EVOLREP - Impact et évolution des répétitions dans les génomes : interactions entre innovation génétique et fonction biochimique

http://wwwabi.snv.jussieu.fr/evolrep

Coordinateur : Joël Pothier

Equipes: (1) Atelier de BioInformatique, Université Pierre et Marie Curie-Paris 6 (ABI), (2) Unité Génétique des Génomes Bactériens, URA CNRS 2171, Institut Pasteur (GGB), (3) Equipe Structure et Dynamique du Génome, Institut Jacques Monod (SDG UMR CNRS 7592), (4) Equipe Helix, INRIA Rhône-Alpes (HELIX), (5) Laboratoire d'Informatique de Paris Nord, Université Paris XIII (PNORD), (6) Laboratoire Adaptation et Pathogénie des Microorganismes, Equipe Contrôle de l'Expression Génique (CEG)

OBJECTIFS

Les duplications sont des éléments qui jouent un rôle important dans la dynamique des génomes en stimulant les évènements de recombinaison ectopiques. Elles sont une des traces de la fluidité des génomes dont l'analyse apporte le plus d'information. Chez *Saccharomyces cerevisiae*, l'étude des duplications intra-chromosomiques au niveau nucléique avait montré une sur-représentation des répétitions directes proches, espacées de moins de 1 kb (DDP).

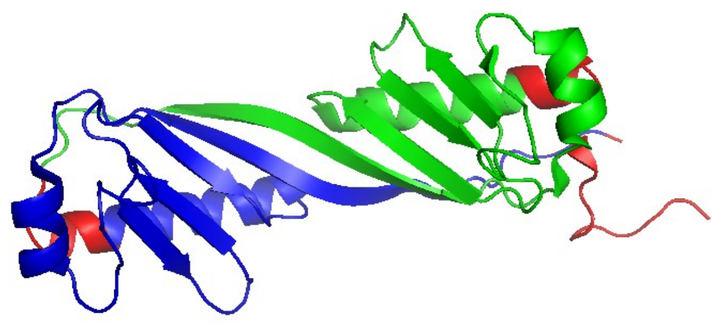
L'objectif principal de ce projet était d'identifier et de caractériser l'ensemble des DDPI (Duplications Directes Proches Intra-géniques) des gènes codants des organismes microbiens complètement séquencés (archaea, bactéries et levures) aux trois niveaux de descriptions que sont l'ADN, les séquences protéiques et leur structure tridimensionnelle.

DEROULEMENT DU PROJET

Méthodes et études intégrées des duplications aux trois niveaux

- Swelfe, un logiciel de recherche « intégrée » de duplications (2003-2005): une même méthode de détection des DDPI pour plusieurs modes de représentation de l'objet biologique (séquences ADN, séquences protéiques, structures 3D protéiques).

Exemple de répétition sur la Tata-box Binding Protein (TBP) d'un archée mesothermophilique, *Sulfolobus acidocaldarius* (1MP9). : a) structure 3D, b) séquence protéique, c) ADN. Les répétitions sont représentées en bleu et vert et le reste de la séquence en rouge.



b) DEIPYKAVVNIENIVATVTLDQTLDLYAMERSVPNVEYDPDQFPGLIFRLESPKITSLIFKSGKMVVTGAKSTDELIKAVKRIIKTLKK
YGMQLTGKPKIQIQNIVASANLHVIVNLDKAAFLLENNMYEPEQFPGLIYRMDEPRVVLLIFSSGKMVITGAKREDEVHKAVKKIFDKLVE
LDCVKPVEEEELE

c) GATGAGATCCCGTATAAAGCAGTCGTAAATATAGAGAATATCGTTGCCACAGTGACTTTGGATCAAACATTGGATTTATATGCGATGGA
AAGAAGCGTACCAAACGTTGAATATGATCCTGATCAATTCCCAGGATTAATATTTAGGCTTGAATCTCCCAAGATAACCTCATTAATATTT
AAATCAGGAAAAATGGTCGTTACTGGAGCTAAAAGTACAGATGAGCTAATAAAGGCTGTAAAACGAATTATAAAAACCCTTAAAAAATATG
GAATGCAACTAACAGGAAAACCTAAGATACAAATACAAAACATAGTCGCATCAGCTAATCTGCACGTTATAGTTAACCTTGATAAAGCAGC
ATTCCTGCTAGAGAATAACATGTACGAACCAGAGCAGTTCCCAGGTCTAATATATAGAATGGATGAGCCCAGAGTTGTTCTATTAATTTTT
AGCAGTGGTAAAATGGTTATTACAGGAGCTAAGAGAGAAGATGAAGTTCATAAGGCTGTTAAAAAAATATTCGATAAACTGGTAGAGTTAG
ATTGTGTAAAGCCCGTTGAAGAAGAAGAGTTAGAA

Mise à la disposition de la communauté des résultats

- MicrOBI : une approche pragmatique de l'intégration de données en biologie (2006)**:** MicrOBI est une base intégrative ayant pour but de mettre en relation les données taxonomiques, génomiques et fonctionnelles des micro-organismes complètement séquencés. Les données sur les répétitions aux 3 niveaux sont insérées dans la base MicrOBI.

- Evolution et dynamique de duplications proches internes – analyse comparative (2004-2006): voir résultats scientifiques.

Nouveaux algorithmes et méthodes de détection des répétitions

-Triades: Motifs relationnels (2003-2006) **:** Recherche de sous-structures similaires 3D basée sur les positions spatiales relatives des amino–acides (Logiciel « Triades »). Recherche des occurrences de motifs relationnels et définition d'un motif complexe par un ensemble partiellement ordonné de motifs.

- RepSeek: recherche de similarités internes longues (2004) **:** Il s'agit un algorithme de recherche de répétitions sur les séquences nucléiques. Le programme est accessible en ligne à http://wwwabi.snv.jussieu.fr/public/RepSeek.

FAITS MARQUANTS

- Collaboration internationale avec le Swiss Institute of Bioinformatics (SIB) de Genève
- Nouvelles collaborations (projet galectine) entre l'équipe Structure et Dynamique du Génome (Isabelle Gonçalves, Denis Houzelstein et Pierre Netter) et le Département Biologie Intégrative (François Bonhomme et Annie Orth, UMR 5554 Université de Montpellier 2)
- Thèse passée en décembre 2005 à l'Atelier de BioInformatique : Mathilde Carpentier
- Thèse en cours d'Anne-Laure Abraham à l'Atelier de BioInformatique sur l'étude des duplications aux 3 niveaux structuraux, protéiques et nucléiques. Cette thèse a débuté en octobre 2005.

RESULTATS SCIENTIFIQUES

-Développement de plusieurs méthodes de détection de répétitions (voir déroulement du projet).

-Résultats biologiques: Environ 15% des protéines contiennent des répétitions détectables à au moins 1 niveau. Certaines répétitions en structures ne sont plus détectables en séquences car celles-ci ont divergé. Certaines répétitions en acides aminés ne sont pas trouvées au niveau des structures 3D, en partie à cause des gaps qui sont rares en structures

- Evolution et dynamique des duplications proches internes – analyse comparative et études expérimentales sur la dynamique des répétitions proches: Analyse phylogénétique d'une famille multigénique (les galectines) chez les vertébrés. Nous avons montré que ces gènes *Lgals* bi-CRD résultent d'une duplication intragénique en tandem antérieure à la divergence des vertébrés. L'analyse de la famille multigénique permet donc d'avoir un exemple de ce que peut-être l'évolution de gènes possédant une duplication interne.

PRINCIPALES PUBLICATIONS

[1] Rocha E.P.C., Cornet E., Michel B. (2005) Comparative and Evolutionary Analysis of the Bacterial Homologous Recombination Systems . PLoS Genet 1(2): e15.

[2] Carpentier M., Brouillet S. and Pothier J. (2005) YAKUSA : a fast structural databases scanning method. Proteins : Structures, Fonctions and Bioinformatics, 61, 137-51

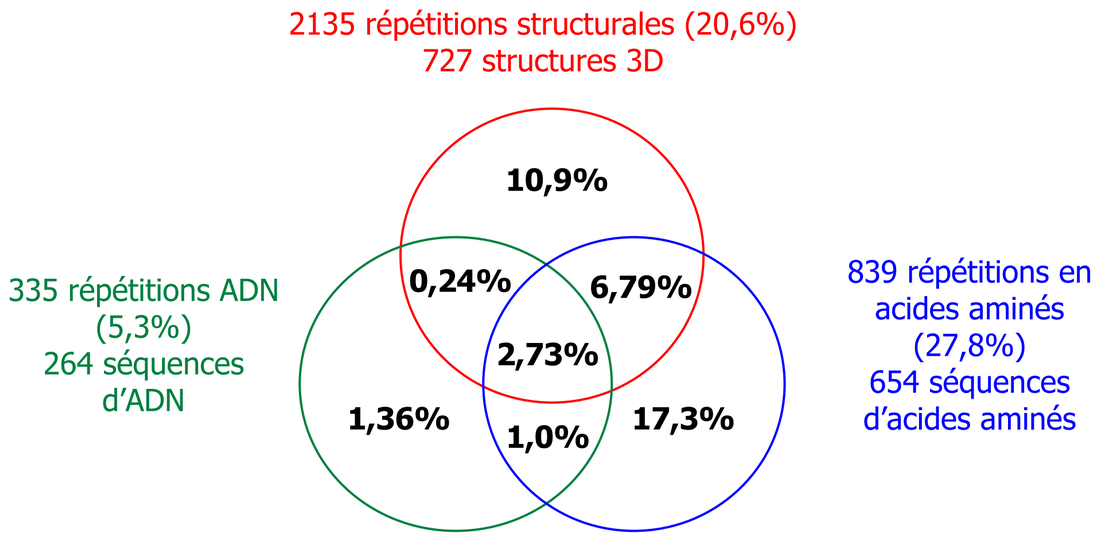
[3] Achaz G., Boyer F., Rocha E.P.C., Viari A., Coissac E. (2007) Repseek, a tool to retrieve approximate repeats from large DNA sequences. Bioinformatics. 23, 119-121

[4] Pisanti N., Soldano H. and Carpentier M. (2005) Incremental Inference of Relational Motifs with a Degenerate Alphabet. In 16th Annual Symposium on Combinatorial Pattern Matching (CPM 2005), Jeju Island, Korea, LNCS 3537. 229-240 Springer-Verlag

[5] Pisanti, N., Soldano H., Carpentier M. and Pothier J. (2006) Implicit and Explicit Representation of Approximated Motifs. in Algorithms for Bioinformatics, C. Iliopoulos and K. Park and K. Steinhofel editors, King's College London Press. Texts in Algorithmics, **6,** 1-14

[6] Houzelstein, D., Goncalves I.R., Fadden A.J., Sidhu S.S., Cooper D.N., Drickamer K., Leffler H., and Poirier F. (2004) Phylogenetic analysis of the vertebrate galectin family. Mol. Biol. Evol. 21:1177-1187

Analyse globale des répétitions



Les répétitions trouvées à chaque niveau ne sont pas les mêmes :

Pour les protéines contenant des répétitions à au moins un niveau : pourcentage de résidus présents dans les répétitions à un ou plusieurs niveaux.

Par exemple, les répétitions trouvées à la fois au niveau des séquences en acides aminés et des structures 3D représentent 6,79% de la longueur totale des protéines ayant des répétitions à au moins un niveau. A chaque fois, nous avons vérifié que ces répétitions ne pouvaient pas être trouvées aux autres niveaux.