfam_scan` est exécuté avec la banque LeARN comme cible. Les résultats éventuels sont visualisés sous la forme d'une fiche descriptive incluant la structure secondaire prédite.

Annotation d'un ARN



La section d'annotation permet à un utilisateur authentifié de corriger les erreurs éventuelles dues a l'annotation automatique mais également d'ajouter des annotations propres à chaque ARN.
Le processus d'annotation d'un ARN se déroule en deux étapes parfois suivies d'une troisième lorsque les bornes de l'ARN ont été modifiées.

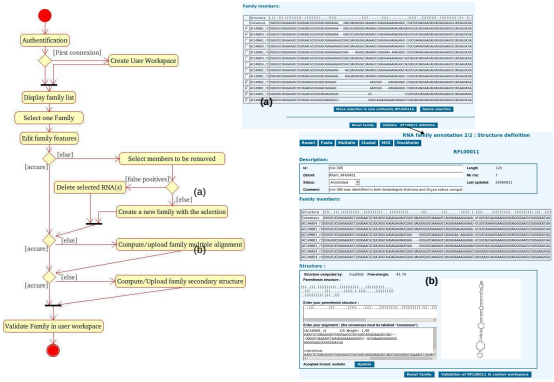
1) La première étape consiste à valider/modifier les informations générales de séquence (a) et de structure (b). Concernant la structure, l'utilisateur peut soit sélectionner un programme de prédiction intégré au système d'annotation soit saisir la structure sous un format parenthésé.

2) La deuxième étape consiste à annoter un ou plusieurs segments de l'ARN (c) avec en permanence la possibilité de visualiser cette annotation sur la molécule d'ARN (d).

3) Enfin, si les bornes de l'ARN ont été modifiées, l'alignement multiple de la famille à laquelle l'ARN appartient a aussi été modifié, ce qui implique une validation explicite de l'alignement multiple de la famille par l'annotateur (e).

Annotation d'une famille

L'annotation d'une famille se décompose en deux étapes: la première consiste dans la validation des membres de la famille (a). Ainsi un annotateur peut être emmené à supprimer certains faux positifs ou à créer une sous famille avec les membres sélectionnés. La seconde étape consiste dans la validation de l'alignement et de la structure (b). A ce niveau, il peut s'agir d'une simple validation des calculs prédéfinis ou de la saisie d'un alignement et/ou d'une structure secondaire calculés par ailleurs.



Bibliographie

[1] Mattick JS. (2003) Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays.*, 25(10):930-9
[2] Barneche F, Gaspin C, Guyot R, Echeverria M. (2001) Identification of 66 box C/D snoRNAs in Arabidopsis thaliana: extensive gene duplications generated multiple isoforms predicting new ribosomal 2'-O-methylation sites. *J Mol Biol.*, 311:57-73
[3] Bonnet E, Wuyts J, Rouze P, Van de Peer Y. (2004) Detection of 91 potential conserved plant microRNAs in Arabidopsis thaliana and Oryza sativa identifies important target genes. *Proc Natl Acad Sci U S A.*, 101:11511-6
[4] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S.R. Eddy, A. Bateman. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, 33:D121-4
[5] S. Griffiths-Jones. (2004) The microRNA Registry. *Nucleic Acids Res.*, 32:D109-11
[6] A.M. Gustafson, E. Allen, S. Givan, D. Smith, J.C Carrington, K.D. Kasschau. ASRP: the Arabidopsis Small RNA Project Database. Nucleic Acids Res., 33(Database issue):D637-40, 2005
[7] Waugh A, Gendron P, Altman R, Brown JW, Case D, Gautheret D, Harvey SC, Leontis N, Westbrook J, Westhof E, Zuker M, Major F (2002) RNAML: a standard syntax for exchanging RNA information. *RNA.* 8:707-17
[8] http://www.genetics.wustl.edu/eddy/infernal/
[9] M. Zuker. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31, 3406-15.
[10] Hofacker IL.(2003) Vienna RNA secondary structure server.*Nucleic Acids Res.* 31:3429-31.
[11] Jossinet,F. and Westhof,E..(2005) Sequence to Structure (S2S): display, manipulate and interconnect RNA data from sequence to structure. *Bioinformatics.*, 21: 3320-1.
[12] Lambert A, Fontaine JF, Legendre M, Leclerc F, Permal E, Major F, Putzer H, Delfour O, Michot B, Gautheret D. (2004) The ERPIN server: an interface to profile-based RNA motif identification. *Nucleic Acids Res.* 32:W160-5.

Application

Le pipeline LeARN a été appliqué à l'analyse de 130Mb de séquences génomiques de légumineuses (*Medicago truncatula* et *Lotus Japonicus*); à partir desquelles nous avons pu détecter 716 ARNs regroupés en 66 familles.

Disponibilité

Une diffusion du package est prévue pour l'automne 2006; les résultats obtenus sur les données génomiques « légumineuses » sont disponibles via l'interface LeARN à l'url
`http://symbiose.toulouse.inra.fr/LeARN/cgi-bin/learn.cgi`

Remerciement

Ce travail a été financé par le projet intégré « EU/FP6 Grain-Legumes»