SERPIN

(Projet de Recherche 2004)

Développement du serveur ERPIN pour l'annotation des motifs ARN

http://tagc.univ-mrs.fr/erpin/

CONTEXTE

La connaissance des ARN non codants (ARNnc) est aujourd'hui un aspect essentiel de l'annotation des génomes. En raison de la spécificité structurale de l'ARN, cette annotation ne peut se faire avec les programmes d'alignement local classiques comme BLAST. Afin de résoudre ce problème, nous avons développé le programme ERPIN, qui utilise le principe des matrices poids-position pour rechercher à la fois séquence et structure secondaire [1]. Le programme ne requiert en entrée qu'un alignement multiple de l'ARN étudié, accompagné d'une liste d'appariements. Il a démontré son efficacité dans plusieurs études, en termes de spécificité, de sensibilité et de rapidité des recherches [3-5].

Sur la base du « moteur » ERPIN, nous avons créé un serveur Web (http://tagc.univ-mrs.fr/erpin/) permettant une recherche en ligne aisée de quelques dizaines de motifs ARN dans tout fragment génomique fourni par l'utilisateur. Ce serveur dispose d'atouts importants par rapport aux méthodes concurrentes fondées sur les grammaires stochastiques, notamment en ce qui concerne la rapidité des recherches, le calcul d'une E-value et le traitement des structures en « pseudonoeud » (impossible par les grammaires).

OBJECTIFS

L'objectif du projet est de soutenir le développement du serveur ERPIN et d'asseoir sa position de ressource internationale pour l'annotation des motifs/gènes ARN. Ceci implique les étapes suivantes:

A. Offrir une évaluation statistique plus rigoureuse des résultats par un calcul amélioré de la E-value.

B. Améliorer le traitement des alignements « pauvres » contenant par exemple moins de 10 séquences, par l'utilisation de pseudo-comptes adaptés aux ARN

C. Permettre la recherche simultanée de tous les motifs définis dans la base

D. Développer de nouveaux alignements afin de couvrir la plupart des motifs ARN connus.

E. Permettre aux utilisateurs d'entrer directement un alignement sur le serveur Web afin de (i) déterminer automatiquement la meilleure région à utiliser pour une recherche ERPIN et (ii) réaliser cette recherche sur une banque fournie

RESULTATS

E-value et pseudocomptes

Le calcul de E-value a été significativement amélioré en terme de précision et de rapidité par l'adoption de produits de convolution. Cette approche permet de calculer directement la distribution des scores attendus pour tout profil sans gap à colonnes indépendantes (fig.1). D'autre part, le traitement des alignements pauvres est maintenant assuré par l'utilisation de pseudocomptes, calculés sur la base des mutations observées dans des alignements d'ARN ribosomique. La nouvelle version de Erpin (4.5) utilisant les pseudocomptes est plus sensible que les versions précédentes qui pénalisaient fortement toute déviation par rapport aux séquences d'entraitement. Les objectifs A et B ont donc été atteints, et ce travail a fait l'objet d'une publication (6).

Optimisation d'ERPIN avant objectifs C et D

Afin de réaliser les objectifs C et D, et notamment permettre la recherche de motifs de grande taille (gène complet de RNAse P ou ARNr), nous devions débloquer un verrou algorithmique dans la gestion des masques et des recherches en plusieurs étapes. En effet, dans certains cas, les positions relatives des éléments détectés à une étape *i* étaient perdues à l'étape *i* +1, entraînant des temps de calculs prohibitifs pour les longs ARN. Ce verrou est maintenant réglé et une version beaucoup plus rapide du programme est en cours de test.

Impact des services

Les statistiques d'utilisation du site ERPIN ne sont pas disponibles pour 2004. Pour 2005 le nombre de recherches effectuées (lancement d'un Run Erpin sur une séquence utilisateur avec un motif donné) s'élève à environ 20/jour (300 à 600/mois, cf statistiques mensuelles Awstat, Fig 2).

LABORATOIRES IMPLIQUES

•INSERM ERM206 TAGC, Technologies avancées pour le génome et la clinique, Marseille
•CNRS UMR 6207, CPT, Centre de Physique Théorique, Marseille
•CNRS UPR 9022, IBMC, Institut de biologie moléculaire et cellulaire, Strasbourg

MOTS CLES

ARN non-codant, annotation génomique, profils, recherche de motifs

CALENDRIER

2004	Marseille	A (E-value) B (pseudocomptes)
2005	Marseille	Recrutement CDD 1 an C (recherche simultanée) D (nouveaux alignements)
2006	Marseille-Strasbourg	E (recherche auto)

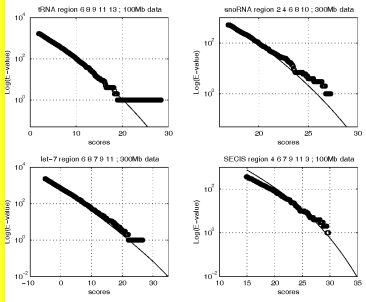

Figure 1. Comparision of computed E-values (solid lines) and number of solutions obtained from simulation on a random database of uniform nucleotide composition (circles), for different RNA motifs (ref. 6)

Total: 67 different pages-url		Viewed	Average size	Entry	Exit
/erpin/cgi-bin/results.pl	fev 2005	501	557.02 KB	4	23
/erpin/cgi-bin/results.pl	mar 2005	616	22.99 KB	1	17
/erpin/cgi-bin/results.pl	avr 2005	313	6.87 KB	2	16

Figure 2. Analyse des logs sur serveur Web ERPIN par Awstat. Accès mensuel au script « result.pl ». Chaque accès (« viewed ») correspond à une recherche de motif ARN dans une banque de séquences fournie par l'usager.

PUBLICATIONS PRINCIPALES

ANTERIEURES AU PROJET

1. Gautheret D. & Lambert A.(2001). Direct RNA definition and identification from multiple sequence alignments using secondary structure profiles. J. Mol. Biol. 313:1003-11.
2. Lambert A., Lescure A., Gautheret D. (2002). A Survey of Metazoan Selenocysteine Insertion Sequences. Biochimie, 84, 953-959.
3. Legendre, M., Gautheret D. (2003). Sequence determinants in human polyadenylation site selection. BMC genomics, 4, 7.
4. 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. Nucl. Acids Res. 32, W160-5.
5. Legendre M, Lambert A, Gautheret D. (2005) Profile-based detection of microRNA precursors in animal genomes. Bioinformatics. 2005 Apr 1;21(7):841-5.

DEPUIS DEMARRAGE

6. Computing Expectation Values for RNA Motifs Using Discrete Convolutions. BMC Bioinformatics. 2005 May 13;6(1):118