UNIVERSITE
PIERRE & MARIE CURIE
LA SCIENCE A PARIS

Mémoire d'habilitation
à diriger les recherches

La fluidité des génomes

Présenté par

Eric Coissac

le 18 Novembre 2005
en présence du jury composé de

Professeur Pierre Netter, Président, Université Pierre et Marie Curie (Paris VI),
Professeur Antoine Danchin, Rapporteur, Institut Pasteur Paris,
Professeur Victor Jongeneel, Rapporteur, Université de Lausanne,
Docteur Jean-Loup Risler, Rapporteur, CNRS, Génopôle d'Evry,
Docteur François Rechenmann, Examinateur, INRIA Rhône-Alpes.

Table des matières

Résumé

Depuis le milieu des années 1990 et la publication des deux premiers génomes complètement séquencés (*Haemophilus influenzae* et *Saccharomyces cerevisiae*), la biologie a franchi une nouvelle étape. Après la révolution de la biologie moléculaire du début des années 1970 et la vision, que certains qualifient de réductionniste, qu'elle a amenée, l'ère de de la génomique fait actuellement évoluer la biologie vers une vision plus intégrative. Ce nouvel engouement pour une biologie dite intégrative a permis de prendre conscience que l'idée selon laquelle l'inventaire complet des gènes d'un organismes permettrait d'appréhender son fonctionnement est une vision simpliste, bien qu'elle ait justifiée en grande partie le développement de nombreux «projets génomes».

J'ai eu la chance de commencer mes travaux de recherche au début des projets génomes et j'ai, dans ce cadre, participé au projet de séquençage du génome de la levure *Saccharomyces cerevisiae*. Je ne pourrais dire si c'est en opposition à l'idée du génome vu comme un simple sac à gènes, mais dès ce moment, j'ai orienté mon travail de recherche vers l'étude de l'évolution de la structure des chromosomes de la manière la plus indépendante possible des gènes qu'ils portent. Il m'importe, au travers de mes travaux, d'essayer de mettre en évidence des contraintes évolutive qui sont liées à la nature même du support de l'information génétique et non à l'information portée.

La stratégie suivie m'a conduit a étudier les mécanismes de duplication à l'origine de nombreux remaniements chromosomiques. Il m'a été ainsi possible de proposer un modèle expliquant l'origine de nombreuses répétitions observables dans les génomes ainsi que leurs évolutions. Ce modèle semble être applicable, pour ses grandes lignes, aux trois super règnes (Eucaryotae, Eubacteriacae et Archae) ce qui montre le caractère ancestrale des mécanismes sous-jacents.

Même si l'exercice présente un intérêt, il ne serait sans doute pas raisonnable de poursuivre ce type de travail sans tenter de croiser les résultats ainsi obtenus avec des données relatives à l'information présente sur les chromosomes, et donc à la fonction des gènes codés par ceux-ci. La mise en place du lien entre les données de répétitions dont je dispose et les données fonctionnelles disponibles relève de l'intégration et donc de la représentation des connaissances. MicrOBI peut être considéré comme ma réponse à ce problème. Aujourd'hui cette base de donnée permet de maintenir cohérent les liens existant entre plusieurs bases de données publiques décrivant différents types d'informations biologiques. L'ajout des données de répétition au schéma actuel permettra de poser au système des requêtes complexes intégrant les différents niveau de données que sont le génome, le protéome et les classifications fonctionnelles.

Abstract

Since the publication of the two first fully-sequenced genomes (*Haemophilus influenzae* and *Saccharomyces cerevisiae*) in the second half of nineties, biology has entered in a new phase. After the era of molecular biology, at times described as reductionistic biology, the new genomic era has incited biologists to adopt a more integrative point of view. This new emphasis on what has been called "integrative biology", has fled to the awareness that the simplistic idea according to which the knowledge of the full list of genes of an organism is enough to understand its functioning is not realistic, even if it has justified many genome projects.

I was lucky to begin my research work at the same time when the genome projects started. I participated tin the yeast genome project leading to the full sequence of the first eukaryotic genome in 1997. I do not know if it is by opposition in the simplistic idea described above, but since at time I have worked on chromosome evolution as independently as possible of the genes carried by them. I have tried through my works to highlight evolutionary constraints associated to DNA structure and not to its information content.

For this reason, I am mainly studying duplication mechanisms which explain many of the chromosomal rearrangements. A model issued from this work has been proposed. It explains the origin and fate of observable duplications in genomes. It seems to be applicable, to a large extend, to the three super-kingdoms (Eukaryota, Eubacteria and Archea), which is for me a good indication of the ancestral aspects of the underlying mechanism.

Even if this approach independently of gene function has its interest, it would undoubtedly not be reasonable to continue this type of work without trying to connect results obtained until now with data on the information carried by the chromosomes and thus with the function of genes encoded by them. Establish a link between data on repeats available to me and functional data available in public database is a challenge involving data integration and knowledge representation. MicrOBI can be regarded as my response to this problem. Today, this database allows the maintenance of coherent existing links between several public databases describing different type of biological information. Adding data on chromosomal repeats to the present schema will allow the construction of complex requests integrating different levels of information such as genome, proteome and functional classification.

Première partie

Curiculum Vitae

État Civil

Date de naissance

27 septembre 1968

Adresse professionnelle :

INRIA-Rhône-Alpes
Projet Helix
655 Avenue de l'Europe
Montbonnot
38 Saint Ismier
France

Téléphone : +33 4 76 61 53 40
E-mail : Eric.Coissac@inrialpes.fr

Adresse personnelle :

Sous le château
73110 Arvillard
France

Téléphone : 33 8 73 09 27 68

Formation universitaire

1996 :

Thèse d'Université en génétique
Directeur : Professeur Pierre Netter
Université Paris VI (Pierre et Marie Curie).

1991 :

Diplôme d'études approfondies de «génétique moléculaire et cellulaire»
Université Paris XI (Paris Sud Orsay)

1990 :

Maîtrise de génétique,
Université Paris XI (Paris Sud Orsay)

1989 :

Licence de biochimie,
Université Paris XI (Paris Sud Orsay)

1988 :

Brevet de technicien supérieur en Biotechnologie,
Ecole Nationale de Chimie, Physique, Biologie (Paris)

Expérience Professionnelle

Septembre 2002 à aujourd'hui :

Délégué par l'Université Paris VI dans le projet Helix à l'INRIA-Rhône-Alpes (*Institut National de Recherche en Informatique et Automatique*).

Septembre 1997 à Septembre 2002 :

Maître de conférences en génétique. En charge des enseignements de bioinformatique à l'Université Paris VI.

Octobre 1995 à Septembre 1997 :

Attaché temporaire d'enseignement et de recherche en génétique à l'Université Paris VI.

Octobre 1992 à Octobre 1995 :

Allocataire de recherche - Moniteur en biologie moléculaire et cellulaire à l'Université Versailles-Saint-Quentin.

Activités de recherche

Je suis maître de conférences en génétique à l'Université Pierre et Marie Curie (Paris VI) depuis 1997. Depuis trois ans, je suis en délégation dans le projet Helix à l'INRIA-Rhône-Alpes (Institut National de Recherche en Informatique et Automatique).

Mon activité de recherche est principalement liée à l'étude de la dynamique et de l'évolution des génomes. Pour observer les phénomènes relatifs à l'évolution des chromosomes, j'ai décidé durant mon doctorat d'abandonner les techniques utilisées habituellement dans un laboratoire de biologie pour utiliser une approche bioinformatique.

Dans les pages suivantes, je décrirai rapidement mes activités de recherche par ordre chronologique décroissant.

Septembre 2002 à aujourd'hui :

Laboratoire :
Projet Helix, Dr François Rechenmann,
INRIA-Rhône-Alpes - Grenoble (France).

Le projet Evolrep.
Le projet Evolrep a pour but d'analyser l'évolution des duplications intragéniques. Ces duplications peuvent être recherchées aussi bien au niveau de la séquence nucléique que protéique, mais aussi lorsque cette information est disponible, au niveau de la structure tridimensionnelle de la protéine. Une analyse comparative d'un même événement de duplication à ces trois niveaux d'observation (ADN, protéine 1D et 3D) et ce simultanément dans un grand nombre d'espèces, est une source inestimable de données pour la compréhention de l'évolution des génomes. Ce projet est mené en collaboration avec deux laboratoires parisiens et un laboratoire grenoblois. Il est financé par l'ACI IMPBIO (2004-2006) du ministère de la recherche.

Le projet MicrOBI.
Pour analyser finement un génome (par exemple lier des données génomiques et des données fonctionnelles), il est indipensable d'intégrer une grande quantité de données hétérogènes, de par leurs natures et leurs origines.
Afin de synchroniser les données issues de différentes sources, j'ai mis en place une base de données relationnelle (MicrOBI) assurant la cohérence des références croisées entre des données génomiques (Genomes EBI), protéiques (*Swiss-Prot*et *HAMAP*), métabolique (KEGG), taxonomique (NCBI) et de classification fonctionnelle (GeneOntology et Enzyme).
Ce travail est maintenant effectué en collaboration avec Anne Morgat de l'Institut Suisse de Bioinformatique (Genève) dans le cadre du projet de réannotation des données métaboliques de la base de données *Swiss-Prot*.

Janvier 1997 à Septembre 2002 :

Laboratoires :
Structure et dynamique des génomes, Prof. Pierre Netter,
Institut Jacques Monod, / CNRS - Paris (France).

Atelier de bioinformatique (ABI) / UPMC - Paris (France).

Analyse des répétitions au niveau chromosomique.
Les origines des répétitions nucléiques observables dans les chromosomes sont multiples (transposons, ADN satellite...). Parmi toutes les répétitions observées, je m'intéresse à celles issues d'«accidents» chromosomiques (*i.e* les répétitions crées par des erreurs de réplication ou de recombinaison). L'idée sous-jacente est que ces accidents sont le moteur principal de l'évolution et que les duplications qu'ils engendrent laissent parfois des traces observables sur les séquences nucléiques.
L'observation des ces événements de duplication chez un grand nombre d'organismes nous a conduit à proposer un modèle expliquant la création et l'évolution de ces répétitions. À ce jour, ce modèle est acceptable pour tous les organismes étudiés (bactérie, archaebactérie ou eucaryote).
Ce travail a été principalement développé par Guillaume Achaz durant son doctorat sous ma responsabilité scientifique.
Le projet EvolRep, exposé précédemment, vise à étendre ce travail.

Juin 1991 à Décembre 1996 :

Travail de thèse :

Directeur de thèse : Prof. Pierre Netter
Laboratoire :
Centre de Génétique Moléculaire / CNRS - Gif-sur-Yvette (France).

- **Projet de séquençage du génome de *Saccharomyces cerevisiae*.**
  J'ai personnellement travaillé à l'obtention de la séquence des régions télomérique et sub-télomérique gauches du chromosome VII.

- **Analyse bioinformatique de l'organisation des duplications dans le génome de la levure.**
  Cette analyse a été réalisée en étudiant l'organisation chromosomique de familles multigéniques. Il a permis de décrire l'organisation du génome de la levure en grands blocs de duplication.

Septembre 1989 à Juin 1990 :

Travail de DEA :

Laboratoire :
Laboratoire du Prof. Piotr Slonimski,
Centre de Génétique Moléculaire / CNRS - Gif-sur-Yvette (France).

- **Analyse de mutant de la protéine senseur d'oxygène *HAP1* chez la levure (*Saccharomyces cerevisiae*).**
  Directeur : Dr. Eric Petrochilo

- **Analyse des interactions entre deux gènes de *Drosophila melanogaster*.**
  Ces gènes, *Shaggy* et *Fused*, codent pour des protéines kinases jouant un rôle dans la mise en place de la polarité segmentale durant le processus de développement de l'embryon.

Directeur : Prof. Bernadette Limbourg-Bouchon

- **Analyse de la relation structure – fonction de la cytochome C** oxydase de la levure *Saccharomyces cerevisiae*. Cette étude a été menée grâce à des techniques de biologie moléculaire et de génétique.
  Directeur : Prof. Pierre Netter

Janvier 1988 à Juillet 1988 :

Directeur : Dr Osamu Kurahashi

Laboratoire :
Ajinomoto - Eurolysine S.A - Centre de recherche, Orsay (France).

Délétion de la région atténuatrice de l'opéron thréonine par mutagenèse dirigée chez *Escherichia coli*.

Direction d'étudiants

Direction de thèse :

Co-direction, avec le Professeur Pierre Netter, de la thèse de Guillaume Achaz :
Sujet :
Etude de la dynamique des génomes : les répétitions intrachromosomiques.

Voir la partie bibliographie page 22 pour la liste des publications relatives à cette direction de thèse (Achaz *et coll.*, 2000, 2001, 2002, 2003).

Octobre 1998 à Juin 2002

Direction de DEA :

Conjointement à mon implication dans l'enseignement de la bioinformatique au sein du DEA de génétique (Université Paris VI, Paris VII, Paris XI et Versailles-Saint-Quentin), j'ai dirigé chaque année de un à trois stages de DEA (encadrement de huit étudiants au total).
Les thèmes de recherches abordés durant ces stages sont :
- Classification des gènes par des modèles de Markov.
- Détermination de la spécificité tRNA - acide aminé.
- Analyses de l'organisation des télomères de la levure.
- Etudes des familles protéiques d'archaebactéries présentant une duplication intragénique.

Direction de DESS :

- Septembre 1997 à Juin 1998

  Directeur de Stéphane Descorps-Declere (*DESS Informatique appliqué à la biologie* Université Pierre et Marie Curie)
  Analyse de la synténie des chromosomes de levure par programmation dynamique.

- Septembre 2001 à Juin 2002

  Directeur de Solène Deude (*DESS Etudes de Génomes : Outils Informatiques et Statistiques (EGOIST* Université de Rouen)
  Développement d'un module Python pour la manipulation des données de biologie moléculaire.

- Avril 2003 à Juillet 2003

  Directeur de Zoya Daneshrad (*DESS Compétences complémentaires en informatique*, Université de Grenoble I)
  Spécification d'une interface web dédiée à la base de données MicrOBI.

Enseignement

J'occupe un poste d'enseignant/chercheur à l'Université Paris VI, depuis octobre 1995.

Module d'école doctorale

Septembre 2001 à Juin 2002

Trois enseignements de 30 heures :
- Manipulation de données sous système Unix.
- Introduction à l'algorithmique pour les biologistes.
- Introduction à l'analyse de séquences.

Université Paris VI (Pierre et Marie Curie).

Enseignement de Master

Enseignement de DEA (Master 2 recherche)

- Mise en place d'un cours d'introduction aux techniques de représentation de connaissances en biologie (30h).

  Septembre à Décembre 2004

  Université Joseph Fourier de Grenoble.

- Coordination et enseignement des cours de bioinformatique du DEA de génétique.

  Septembre 1999 à Juin 2001

  Durant cette période, j'ai été membre du jury de DEA et responsable de l'enseignement de bioinformatique (40 heures) pour l'ensemble des étudiant. J'étais, de plus, responsable du stage des étudiants qui suivaient l'option bioinformatique du DEA.

  Université Paris VI (Pierre et Marie Curie).

- Responsable du DEA *Analyse des génomes et modélisation moléculaire* pour l'Université Paris VI.

  Septembre 2000 à Juin 2002

  Université Paris VI (Pierre et Marie Curie) et Université Paris VII (Denis Diderot).

Enseignement de DESS (Master 2 professionnel)

- DESS EGOIST (*Etudes de Génomes : Outils Informatiques et Statistiques*)

  1999 à 2002

  Cours/TD sur les alignements de séquences (algorithmes exacts et heuristiques)

  Université de Rouen.

Enseignement de Maîtrise - Master 1

– Coordination et enseignement des cours de bioinformatique en licence de biologie cellulaire et physiologie - option génétique

Septembre 1997 à juin 2002

Ce cours est intégré dans le module : «Structure et dynamique des génomes» (70 heures sur 130 pour le module complet)
– Bases de données en biologie
– Alignements de séquences
– Principes de base de la phylogénie moléculaire
Université Paris VI (Pierre et Marie Curie).

– Enseignement dans les enseignement de biologie moléculaire et cellulaire

Octobre 1995 à Juin 1998

Université Paris VI (Pierre et Marie Curie).

Enseignement de Licence

Enseignement en Licence 3 - *Licence*

– Enseignement de bioinformatique pour les étudiants de licence d'informatique.

Janvier à mars 2002

Présentation du problème de l'alignement de séquences biologiques. Aspects algorithmiques et mise en relation avec le modèle biologique.
– Définition de la similarité entre séquences et distances associées.
– Algorithmes exacts d'alignement de séquences (programmation dynamique)
– Approches heuristiques de l'alignement (algorithmes des programmes BLAST et FASTA)
– Détection de gènes chez les procaryotes
Université Paris VI (Pierre et Marie Curie).

Enseignement de Licence 1 et 2 - *DEUG*

– Introduction à l'informatique pour les étudiants biologistes.

Septembre 1996 à juin 2002

– Introduction à la modélisation en biologie
– Cours de programmation (PASCAL)
Université Paris VI (Pierre et Marie Curie).

– Biologie Moléculaire et cellulaire
64 heures par an durant mon doctorat (moniteur).
Université de Versailles - Saint Quentin.

Formation permanente

– Organisation de trois cours de formation permanente en bioinformatique

Janvier à Juin 2005

– Introduction à l'analyse de séquences (30h)
– Introduction à UNIX pour les biologistes (20h)

– Introduction à la représentation de connaissances en biologie (30h)
Université Joseph Fourier de Grenoble.

– Formation permanente en bioinformatique : *Génomique : utilisation de l'outil informatique*

Septembre 1997 à juin 2002

40 heures par an
Université Paris VI (Pierre et Marie Curie).

– Formation continue en biologie moléculaire pour le diagnostique médical
au centre biomédical des Cordeliers
Université Paris VI (Pierre et Marie Curie).

Autres activités scientifiques

Comité scientifique :

– Membre du comité directeur de JOBIM (Journée Ouverte de Biologie, Informatique et Mathématique), depuis 2000.

Publications

Articles publiés dans des revues avec comité de lecture

- Achaz G., Coissac E., Netter P. and Rocha EPC.
  Associations between inverted repeats and the structural evolution of bacterial genomes
  Genetics. 2003 Aug ;164(4) :1279-89.

- Baudin-Baillieu A., Fernandez-Bellot E., Reine F., Coissac E. and Cullin C.
  Conservation of the prion properties of *Ure2p* through Evolution,
  Mol Biol Cell. 2003 Aug ;14(8) :3449-58.

- Achaz G, Rocha EPC, Netter P, Coissac E.
  Origin and fate of repeats in bacteria.
  Nucleic Acids Res, 2002. 30(13) : p. 2987-94.

- Achaz G., Netter P. and Coissac E.,
  Study of intrachromosomal duplications among the eukaryote genomes.
  Mol Biol Evol, 2001. 18(12) : p. 2280-8.

- Achaz G, Coissac E, Viari A, Netter P..
  Analysis of intrachromosomal duplications in yeast *Saccharomyces cerevisiae* : a possible model for their origin.
  Mol Biol Evol, 2000. 17(8) : p. 1268-75.

- Coissac E., Maillier E., and Netter P.,
  A comparative study of duplications in bacteria and eukaryotes : the importance of telomeres.
  Mol Biol Evol, 1997. 14(10) : p. 1062-74.

- Figueroa-Bossi N, Coissac E, Netter P, Bossi L.
  Unsuspected prophage-like elements in *Salmonella typhimurium*.
  Mol Microbiol, 1997. 25 : p. 161-173.

- Tettelin H, Agostoni Carbone ML, Albermann K, Albers M, Arroyo J, Backes U, Barreiros T, Bertani I, Bjourson AJ, Bruckner M, Bruschi CV, Carignani G, Castagnoli L, Cerdan E, Clemente ML, Coblenz A, Coglievina M, Coissac E, Defoor E, Del Bino S, Delius H, Delneri D, de Wergifosse P, Dujon B, Kleine K, *et al.*,
  The nucleotide sequence of *Saccharomyces cerevisiae* chromosome VII.
  Nature, 1997. 387(6632 Suppl) : p. 81-4.

- Coissac E, Maillier E, Robineau S, Netter P.,
  Sequence of a 39,411 bp DNA fragment covering the left end of chromosome VII of *Saccharomyces cerevisiae*.
  Yeast, 1996. 12(15) : p. 1555-62.

- Barny MA, Guinebretiere MH, Marcais B, Coissac E, Paulin JP, Laurent J.
  Cloning of a large gene cluster involved in *Erwinia amylovora* CFBP1430 virulence.
  Mol Microbiol, 1990. 4(5) : p. 777-86.

Autres articles

- Anne Morgat et Eric Coissac
  La gestion des connaissances biologiques
  Biofutur, Janvier 2005 : p. 20-22

Deuxième partie

Les duplications, marqueurs de la fluidité génomique

La description du monde vivant par la biologie fait apparaître de nombreux niveaux d'organisations allant de l'écosystème à la molécule en passant notamment par la population, l'individu, la cellule. L'évolution des organismes permet notamment leur adaptation au monde qui les entoure. Même si la théorie neutraliste a montré que l'évolution est avant tout un phénomène aléatoire, les contraintes imposées par le milieu naturel imposent des pressions de sélection conduisant au maintien de telles ou telles caractéristiques chez un organisme. La vision la plus simple de ce phénomène est d'imaginer une transmission de ces contraintes du niveau d'organisation le plus élevé, l'écosystème, vers le niveau le plus bas, le gène, à travers chacun des niveaux intermédiaires. Un changement climatique, conduira à sélectionner les populations les plus adaptées aux nouvelles conditions, au sein de celles-ci, les individus les plus résistants seront eux-mêmes sélectionnés, et ainsi de suite jusqu'à la sélection des mutations permettant la meilleure adaptation. Il y a ce que l'on pourrait appeler une propagation descendante des contraintes.

En bas de cette échelle, il y a l'ADN qui porte l'information génétique. Cette molécule possède ces propres règles évolutives liées à sa structure chimique et aux mécanismes qui régissent son maintien (réplication, réparation, recombinaison). Ces caractéristiques imposent elles aussi des contraintes sur l'évolution des organismes en induisant préférentiellement certains types de mutations. De manière similaire, certaines séquences protéiques pouvant en théorie être décrites par un gène ne sont certainement pas réalistes d'un point de vue thermodynamique. Il est donc possible de décrire de manière analogue à la propagation descendante des contraintes, une série de contraintes se propageant de manière ascendante dans les différents niveaux d'organisation.

C'est la recherche des contraintes ascendantes et principalement des contraintes imposées par l'ADN sur l'évolution des protéines qui est sous-jacente à la plus grande partie de mon travail de recherche.

Chapitre 1

L'ère du séquençage des génomes complets

Nous venons de célébrer le cinquantième anniversaire de la découverte de la structure de l'ADN [1] par Watson et Crick (1953) et le troisième anniversaire de la publication annonçant le séquençage complet du génome humain. Ces deux dates illustrent l'immensité du pas franchi par la biologie, et plus particulièrement par la biologie moléculaire, durant la seconde moitié du vingtième siècle. J'ai eu la chance de commencer mon travail de recherche en génétique au début des années 1990, une période charnière pour la biologie, avec l'avènement des projets de séquençage complet de génomes.

Des protéines à l'ADN

L'ADN a longtemps été perçue comme une molécule biologique de second ordre. Considérée comme un simple polysaccharide, on lui assignait au mieux un rôle de structuration du compartiment cellulaire. Les protéines étaient alors considérées comme les composants cellulaires principaux assurant les rôles fondamentaux dans le fonctionnement de la cellule. Leur structure complexe permettait de supposer qu'elles étaient le type de molécule idéale pour stocker et transmettre l'information génétique. La première démonstration du rôle de l'ADN dans le transfert de l'information génétique par Avery et McCarty (1944) s'appuyait sur des expériences de restauration du type S à des souches R de Streptocoques (*Streptococcus pneumoniae*) [2]. On notera qu'elle ne précède que de dix ans les travaux de Watson et Crick. Il faudra attendre 1952, et les travaux de Chase et Hershey sur les bactériophages, pour finir de convaincre, après une longue période de controverse, la communauté des biologistes de ce fait. Cette nouvelle fonction, de support de l'information génétique, attribuée à l'ADN donnait un support moléculaire au concept théorique de gène développé par Gregor Mendel dans son manuscrit « *Versuche über Plflanzenhybriden*» (Expériences sur l'hybridation des plantes) présenté en 1865 lors d'une réunion de la société d'histoire naturelle de Brünn (Mendel, 1865).

Malgré cette découverte, l'influence importante de la communauté des biochimistes et la maîtrise de nombreuses techniques d'étude des protéines continuèrent de donner à celles-ci

[1] acide désoxyribonucléique

[2] Les *streptocoques pneumoniae* existent sous deux formes. La forme R, pour *Rough* ou rugueuse, est saprophyte. La forme S, pour *Smooth* ou lisse, est pathogène. La différence entre ces deux souches tient à l'expression d'un gène dans la forme S, permettant la synthèse d'un muco-peptide donnant aux colonies bactériennes sur une boîte de Pétri un aspect lisse et brillant. Cette enveloppe mucoïde a avant tout comme propriété de permettre aux streptocoques d'échapper au système immunitaire. Dans ses expériences, Avery a mis en contact des streptocoques de type R avec un extrait purifié d'ADN provenant d'une souche S. Après culture de ces bactéries ainsi traitées sur des boîtes de pétrie, il a pu constater la présence de colonies de type S.

Année	Protéine	ARN	ADN	Nombre de résidus
1935	Insuline			1
1945	Insuline			2
1947	Gramicidine S			5
1949	Insuline			9
1955	Insuline			51
1960	Ribonucléase			120
1965		ARNt_{Ala}		75
1967		ARNr 5S		120
1968			Bactériophage λ	12
1978			Bactériophage $\phi X174$	5386
1981			Mitochondrie	16569
1982			Bactériophage λ	48502

TAB. 1.1 – **Historique des travaux de séquençage** : d'après le «Laboratory of Molecular Biology» à Cambridge http ://www2.mrc-lmb.cam.ac.uk/archive/Sanger58.html

un rôle primordial dans le nouveau modèle que les biologistes se faisaient du vivant. Pour beaucoup d'entre eux, le génome n'était qu'une sorte de «sac à gènes».

Le principal frein à la considération de l'ADN comme un des éléments essentiels de la biologie est certainement lié aux difficultés techniques rencontrées pour le séquencer, alors que depuis le milieu des années 1950, la détermination de la structure primaire de l'insuline par Sanger *et coll.* (1955) avait ouvert la voie du séquençage des protéines. Les travaux de Pehr Edman, initiés dès le début des années 1960 (Edman et Begg, 1967) et la commercialisation quasi simultanée par la société Beckmann d'un appareil automatique de séquençage des protéines, basé sur ces travaux, finirent de démocratiser cette technique.

L'accès à l'information génétique était plus que fastidieux. Les premières séquences d'ADN obtenues au début des années 1970 étaient réalisées indirectement en séquençant, par des digestions enzymatiques ménagées, une copie ARN de la molécule d'ADN à séquencer. La molécule d'ARN était synthétisée par transcription *in vitro*. À titre d'exemple, le séquençage de l'opérateur de l'opéron lactose par Gilbert et Maxam en 1973 demanda, en plus de constructions génétiques fastidieuses à obtenir, plus de cent litres de culture bactérienne et une millicurie de radioactivité. Tous ces efforts ne permirent de déterminer que vingt-quatre paires de bases d'une molécule d'ADN[3] (voir TAB. 1.1).

Malgré les travaux prometteurs de Sverdlov *et coll.* (1973) qui permettaient de déterminer par une série de réactions chimiques l'enchaînement des deux types de nucléotide (purine et pyrimidine), il faudra attendre 1977 pour voir décrit dans la littérature deux méthodes efficaces de séquençage. La première, publiée par Maxam et Gilbert (1977), peut être considérée comme l'aboutissement des travaux de Sverdlov. La seconde, publiée six mois plus tard par Sanger *et coll.* (1977), s'appuie sur une synthèse *in vitro* de l'ADN en présence d'agents inhibiteurs. C'est cette dernière méthode qui a permis le réel développement du séquençage de l'ADN

Ce décalage dans la maîtrise des techniques liés à la manipulation des deux grandes macromolécules du vivant se poursuivit jusqu'au début des années 1980. Durant cette période, le séquençage des protéines est nettement plus fréquent que celui de leur gène. On peut dater le passage à notre mode de travail actuel, où la séquence d'une protéine est déduite de la séquence nucléique de son gène, avec la publication de la séquence du gène de la β galactosidase (Kalnins *et coll.*, 1983) qui démontra la qualité des séquences ainsi obtenues.

Un attrait nouveau est donné à l'ADN notamment par les travaux de François Jacob, Jacques Monod et André Lwoff, qui durant les années 1960 démontrent l'existence de si-

[3]Un gène d'une bactérie est composé souvent de plus de mille paires de bases. Le génome d'un virus à une taille de plusieurs dizaines de milliers de paires de bases, celui d'une bactérie de quelques millions de paires de bases et celui de l'homme de 3,2 milliards de paires de bases (Venter *et coll.*, 2001).

gnaux, portés par la molécule d'ADN, participant à la régulation l'expression des gènes. C'est le début de la biologie moléculaire (ou de la génétique moléculaire).

Le séquençage des génomes complets

Les premières séquences publiées après la mise au point des méthodes modernes de séquençage de l'ADN étaient principalement de petite taille et correspondaient biologiquement à la séquence d'un gène et de ses régions régulatrices proches (promoteur). L'idée de séquencer des génomes complets est certainement née de manière concomitante avec cette avancée technologique. Il suffit pour s'en convaincre de ne prendre en compte que la publication de la séquence complète du génome du bactériophage ϕX174 en 1978 (Sanger *et coll.*, 1978), soit moins d'un an après la publication des méthodes de séquençage de l'ADN. Certes, ce génome viral a une très petite taille (5386 pb), mais l'idée selon laquelle la connaissance de l'information génétique complète d'un organisme était intéressante scientifiquement était née.

Durant la seconde moitié du vingtième siècle, la biologie a vécu deux grandes périodes de son histoire. Elle a d'abord mis en évidence le support moléculaire de plusieurs des éléments qu'elle décrivait dans ces modèles. L'accessibilité à l'expérience de ces molécules permit l'essor de la biologie moléculaire, qui eut comme but principal de décrire des mécanismes de plus en plus fin. L'étude des mécanismes moléculaires a permis d'une part de démontrer l'unicité du monde vivant à travers ces mécanismes fondamentaux, mais parfois elle a aussi malheureusement conduit à la perte de la vision d'ensemble de l'organisme étudié. Ce travail de dissection moléculaire d'un mécanisme précis a connu son apogée au début des années 1990. Les techniques de biologie moléculaire, devenues très performantes, mettaient à portée de main des scientifiques des rêves impossibles quelques années plus tôt comme le séquençage complet de grands génomes. Outre-Atlantique, des projets de séquençage du génome humain s'échafaudaient. En Europe, André Goffeau mit en place, dès 1989, un consortium pour séquencer le génome de la levure *Saccharomyces cerevisiae*. C'est dans ce contexte que ma thèse débuta avec comme objectif le séquençage des régions télomérique et sub-télomérique gauches du chromosome VII de la levure.

Les «projets génomes», constitués du projet de séquençage par lui même et des analyses globales du transcriptome ou du protéome, ont ouvert de nouvelles perspectives en offrant à nouveau une vision globale de l'organisme, mais cette fois-ci au niveau moléculaire. Cette photographie générale et très détaillée de l'organisme étudié fournit tellement d'informations simultanément que les méthodes de raisonnement utilisées durant les vingt premières années de la biologie moléculaire ne peuvent plus suffire. Pour tirer profit de ces nouveaux résultats, il est nécessaire de réintégrer les connaissances acquises de manières éparses, de les formaliser afin de pouvoir les analyser de manière plus rationnelle. Une forme de modélisation, plus formelle que celle utilisée en biologie moléculaire, est donc en train de reprendre une place importante en biologie. Place importante qu'elle tenait déjà, notamment au début de la génétique et qu'elle n'a jamais cessé d'occuper dans d'autres disciplines comme l'écologie ou la phylogénie.

Cette nécessité de formalisation a conduit au développement d'interfaces entre la biologie et d'autres disciplines scientifiques comme les statistiques et l'informatique. Cette nouvelle approche de la biologie est généralement identifiée sous le nom de bio-informatique. Elle a pour but d'utiliser les formalismes décrits dans ces autres disciplines pour décrire de manière formelle les phénomènes biologiques et ainsi permettre l'utilisation d'outils aptes à répondre à ces questions biologiques reformulées dans un autre langage. Dans les pages suivantes de ce manuscrit, j'aborderai avec une approche bio-informatique le problème de l'évolution et de la fluidité des génomes. Mes propos suivront ce que je pense être une démarche bio-informatique typique et se diviseront en une présentation de la problématique biologique, une description des modèles informatiques utilisables pour aborder cette question biologique et enfin un retour sur la biologie avec une analyse des principaux résultats obtenus.

Chapitre 2

La fluidité de génomes

La fluidité des génomes est le fil conducteur de mon travail de recherche. Pour reprendre l'idée de «bricolage moléculaire» proposée par François Jacob (1981), je pense que les remaniements chromosomiques sont un des moteurs essentiels de l'évolution. Ils sont issus d'accidents moléculaires qui offrent, pour certains d'entre eux, de nouvelles potentialités fonctionnelles à l'organisme alors que d'autres conduisent à des dysfonctionnements plus ou moins importants. L'accès récent à la séquence de génomes complets permet d'étudier ces remaniements en appréhendant certains des mécanismes à leur origine. De plus, le nombre croissant de génomes séquencés permettra certainement, par une étude de génomique comparative, l'étude de l'impact des remaniements sur les fonctions cellulaires.

Les duplications, marqueurs de la fluidité génomique

Les duplications dans les génomes intéressent les biologistes depuis très longtemps. La plus vieille duplication étudiée est certainement la mutation *BAR* de *Drosophila melanogaster* (Tice, 1914). Caractérisée au niveau moléculaire en 1989 comme issue d'une recombinaison illégitime entre deux éléments B104 localisés en 16*A*1 et 16*A*7 (Tsubota *et coll.*, 1989), elle fut d'abord identifiée, en 1936 par deux laboratoires indépendants, comme une duplication de la région 16*A* du chromosome X (Bridges, 1936; Muller *et coll.*, 1936). Les séquences d'ADN étant bien évidemment indisponibles à l'époque, c'est par l'observation des chromosomes polythènes des glandes salivaires que cette caractérisation fut réalisée. Cette mutation *BAR* et les duplications associées à ces divers allèles ont été étudiées de manière beaucoup plus précise. Cela a permis d'identifier non seulement le cas de duplication, mais aussi de triplication de la région 16*A* (Sutton, 1943).

Au départ, les duplications étaient perçues comme des événements rares, surtout chez les organismes possédant un génome de petite taille comme les bactéries et les archaebactéries. C'est l'avènement de l'ère de la génomique qui a permis d'estimer la grande importance de ces événements. Jusqu'au début des années 1990 et le démarrage du projet de séquençage du génome de la levure *Saccharomyces cerevisiae*, seules quelques familles multigéniques étaient décrites pour cet organisme. La levure, bien qu'étant un eucaryote, était considérée comme un organisme possédant un génome très compact où les duplications de gènes n'avaient que peu leur place. Aussi à part les deux iso-cytochromes C (Downie *et coll.*, 1977), les gènes *CUP* de résistance au cuivre (Karin *et coll.*, 1984) et les nombreuses copies des gènes ribosomiques (Venema et Tollervey, 1999), peu de choses étaient connues.

Durant ma thèse, j'ai participé au séquençage du génome de la levure. J'avais en charge la détermination de la séquence de la région télomérique et sub-télomérique gauche du chromosome VII (Coissac *et coll.*, 1996). Lors de ce travail, j'ai pu mettre en évidence un gène codant pour une protéine de 120 acides aminés. Deux autres copies de ce gène avaient déjà été identifiées, l'une sur le chromosome III déjà complètement séquencé, l'autre sur le chromo-

some II, alors en cours de séquençage. Nous avions ainsi mis en évidence l'une des premières triplications dans le génome de la levure. Au fur et à mesure de l'avancement du travail de séquençage, cette famille a vu sa taille croître jusqu'à 23 membres dans le génome complet (Viswanathan *et coll.*, 1994). Le nombre de familles multigéniques a également augmenté de manière très importante durant cette période. Rapidement, des estimations indiquant que 30 à 50% des gènes de la levure appartenaient à des familles multigéniques ont été avancées. Hormis l'intérêt scientifique directement lié à la mise en évidence du niveau important de redondance génique, la présence de ces nombreuses familles permettait d'aborder le problème de la fluidité des génomes. L'idée était de considérer ces familles uniquement comme des marqueurs d'événements de remaniement chromosomique. Dans ce cas, l'étude de leur organisation topologique sur le chromosome pouvait certainement apporter des informations sur les mécanismes ayant conduit à leur création.

Ce travail m'a permis de mettre en évidence une organisation du génome de la levure en grands blocs de duplication (*voir* Fig. 2.1, Coissac *et coll.*, 1997). Un travail similaire a été publié en parallèle par Wolfe et Shields (1997). Ces travaux simultanés ont conduit à l'élaboration de deux modèles permettant d'expliquer cette structure particulière du génome. Le premier reposait sur une duplication complète du génome, soit du fait d'une méiose aberrante, soit du fait d'une fusion de deux cellules (Wolfe et Shields, 1997). Le second posait l'hypothèse d'un mécanisme permanent conduisant à des duplications partielles de grandes régions chromosomiques (Coissac *et coll.*, 1997). Compte tenu des informations disponibles à l'époque rien ne permettait de trancher entre les deux hypothèses. Des travaux récents ont validé l'hypothèse d'une duplication complète du génome suivi d'une perte partielle des gènes dupliqués (Kellis *et coll.*, 2004). L'idée d'une dynamique permanente de l'organisation des chromosomes ne doit pas pour autant être négligée. Des travaux expérimentaux ont montré qu'il était possible de sélectionner des mutants de la levure *S. cerevisiae* présentant des duplications partielles d'un grand fragment d'un de leurs chromosomes (Casaregola *et coll.*, 1998). Je pense donc que si la structure particulière du génome de la levure s'explique par une étape de polyploïdisation, celle-ci n'est pas à l'origine de l'ensemble des remaniements chromosomiques existant dans ce génome et qu'elle n'est qu'un élément parmi tant d'autres.

Peu de séquences de génomes complets étaient disponibles avant la fin des années 1990. Le génome d'*Haemophilus influenzae* a été le premier génome bactérien complètement séquencé (Fleischmann *et coll.*, 1995). À cette même période, les programmes de séquençage des génomes des deux bactéries modèles : *Escherichia coli* et *Bacillus subtillis*, étaient en cours de réalisation. Aussi, les travaux contemporains visant à essayer de comprendre les mécanismes évolutifs des chromosomes ne pouvaient pas se baser sur une approche de génomique comparative. Il était donc nécessaire de mettre en place des approches méthodologiques visant à extraire de l'information à partir des données d'un seul génome. J'ai donc choisi l'analyse des duplications pour atteindre ce but.

Une analyse fine des duplications au niveau nucléique

Les études précédemment citées (Wolfe et Shields, 1997; Coissac *et coll.*, 1997; Kellis *et coll.*, 2004) ont comme point commun d'avoir été réalisées en se basant sur l'organisation chromosiques des gènes dupliqués codant pour des protéines. De ce fait, elles prenaient comme unité de duplication le gène dans son intégralité. Cette approche offre plusieurs avantages. Parmi ceux-ci il faut citer la possibilité d'effectuer les analyses de similarité au niveau des séquences protéiques. La plus grande sensibilité des méthodes d'alignement lorsqu'elles sont utilisées sur des séquences polypeptidiques fait qu'il est possible d'inférer l'homologie entre deux séquences protéiques dès lors que leur alignement met en évidence de 30% à 35% d'identité. Une simulation réalisée sur des séquences nucléiques (*voir* Fig. 2.2) montre que pour l'ADN, il faut atteindre plus de 65% d'identité pour considérer que les deux séquences sont homologues. Travailler au niveau du produit des gènes permet donc d'identifier des événements anciens.

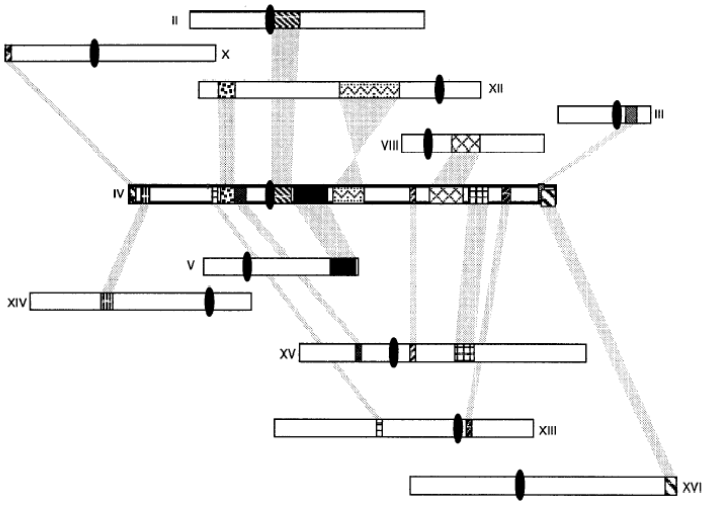


FIG. 2.1 – **Organisation en mosaïque du chromosome IV de la levure** – l'ensemble du génome de *Saccharomyces cerevisiae* est couvert de blocs synténiques présents en deux exemplaires et associant un grand nombre de chromosomes entre eux. Ici, à titre d'exemple, nous voyons comment le chromosome IV partage des éléments de séquences avec 10 autres chromosomes. Un schéma complet incluant tous les chromosomes montrerait qu'aucun des seize chromosomes de la levure n'est exclu de ce réseau.

L'autre avantage apporté par le travail au niveau des protéines est sa plus faible nécessité en puissance de calcul. Même si les génomes bactériens et le génome de la levure sont codants dans une très grande proportion (72% pour *Saccharomyces cerevisiae* et plus de 80% pour la plupart des bactéries) nous pouvons considérer que la longueur totale des séquences protéiques est inférieur au tiers de la longueur de la séquence du chromosome (3 nucléotides $\rightharpoonup$ un codon $\rightharpoonup$ un amino-acide). Ce compactage de l'information réduit d'un facteur 10 le temps de calcul puisque les programmes d'alignement optimaux possèdent une complexité en $O(N^2)$ où N représente la taille des séquences alignées. Cette réduction de temps est loin d'être négligeable puisqu'en 1997 (Coissac *et coll.*, 1997) il fallait plus de trois semaines de calcul pour classer en groupes de séquences similaires les 6000 gènes de la levure en utilisant l'algorithme d'alignement global de Needleman et Wunsch (1970).

La détection des duplications réalisée par l'analyse des séquences traduites permet donc, dans un temps de calcul raisonnable, d'obtenir des résultats d'une grande sensibilité. Cela permet notamment d'identifier des événements de duplication anciens. Cette approche possède, par contre, l'inconvénient de n'offrir qu'une vue macroscopique des événements. Les duplications n'ont aucune raison d'apparaître en suivant les limites des gènes. Il est donc intéressant de rechercher les duplications sur les chromosomes indépendamment du caractère codant ou non codant des régions considérées. J'ai donc, en dirigeant de travail de thèse de Guillaume Achaz, cherché à mettre en place une méthodologie permettant une analyse des duplications directement au niveau de la séquence nucléique des chromosomes.

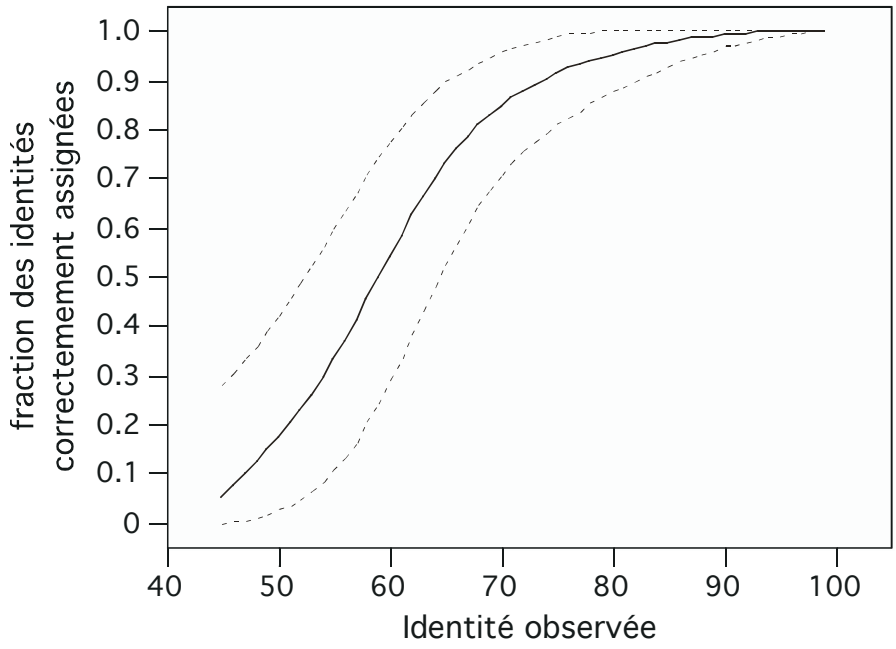


FIG. 2.2 – **Analyse de la pertinence des alignements de séquences nucléiques** – Une évolution simulée d'une séquence selon le modèle suivant a été réalisée : répartition uniforme des mutations le long de la séquence, équiprobabilité de substitution vers chacune des trois autres bases possibles, des insertions – délétions d'une seule paire de base, 10 fois plus de substitutions que d'insertions – délétions. Pour différents temps évolutifs, la séquence générée est alignée contre la séquence initiale. La courbe représente la fraction des bases identiques dans l'alignement correspondant réellement à des base homologues (correspondant à la même base ancestrale) en fonction du taux d'identité observé entre les deux séquences d'après cet alignement. La courbe dessinée en noir correspond à la moyenne de 1000 expériences, les deux courbes en pointillées délimitent les quantiles 5% et 95%. Il apparaît clairement qu'en dessous de 75% d'identité une forte proportion des nucléotides alignés le sont par le hasard

Chapitre 3

La détection des répétitions au niveau de l'ADN

La détection des répétitions directement au niveau des séquences nucléiques complètes des chromosomes pose plusieurs types de problèmes. Ces problèmes tiennent à la complexité algorithmique des solutions envisageables, à la sensibilité des méthodes d'alignements appliquées aux séquences nucléiques et au fort impact du biais de composition des séquences sur les résultats de l'alignement.

Le temps de calcul nécessaire à l'identification des répétitions est certainement le paramètre le plus limitant. La seule approche exacte permettant la détection des répétitions est basée sur la recherche des d'alignements locaux sub-optimaux par programmation dynamique. Proposée par Waterman et Eggert (1987) dans le cadre de l'étude des ARN de transfert (ARNt) et des ARN ribosomiques (ARNr), cette approche a pour inconvénient majeur de posséder une complexité en $O(N^2)$ avec N la taille de la séquence analysée. Cette complexité rend inutilisable cet algorithme pour des séquences complètes de chromosomes. Seules des heuristiques permettent d'obtenir des résultats dans un temps de calcul raisonnable.

Plusieurs stratégies ont été envisagées pour mettre en évidence les répétitions. Bien qu'identifiées dans la littérature sous le même intitulé de «détection de répétitions dans les séquences biologiques», elles ne répondent absolument pas aux mêmes problèmes.

La première stratégie vise à identifier dans une séquence les occurrences de séquences biologiques fortement répétées telles que les éléments transposables, les mini ou microsatellites. Elle s'appuie sur des banques de données répertoriant des séquences déjà connues correspondant à ces différentes catégories de répétitions et utilise des programmes de type BLAST (Altschul *et coll.*, 1997) pour les identifier sur la séquence analysée. L'implémentation la plus connue de ce type d'approche est *RepeatMasker* (Smit *et coll.*, 2004). L'inconvénient principale de cette méthode de détection est que seuls les éléments répétés déjà connus et inclus dans la banque de référence peuvent être mis en évidence. Elle est donc inutilisable pour identifier les répétitions fortuites provenant de remaniements chromosomiques.

La deuxième stratégie ne s'attache à identifier qu'une sous catégorie de répétitions : les répétitions multiples en tandem d'un court motif. Ce type de répétitions correspond, d'un point de vue biologique, aux microsatellites ou aux minisatellites. Plusieurs solutions à ce problème ont été proposées, certaines correspondent à des algorithmes exacts (Landau *et coll.*, 2001), d'autres à des heuristiques (Benson, 1997; Kolpakov *et coll.*, 2003; Delgrange et Rivals, 2004). Bien que ces programmes ne se limitent pas à identifier des familles de répétitions déjà connues, la restriction imposée par leurs algorithmes sur le type des répétitions identifiées les rend inadaptés à l'étude des remaniements chromosomiques.

Enfin, la dernière stratégie consiste à essayer de détecter l'ensemble des répétitions présentes dans un génome. La plupart des programmes réalisant ce travail se restreignent à

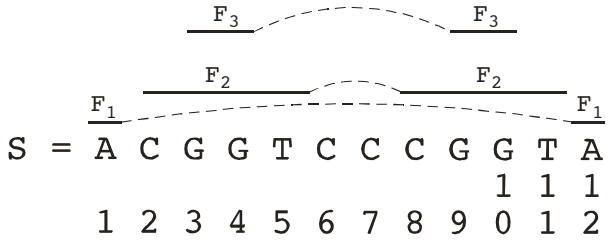


FIG. 3.1 – **Description d'un facteur répété maximal** – Un facteur répété maximal est un facteur présent au moins deux fois dans une séquence et qui ne peut être étendu ni à gauche, ni à droite. La séquence S prise comme exemple possède 12 symboles numérotés de 1 à 12 (en dessous). Cette numérotation sert de base à la dénomination des facteurs F_x où x est la position du premier symbole du facteur en partant de la gauche. Pour cette séquence, les facteurs répétés $F_1 = A$ et $F_2 = CGGT$ répondent à la définition du facteur maximal répété. Par contre le facteur F_3 n'est pas maximal car il peut être étendu à gauche et à droite pour donner le facteur répété maximal F_2.

rechercher les répétitions exactes (sans erreur) présentes dans les séquences. Ils utilisent soit un algorithme exact, du type arbre des suffixes comme le programme *Reputer* (Kurtz et Schleiermacher, 1999), soit une heuristique comme le programme *ForReapeat* (Lefebvre *et coll.*, 2003).

À ce jour, la seule stratégie envisageable permettant de détecter des répétitions approchées dans un génome complet est d'utiliser les résultats des programmes de détection de répétitions exactes décrits ci-dessus et de poursuivre l'analyse en alignant par programmation dynamique les régions flanquantes des répétitions strictes identifiées. C'est cette stratégie, que j'ai utilisée dans mes analyses des répétitions.

Détection des répétitions exactes

La détection des répétitions exactes à l'intérieur d'un texte a fait, et fait toujours, l'objet de nombreux travaux de recherches. Je décrierais plus précisément sur trois structures de données et/ou algorithmes permettant la recherche des répétitions de ce type : l'arbre des suffixes, *KMR* et la table des suffixes.

Le problème informatique qui correspond à la question : « Quelles sont les répétitions exactes présentes dans une séquence d'ADN ? » est d'énumérer l'ensemble des facteurs répétés maximaux (*voir* FIG. 3.1) et de localiser leurs occurrences dans un texte de taille N construit sur un alphabet de taille fini ($\{A, C, G, T\}$ pour une séquence d'ADN).

L'arbre des suffixes

L'arbre des suffixes est une structure de données très intéressante, notamment pour la recherche des répétitions à l'intérieur d'un texte. Elle offre l'avantage de pouvoir être construite dans un temps et un espace mémoire linéaires avec la taille de la séquence (Weiner, 1973; McCreight, 1976; Ukkonen, 1992). Une fois construite, cette structure de données permet d'énumérer simplement les facteurs maximaux répétés dans la séquence. En effet, tout chemin partant de la racine et allant jusqu'à un noeud interne est un facteur maximal. Il est possible de connaître le nombre d'occurrences et la localisation de ces facteurs en comptant le nombre de feuilles présentes sous le noeud interne considéré et en lisant les étiquettes associées à ces feuilles (*voir* FIG. 3.2).

Le seul problème associé à cette structure de données est la taille mémoire nécessaire. Car, si la complexité en temps et en espace de l'algorithme de construction est bien en $O(N)$, la place nécessaire au stockage d'un noeud de l'arbre est relativement importante. Cependant, comme la place nécessaire au stockage d'un noeud est fonction de la taille de l'alphabet, Kurtz et Schleiermacher (1999) ont proposé une implémentation optimale de cette structure pour l'alphabet décrivant l'ADN $\{A, C, G, T\}$. Dans cette implémentation, la place mémoire occupée est de 12,5 octets par nucléotide.

KMR

Richard Karp, Raymond Miller et Arnold Rosenberg ont proposé en 1972 un nouvel algorithme capable de mettre en évidence des motifs répétés dans un texte, un arbre ou un tableau (Karp *et coll.*, 1972). Cet algorithme, communément nommé *KMR* , permet de résoudre deux problèmes :

Problème 1 : Lister l'ensemble des facteurs répétés de taille k.
Problème 2 : Identifier le plus long facteur répété (de taille L_{max}).

L'algorithme *KMR* , bien que moins efficace en temps de calcul que l'arbre des suffixes, possède des propriétés intéressantes qui rendent ce désavantage théorique non rédhibitoire.

Une description simple de l'algorithme *KMR* est de considérer un texte contenant des facteurs répétés. Dans ce texte, tout facteur répété de longueur $L > 1$ peut être divisé en deux facteurs de longueur $L/2$. Réciproquement deux facteurs répétés de longueur $L/2$ contigus sur la séquence forment par concaténation un seul facteur de longueur L (*voir* FIG. 3.3).

À partir de ce principe de base, il est facile d'expliquer intuitivement l'algorithme de *KMR* :

- énumérer les positions de tous les facteurs de taille 1 (les mono-nucléotides)
- concaténer ces facteurs pour former ceux de taille 2
- concaténer les facteurs de taille 2 pour former ceux de taille 4
- recommencer pour former ceux de taille 8 puis 16 et ainsi de suite jusqu'à k (problème 1) ou jusqu'à ce qu'il n'y ait plus de facteurs à concaténer (problème 2).

Si k ou L_{max} n'est pas une puissance de 2, une dernière étape particulière devra être réalisée pour l'ajustement à la longueur désirée.

Pour le problème 1, la complexité de cet algorithme est en $O(N.\log k)$ en temps et $O(N)$ en espace mémoire. D'un point de vue pratique, k est petit devant N (voir page 39, *Statistiques associées aux répétitions exactes*). De ce fait, cet algorithme, théoriquement non linéaire, reste très efficace dans notre cas particulier. Dans la suite de cet exposé, les recherches des répétitions exactes seront réalisées sur des séquences chromosomiques à l'aide de cet algorithme.

Dans le cas de l'analyse de très grandes séquences (par exemple, pour les chromosomes humains) l'espace mémoire est un facteur limitant. L'intérêt de *KMR* est d'offrir la possibilité d'être implémenté de manière plus économe en mémoire que l'arbre des suffixes. L'implémentation habituelle de *KMR* repose sur un système de piles. Pour gagner en compacité, nous avons réalisé une nouvelle implémentation basée sur un système de listes chaînées ne demandant que 8 octets par nucléotide analysé (Achaz *et coll.*, 2005), ce qui représente une amélioration de 25% relativement à l'implémentation de l'arbre des suffixes de Kurtz et Schleiermacher (1999) dans *Reputer*. De plus, dans *KMR* la complexité en espace est indépendante de la taille de l'alphabet dans lequel est écrit le texte.

La table des suffixes

Une nouvelle structure nommée «table des suffixes» a été proposée par Manber et Myers (1990). Elle permet de retrouver les occurrences d'un motif à l'intérieur d'un texte par recherche dichotomique, avec une complexité en temps de $O(P+\log N)$, où P est la longueur

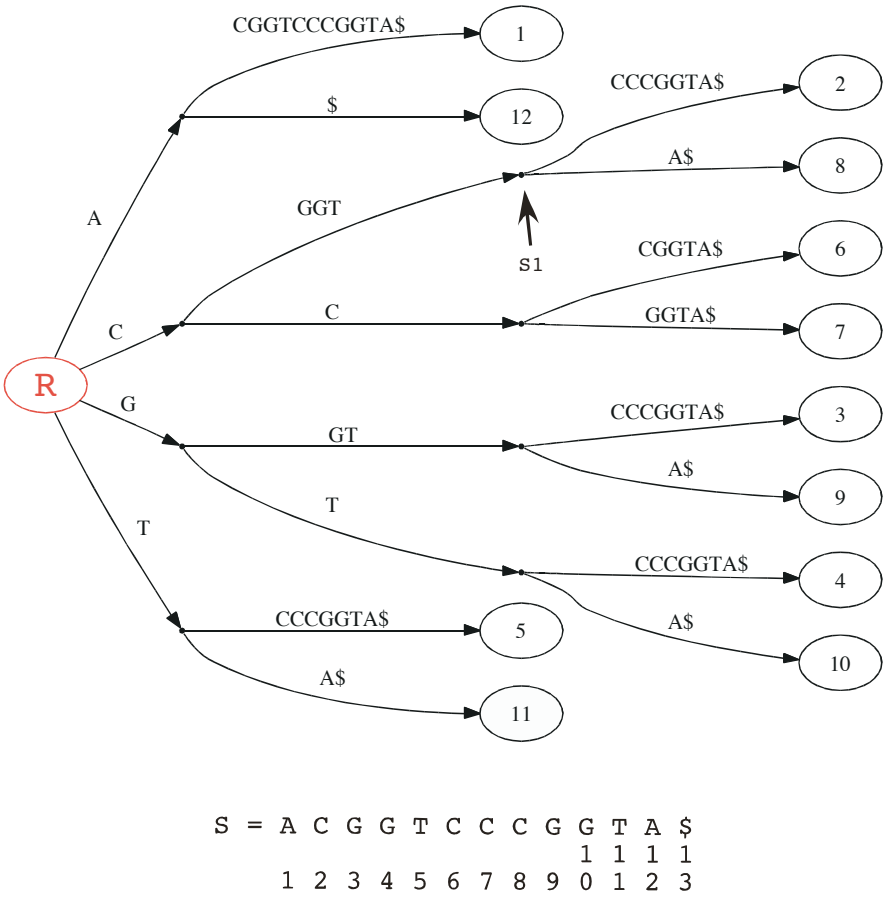


FIG. 3.2 – **Structure d'un arbre des suffixes** – L'arbre des suffixes a été construit dans cette figure pour la séquence S. La séquence d'origine a été allongée d'un symbole supplémentaire (ici le \$) non présent dans la séquence d'origine. Cette modification de la séquence assure qu'aucun suffixe de la séquence n'est égal à un de ces préfixes. Cette propriété est indispensable pour la construction de l'arbre. Chaque feuille de l'arbre représente un suffixe et est étiquetée par la position de début du suffixe correspondant dans la séquence. Le chemin allant de la racine au noeud interne S_1 représente par la concaténation des deux étiquettes d'arc C et GGT le facteur $CGGT$ que l'on sait répété deux fois à la position 2 et 8 car les deux feuilles correspondant à ces suffixes sont filles du noeud S_1.

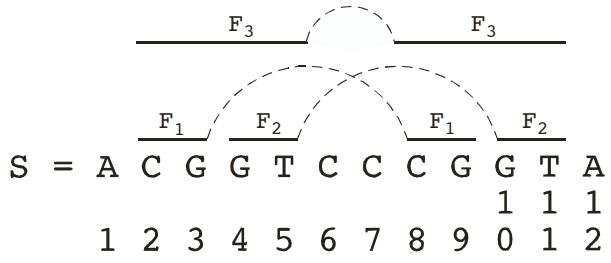


FIG. 3.3 – **Principe de base de l'algorithme *KMR*** – Dans cette séquence exemple F_1 est un premier facteur de deux lettres CG présent en position 2 et 8. F_2 est un second facteur de deux lettres aussi GT présent également en deux copies. Les deux copies de F_2 étant contiguës aux deux occurrences de F_1, il est possible de les concaténer pour former un facteur répété F_3 de quatre lettres $CGGT$. De la même façon, le facteur F_1 peut être vu comme la concaténation de deux facteurs répétés de une lettre C et G

.

du motif et N la longueur du texte. Ce résultat est comparable, voire dans certains cas meilleur, que celui obtenu avec les arbres de suffixes. Mais, le réel avantage des tables de suffixes est qu'elles sont beaucoup moins exigeantes en mémoire.

Une table des suffixes est un simple tableau d'entiers contenant la position de début de chacun des suffixes d'une séquence ordonnée par ordre lexicographique (*voir* FIG. 3.4). Elle occupe donc 4 octets par nucléotide dans notre cas. Une table des suffixes peut être construite avec une complexité en temps $O(N \log N)$ dans le pire cas. Il est possible de construire une table des suffixes en temps linéaire, mais au prix d'un surcoût de mémoire qui rend cette option moins intéressante dans notre cas (Manber et Myers, 1990; Kärkkäinen et Sanders, 2003).

Pour ce qui est de la recherche des répétitions, une forme améliorée des tables de suffixe a été proposée par Abouelhoda *et coll.* (2002). L'amélioration correspond à l'ajout de tables annexes à la table des suffixes proprement dite qui permettent une recherche plus efficace des motifs répétés. Parmi ces tables annexes, la table des LCP pour «*longest common prefix*» est certainement la plus importante (*voir* FIG. 3.4). La structure ainsi modifiée continue d'occuper une place réduite estimée à 5 octets par nucléotide analysé (en moyenne). Cette compacité en mémoire donne à cette structure de données un intérêt justifiant d'envisager son utilisation en lieu et place de *KMR* pour nos recherches de répétitions exactes.

Statistiques associées aux répétitions exactes

Lorsqu'on recherche des répétitions strictes dans une séquence, il est important de déterminer la longueur minimale d'une répétition dont l'occurrence ne peut pas être considérée comme le fruit du hasard. D'un point de vue algorithmique, rien n'interdit aux méthodes de recherche précédemment décrites d'identifier des facteurs répétés maximaux de longueur 1. Il est trivial, dans ce cas extrême, de se rendre compte que ces répétitions ne sont pas significatives, puisque dans une séquence d'une longueur $N \gg 4$ chaque symbole d'un alphabet à quatre lettres, comme l'alphabet nucléique, se retrouvera un grand nombre de fois dans cette séquence.

L'avantage des répétitions strictes, d'un point de vue statistique, est qu'il existe une expression formelle de la distribution de la variable aléatoire $L^{(N)}_{max\ r}$, qui est défini comme la longueur de la plus grande séquence présente r fois dans une séquence de taille N (par

```
S = A C G G T C C C G G T A
                    1 1 1
    1 2 3 4 5 6 7 8 9 0 1 2
```

LCP	POS	Suffixe
0	12	A
1	1	A C G G T C C C G G T A
0	6	C C C G G T A
2	7	C C G G T A
1	8	C G G T A
4	2	C G G T C C C G G T A
0	9	G G T A
3	3	G G T C C C G G T A
1	10	G T A
2	4	G T C C C G G T A
0	11	T A
1	5	T C C C G G T A

FIG. 3.4 – **Structure d'une table des suffixes** – La table des suffixes est un tableau d'entiers représentant la position de début de chacun des suffixes d'une séquence S classés par ordre lexicographique. Elle peut être associée à d'autres tables annexes comme celle des LCP (Longest Common Prefix) qui indique la longueur du préfixe commun entre le suffixe débutant en position S_i et S_{i-1}. Les LCP sont particulièrement utiles dans le cas de la recherche de facteurs répétés.

la suite $L^{(N)}_{max\ 2}$ sera noté L_{max}). Cette distribution décrite par Karlin et Ost (1985) est basée sur un modèle d'indépendance des positions, c'est-à-dire qu'en chaque position de la séquence la probabilité d'apparition d'un des symboles est la même et cela indépendamment des symboles présents aux autres positions.

Dans cette publication Karlin et Ost montrent que, pour une grande séquence de taille N, $L^{(N)}_{max\ 2}$ suit une loi normale dont la moyenne est décrite par la formule (3.1) et la variance par la formule (3.2),

$$E\left(L^{(N)}_{max\ r}\right) = -\frac{1}{\log \lambda^{[r]}}\left[\log C_N^r + \log\left(1-\lambda^{[r]}\right) + \log \lambda^{[r]} + 0,5772\right] + 0.5 \tag{3.1}$$

$$Var\left(L^{(N)}_{max\ r}\right) = 1.645\left(\frac{1}{\log \lambda^{[r]}}\right)^2 \tag{3.2}$$

avec P_j la fréquence relative du nucléotide j (A, C, G ou T) dans la séquence.
et avec λ définie de la façon suivante :

$$\lambda^{[r]} = \sum_{j\in\{A,C,G,T\}} P_j^r \tag{3.3}$$

Par exemple, en s'imposant un risque de 1/1000, il est possible de calculer L_{max}, la longueur de la plus grande répétition pouvant apparaître par hasard dans notre séquence, et d'utiliser ce seuil comme taille minimale (L_{min}) en dessous de laquelle les répétitions ne seront pas conservées, car considérées comme fortuite. À titre d'exemple, L_{min} varie entre 21 et 27 pour les chromosomes bactériens (Rocha *et coll.*, 1999).

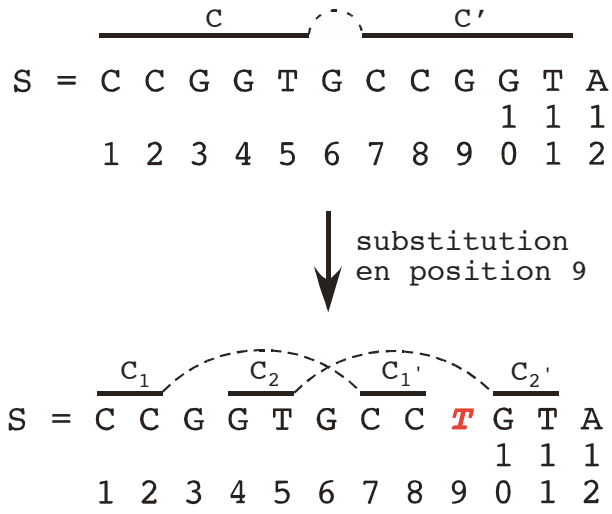


FIG. 3.5 – **Action de l'évolution sur les répétitions strictes** – Lorsqu'une mutation est introduite dans l'une des deux copies d'une répétition stricte (en position 9 de la séquence S) la répétition initialement identifiée comme formée de deux copies C et C' est maintenant détectée, par les algorithmes de détection de répétitions exactes, comme deux répétitions formées des copies C_1 et C'_1 d'une part et C_2 et C'_2 d'autre part.

Détection des répétitions approchées

Les répétitions dans les séquences de chromosomes sont associées à des mécanismes dynamiques les créant et les effaçant. Nous pouvons supposer que juste après la genèse d'une duplication les deux copies sont strictement identiques. Elles peuvent donc parfaitement être détectées par les algorithmes décrits précédemment. Par contre, sitôt qu'une mutation est fixée dans l'une des deux copies, ces algorithmes trouveront deux séquences répétées à la place d'une seule (*voir* FIG. 3.5). Ainsi, au fur et à mesure que des mutations seront accumulées dans les copies, la répétition sera identifiée sous la forme d'un nombre croissant de petites répétitions strictes. Il est clair que d'un point de vue biologique chacune des petites répétitions identifiées, dans ce cadre, ne peuvent pas être considérée comme des événements de duplications indépendants. Il est donc indispensable d'introduire la notion de répétitions approchées, où les deux copies ne sont pas strictement identiques.

Les algorithmes exacts

Les algorithmes exacts pour la détection des répétitions approchées dépendent du type de mutation que l'on souhaite prendre en compte. Si l'on ne souhaite considérer que les substitutions, il existe des modifications des algorithmes de recherche de répétitions exactes qui introduisent la notion d'erreur dans leur détection. Par exemple, il est possible de modifier *KMR* en introduisant une relation de similarité non transitive entre les symboles (Soldano *et coll.*, 1995). Ceci permet de retrouver des motifs répétés non strictement identiques.

Recherche par programmation dynamique

Le plus souvent, il est souhaitable de prendre en compte en plus des substitutions, les mutations de type insertion – délétion. Ces dernières ne peuvent être prises en compte que par l'intermédiaire de méthodes d'alignement utilisant les principes de la programmation dynamique. Pellegrini *et coll.* (1999) ont proposé une méthode de recherche des répétitions

inexactes basée sur une version modifiée de l'algorithme de Waterman et Eggert (1987) permettant d'identifier l'ensemble des alignements locaux optimaux et sub-optimaux existant entre deux séquences. L'idée de cet algorithme est d'aligner une séquence avec elle-même en utilisant un algorithme de recherche du meilleur sous-alignement local. Le meilleur alignement local existant entre une séquence et elle-même est de manière triviale l'alignement de chacune des positions de la séquence avec elle-même. Cet alignement est sans intérêt. Il faut donc l'éliminer et prendre l'alignement local sub-optimal suivant. Ce nouvel alignement correspondra, pour une séquence S de longueur l, à l'alignement de la sous séquence S_{i-j} débutant à la position i et finissant à la position j où $0 \leqslant i \leqslant j \leqslant l$ et la sous séquence S_{m-n} avec $m \neq i$ et $n \neq j$. Si cet alignement est détecté, c'est que ces deux sous-séquences se ressemblent et qu'elles correspondent à une répétition dans la séquence S. Il est possible de rechercher les sous-alignements sub-optimaux suivants afin de détecter les autres répétitions présentes dans la séquence S mais possédant un score d'alignement plus faible.

Cette approche, basée sur un algorithme d'alignement par programmation dynamique, possède une complexité en temps et en espace en $O(N^2)$ où N est la taille de la séquence analysée. Elle ne peut donc être appliquée qu'à des séquences de taille modeste (de l'ordre de quelques milliers de symboles). Elle n'est donc pas utilisable pour détecter les répétitions au niveau de la séquence d'un chromosome bactérien qui possède une taille de plusieurs millions de paires de bases. Par contre, nous utilisons une version légèrement modifiée de l'algorithme décrit par Pellegrini *et coll.* (1999) dans le projet *EvolRep* (voir page 57) qui vise à étudier les répétitions présentes à l'intérieur d'un gène, donc sur une séquence courte.

Pour mémoire, Smith et Waterman (1981) décrivaient la recherche du meilleur sous alignement entre une séquence A de longueur l_A et une séquence B de longeur l_B en utilisant un algorithme de programmation dynamique nécessitant la construction d'une matrice de scores S de dimension $(l_A{+}1, l_B{+}1)$. Si l'on considère le cas simple où les insertions/délétions sont réalisées caractère par caractère, les valeurs présentent dans la matrice S peuvent être décrites par la formule de récurrence (3.4).

$$S_{i,j} = \begin{cases} \max \begin{cases} S_{i-1,j-1} + score(A_i, B_j), \\ S_{i-1,j} + \omega, \\ S_{i,j-1} + \omega, \\ 0 \end{cases} & \text{si } i \neq 0 \text{ ou } \ j \neq 0, \\ \\ 0 & \text{si } i = 0 \text{ et } \ j = 0 \end{cases} \qquad (3.4)$$

$$\text{avec } 0 \leqslant i \leqslant l_A \\ \text{et } 0 \leqslant j \leqslant l_B$$

- $S_{i,j}$ est le score, indiqué dans la matrice S, à la position correspondant au i^{me} caractère de la sequence A et au j^{me} caractère de la séquence B.
- $score(A_i, B_j)$ est une fonction de score indiquant la similarité entre les symboles A_i et B_j tel que $score(A_i, B_j) > 0$ si $A_i = B_j$ et $score(A_i, B_j) < 0$ pour au moins certains cas ou $A_i \neq B_j$.
- ω est le coût d'une insertion (gap symbolisé par le caractère '-') dans une séquence ($\omega < 0$)

Dans le cas particulier de l'alignement d'une séquence avec elle-même, la séquence A étant identique à la séquence B il apparaît une symétrie dans la matrice des scores S puisque les cases $S_{i,j}$ et $S_{j,i}$ mettent en relation les deux mêmes symboles A_i et A_j. La figure 3.6 montre cette propriété en présentant les chemins correspondant aux trois meilleurs alignements d'une séquence possédant une répétition interne. La répétition peut donc être identifiée à partir des deux chemins dessinés en vert. Il devient clair au vue de ce schéma que la formule de

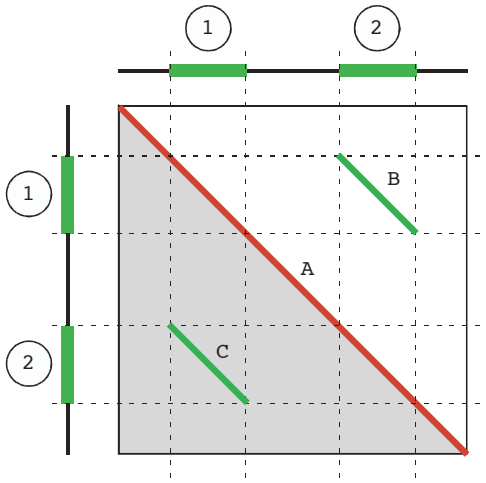


FIG. 3.6 – **Recherche des répétitions approchées par programmation dynamique** – Les premiers chemins optimaux dans la matrice des scores correspondant à l'alignement d'une séquence présentant une répétition peuvent être schématisés comme ci-dessus. La diagonale rouge (A) indique le chemin de l'alignement de la séquence avec elle-même sur l'ensemble de sa longueur. Cet alignement qui correspond à l'alignement de meilleur score n'a aucun intérêt dans le cadre de la recherche de répétitions. Par contre, les deux chemins sub-optimaux dessinés en vert correspondent eux à la mise en évidence d'une répétition matérialisée par des segments numérotés 1 et 2 sur la séquence. Cependant, le chemin B qui représente l'alignement de la copie 1 de la séquence en ordonnée avec la copie 2 de la séquence en abscisse et le chemin C qui représente l'alignement de la copie 2 de la séquence en ordonnée avec la copie 1 de la séquence en abscisse identifient la même répétition. C'est la symétrie de la matrice selon le chemin optimal rouge qui induit cette double détection. Il est dont inutile de tenir compte de la partie grisée de la matrice des scores.

récurrence (3.4) peut, pour la résolution du problème de recherche des répétitions internes, être avantageusement remplacée par la formule (3.5).

$$S_{i,j} = \begin{cases} \max \begin{cases} S_{i-1,j-1} + score(A_i, A_j), \\ S_{i-1,j} + \omega, \\ S_{i,j-1} + \omega, \\ 0 \end{cases} & \text{si } i > j, \\ \\ 0 & \text{si } i \leqslant j \end{cases} \tag{3.5}$$

La différence entre la formule (3.4) et la formule (3.5) est que $S_{i,j}$ prend systématiquement une valeur nulle :

- pour toutes les cases de la diagonale principale ($i = j$) ce qui a pour effet d'éliminer de fait l'alignement trivial de la séquence avec elle-même.
- pour toutes les cases ou $i < j$, cases correspondant à la partie grisée de la matrice dans la figure 3.6, annulant ainsi l'effet de symétrie.

Cette différence rend donc inutile le calcul de toutes les cases $S_{i,j}$ où $i \leqslant j$ d'où le gain d'un facteur deux en temps et en espace sur la complexité de l'algorithme.

Les heuristiques

Le problème de la détection des répétitions approchées sur une grande séquence reste donc complet puisque la complexité en $O(N^2)$ des algorithmes exacts rend ceux-ci inutilisables

pour des séquences de taille supérieure à quelques milliers de caractères (nucléotides ou acides aminés). Seule une méthode de détection basée sur une heuristique reste donc possible.

Heuristique basée sur un post-traitement des résultats de BLAST

Des tentatives de résolution de ce problème par une utilisation détournée du programme BLAST (Altschul *et coll.*, 1997) ont conduit à la publication de programme comme BLASTER (Quesneville *et coll.*, 2003). Dans ce programme, les répétitions sont recherchées en construisant une banque de séquences constituée de fragments chevauchants de la séquence à analyser. Chacun des fragments constituant la banque est ensuite comparé par BLAST avec tous les autres fragments de la banque. Une étape de post-traitement est ensuite réalisée sur les sorties de BLAST pour reconstruire les répétitions. Cette solution, qui offre l'avantage d'une mise en oeuvre simple, présente plusieurs limites dont la plus significative est certainement le cas des grandes répétitions s'étendant sur plusieurs fragments.

Heuristique de type «auto-BLAST»

La stratégie la plus souvent utilisée pour détecter les répétitions approchées dans les grandes séquences pourrait être qualifiée du nom d'«auto-BLAST» . Dans son principe, elle peut être divisée en deux grandes étapes :

1. la détection des répétitions exactes par un des algorithmes décrits précédemment.
2. l'utilisation de ces répétitions strictes comme amorce à une étape d'alignement des régions flanquant la répétition stricte par un programme d'alignement autorisant les insertions/délétions

Plusieurs travaux développent une approche de ce type pour détecter les répétitions (Kurtz *et coll.*, 2000, 2004; Lefebvre *et coll.*, 2003). Nous avons nous-mêmes développé le programme *RepSeek* pour répondre à ce problème (Achaz *et coll.*, 2005). Dans *RepSeek*, nous avons implémenté une stratégie complète de détection des répétitions approchées incluant l'identification des amorces (répétitions strictes) qui utilise notre implémentation compacte de *KMR* et une phase d'extension de ces graines par programmation dynamique. Nos travaux sur les répétitions nous ont permis de mettre en évidence différents problèmes liés à la détection des répétitions dans des génomes complets. Parmi ceux-ci, les deux principaux sont :

- l'influence du biais de composition des séquences sur la longueur des répétitions détectées.
- la significativité statisitique des répétitions identifiées.

Plusieurs solutions à ces problèmes ont été incluses dans le programme *RepSeek*.

Correction des biais de composition

Lorsque deux séquences sont alignées, les algorithmes utilisés cherchent à maximiser la concordance des symboles appariés entre les deux séquences. La mise en correspondance de deux symboles dans un alignement peut être due à deux causes : l'homologie des deux symboles ou le hasard. L'homologie entre deux symboles signifie qu'ils correspondent au même nucléotide (ou amino-acide pour les protéines) dans la séquence ancestrale. Seuls les appariements dus à l'homologie sont intéressants du point de vue de la biologie, mais comme le montre la figure 2.2, il n'est pas possible d'ignorer les appariements dus au hasard. Ce fait est d'autant plus vrai si l'on travaille sur des séquences nucléiques qui sont décrites avec un alphabet de taille 4 (A, C, G, T) et s'il existe un biais de composition dans la séquence éloignant la distribution des nucléotides de l'équiprobabilité des symboles $(0,25; 0,25; 0,25; 0,25)$.

Dans un modèle purement aléatoire, la probabilité d'aligner deux symboles identiques est égale à P_X^2 où X est le nucléotide considéré et P_X la probabilité d'occurrence du symbole.

En cas de déséquilibre de la distribution, la probabilité d'aligner, entre eux par hasard, le ou les symboles majoritaires est donc plus élevée.

$$Subst(i,j) = \begin{cases} +1 & \text{si } i = j, \\ -1 & \text{si } i \neq j \end{cases} \tag{3.6}$$

avec i et j deux symboles de la séquence

$$Subst(N,i) = 1/4 \tag{3.6'}$$

$$Gap_{Open} = -4 \tag{3.7}$$

$$Gap_{Ext} = -1 \tag{3.8}$$

Dans un système de score d'alignement standard pour les acides nucléiques toutes les identités sont pondérées de manière identique (voir equations 3.6 à 3.8). Il s'en suit des gains de score dus à ces appariements aléatoires importants lorsque la composition en nucléotides est déséquilibrée. Lors de la construction d'un alignement local par un algorithme de type Smith et Waterman cela provoque un allongement significatif du sous-alignement identifié (voir la courbe rouge, FIG. 3.7). Le phénomène est identique lors de la recherche de répétitions qui est d'un point de vue algorithmique similaire à un alignement local.

Afin de limiter ce phénomène, il est nécessaire d'utiliser un modèle de rétribution tenant compte du biais de composition. Empiriquement, plusieurs modèles de matrices de substitutions ont été essayés. Dans le cas de la recherche de répétitions, le modèle décrit par les équations (3.9) et (3.9') permet d'obtenir, sur des séquences aléatoires, des longueurs de répétitions similaires pour des variations de taux de GC correspondant à ceux observés dans les génomes bactériens (*voir* FIG. 3.7).

$$Subst(i,j) = \frac{\sigma(i,j)}{2}.\log_4(p_i.p_j)$$
$$\text{où } \sigma(i,j) = \begin{cases} +1 & \text{si } i = j, \\ -1 & \text{si } i \neq j \end{cases} \tag{3.9}$$

$$Subst(N,i) = pi.Subst(i,i) = \frac{p_i}{2}.\log_4(p_i^2) = p_i.\log_4 p_i \tag{3.9'}$$

Statistiques associées aux répétitions approchées

Le problème de la pertinence biologique des répétitions détectées est posée pour les répétitions approchées comme il existe pour les répétitions exactes. La différence avec les répétitions approchées est qu'il n'y a pas de modèle statistique analytique permettant d'affecter à une répétition approchée la probabilité qu'elle soit due au hasard. Il existe par contre plusieurs approximations permettant d'estimer cette probabilité. Leur mise en oeuvre nécessite généralement un temps de calcul non négligeable relativement au temps de détection de la répétition puisqu'elle nécessite l'ajustement de paramètres sur une distribution empirique de scores d'alignement.

Pour valider les répétitions approchées que nous avons détectées dans les séquences des chromosomes, nous avons utilisé deux approches différentes : l'une basée sur les statistiques décrites pour les répétitions exactes, l'autre sur une simulation.

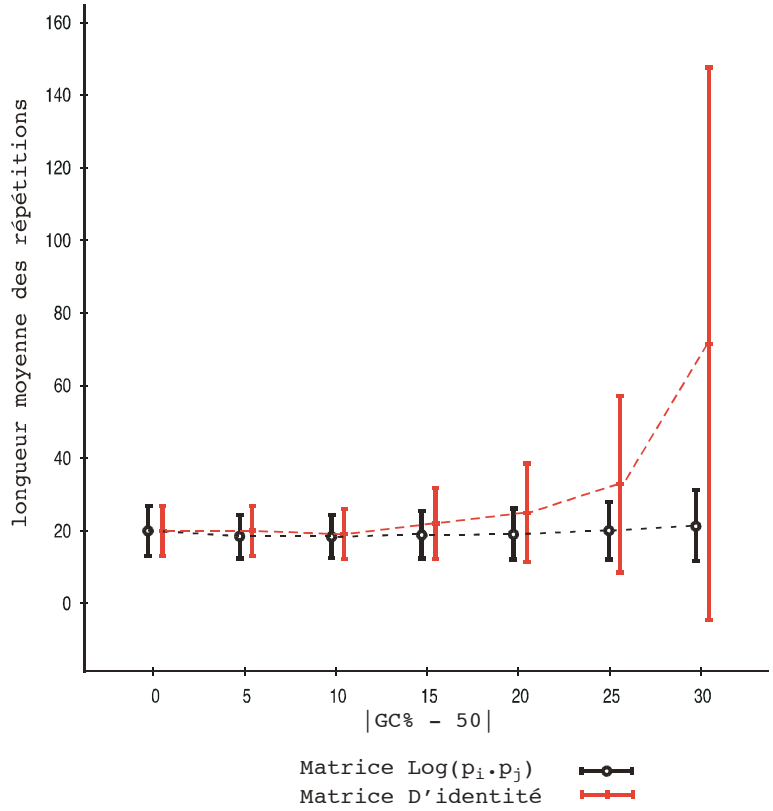


FIG. 3.7 – **Rapport entre longueur des répétitions et biais de composition de la séquence** – La longueur moyenne des répétitions et l'écart type associé sont donnés en fonction du biais de composition de la séquence nucléotidique associée. Pour chaque composition (|GC% -50| en abscisse), dix séquences aléatoires de 10 *Mb* ont été générées et les répétitions présentes sur ces séquences ont été recherchées en utilisant *RepSeek* avec $L_{min} = 15$. L'alignement (pour étendre les répétitions strictes) a été réalisé soit en utilisant la matrice d'identité (courbe rouge) ou la matrice log (courbe noire).

Statistiques basées sur les répétitions exactes

La technique de détermination des répétitions approchées utilisée pour l'analyse des séquences de chromosomes est divisée en deux étapes. La première de ces deux étapes consiste à identifier des répétitions exactes qui serviront de graines pour la détection des répétitions approchées.

Pour valider statistiquement les répétitions approchées, il est possible de s'appuyer sur cette première partie en posant comme principe que toute répétition approchée construite à partir d'une répétition exacte significative sera elle-même significative. Cette vision est conservatrice puisqu'à l'inverse il est simple d'imaginer qu'une répétition approchée fortement dégénérée ne possède aucune répétition exacte qui puisse être considérée comme significative. Elle offre, par contre, l'avantage de permettre l'utilisation du modèle de Karlin et Ost (1985) tel qu'il a été présenté précédemment (voir page 39, Statistiques associées aux répétitions exactes).

Sa mise en place est simple puisqu'il suffit de rechercher les répétitions exactes de longueur $l > L_{min}$ et de ne considérer que ces graines pour la seconde phase d'extension.

Statistiques basées sur l'utilisation d'un modèle empirique

Une autre façon de tester la pertinence d'une répétition approchée est de s'appuyer sur des simulations pour déterminer le score d'alignement minimum permettant de la valider.

Pour construire un test comparable à celui réalisé lors de l'utilisation du modèle de Karlin et Ost où la probabilité déterminée correspond à l'existence de la plus grande répétition présente dans la séquence, il faut déterminer pour un ensemble de génomes aléatoires la plus longue répétition approchée présente. De par ce principe, chaque génome aléatoire ne fournit qu'un seul point à la distribution empirique. Pour obtenir une précision permettant d'estimer des *p-value* de l'ordre de 10^{-3}, il est nécessaire d'analyser plusieurs milliers de génomes aléatoires ce qui n'est pratiquement pas possible.

Une solution pour rendre cette approche empirique réalisable est de changer la question associée au test statistique qui était initialement :

> *Quelle est la probabilité d'identifier au moins une répétition avec un score $s \geqslant S$ dans un génome aléatoire de taille L ?*

de la manière suivante :

> *Quelle est la chance d'obtenir une répétition dans une séquence aléatoire avec un score $s \geqslant S$?*

Dans ce cas, le test ne porte plus sur le génome complet mais sur chaque répétition individuellement et chaque génome aléatoire apporte autant de points à la distribution empirique qu'il comporte de répétitions.

Si nous prenons comme exemple *Escherichia coli K12*, les résultats intermédiaires du programme *RepSeek* montrent que si l'on utilise des mots répétés de taille 15 comme graines pour initier la recherche des répétitions, il est possible d'identifier dans ce génome de $4,64$ Mb, 40376 répétitions approchées. Cela indique qu'il sera nécessaire de réaliser $N = 40376$ tests pour filtrer chacune de ces répétitions. Si nous considérons pour chaque répétition un risque $\alpha = 10^{-3}$, pour l'ensemble du génome le risque réel est de :

$$\begin{aligned} \alpha_{E.\ coli} &= 1 - (1-\alpha)^N \\ &= 1 - 0,999^{40376} \\ &\approx 1 \end{aligned} \tag{3.10}$$

Ce résultat n'est pas surprenant puisque pour chaque test, nous acceptons de nous tromper une fois sur mille et que nous réalisons plus de 40000 tests. En moyenne, nous nous

attendons donc à accepter comme de bonnes répétitions $40376 \times 10^{-3} \approx 40$ répétitions dues au hasard (faux positifs).

Sur ces 40376 répétitions identifiées, 12434 sont significatives d'après le test statistique basé sur une distribution empirique construite à partir de deux génomes aléatoires soit 49610 répétitions. Parmi ces 12434 répétitions conservées par le programme, 40 sont des faux positifs, soit $0,3\%$ des répétitions. Cette très faible proportion de faux positifs n'est absolument pas gênante pour la plupart des analyses réalisées.

A titre de comparaison, la même recherche basée sur les statistiques associées aux répétitions exactes permet d'identifier 3596 répétitions directes pour $P = 10^{-3}$ soit $L_{min} = 24$. Nous voyons que le gain de sensibilité est loin d'être négligeable en comparaison du risque d'identification de fausses répétitions.

Chapitre 4

Un modèle pour la genèse et l'évolution des duplications

L'analyse de l'organisation des familles multigéniques le long des chromosomes de *Saccharomyces cerevisiae* a permis de construire des modèles expliquant certains modes de remaniement du génome de la levure (Coissac *et coll.*, 1997; Wolfe et Shields, 1997; Kellis *et coll.*, 2004). Ces travaux ne considèrent que de grands réarrangements impliquant des régions chromosomiques couvrant plusieurs gènes. De ce point de vue très macroscopique, il n'est pas étonnant que le fait prouvé le plus marquant : la duplication complète du génome (Kellis *et coll.*, 2004), possède lui-même cet aspect de globalité. Le but d'une analyse plus fine, réalisée directement au niveau de la séquence nucléotidique, est d'obtenir des informations sur des remaniements éventuellement plus petits qui permettront une analyse plus mécanistique de la dynamique des répétitions.

L'analyse du génome de *Saccharomyces cerevisiae*

Comme cela c'est chronologiquement passé, je commencerai cette présentation du modèle de «genèse et d'évolution des duplications» à travers l'analyse des répétitions présentes dans le génome de *Saccharomyces cerevisiae*.

Pour des raisons techniques, l'analyse des répétitions au niveau nucléique n'a pas été menée globalement sur l'ensemble du génome, mais indépendamment pour chacun des seize chromosomes de la levure. Toutes les répétitions analysées sont donc des répétions intra-chromosomiques (les deux copies de la répétition sont localisées sur le même chromosome). Les répétitions inter-chromosomiques n'ont pas été recherchées.

Les duplications sont recherchées simultanément sur les deux brins de la séquence d'ADN. Deux grandes classes de répétitions sont donc décrites.

- La première où les deux copies de la séquence répétée sont localisées sur le même brin d'ADN. Elles seront nommées «répétitions directes» ou DR.
- La seconde classe correspond aux répétitions pour lesquelles les deux copies sont localisées chacune sur un brin. Nous les qualifierons de «répétitions inversées» ou IR.

Une fois cette dichotomie réalisée, trois caractéristiques principales de chacune des duplications (directes ou inverses) sont analysées (*voir* Fig. 4.1) :

ℓ : la longueur d'une copie de la répétition
Δ : la distance en nucléotides séparant les deux copies
% *d'identité* : indiquant la similarité entre les deux copies

Le modèles décrivant la dynamique des répétitions a été construit en se basant sur les relations de dépendances existant entre ces paramètres. Dans les paragraphes suivants je présenterai succinctement les caractéristiques des répétitions observées et le modèle construit à partir de ces observations.

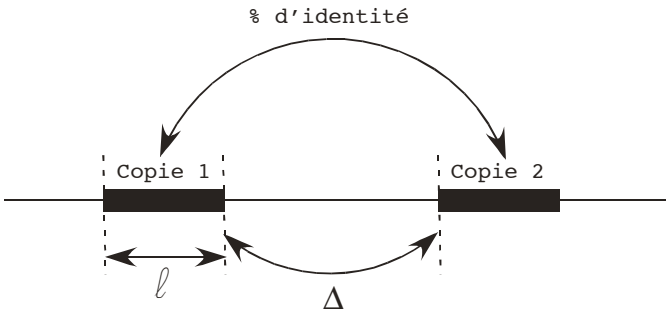


FIG. 4.1 – **Paramètres caractérisant une répétition** – ℓ indique la longueur de la séquence répétée en paires de bases. Comme les deux copies ne sont pas forcément identiques et que l'algorithme utilisé pour leur détection autorise les opérations d'insertions et de délétions, la longueur des deux copie n'est pas non plus obligatoirement identique. Cependant, ces deux longueurs sont fortement corrélées et l'utilisation de l'une ou de l'autre n'influe pas sur les résultats finaux. **Δ** indique, en paires de bases, la distance physique séparant les deux copies le long du chromosome. Cette distance étant mesurée de la fin de la copie 1 au début de la copie 2, elle n'est pas intrinsèquement influencée par ℓ. Par contre, elle peut être négative si les deux copies se chevauchent. % **d'identité** permet d'estimer la similarité entre les deux copies de la répétition.

Analyse de la distance séparant les deux copies d'une répétition

L'analyse de la distance entre les deux copies d'une répétition (Δ) permet d'identifier un comportement différent des répétitions directes et inversées. Les distributions de Δ pour ces deux populations de répétitions, telles qu'elles sont observables dans le génome de *Saccharomyce cerevisiae*, comparées à des distributions équivalentes obtenues à partir de génomes aléatoires montrent que les répétitions directes (DR) présentent une sur-représentation des faibles Δ. Les IR se comportent, par contre, de la même manière que les répétitions du génome aléatoire (*voir* FIG. 4.2). La sous-population des DR présentant un Δ faible sera dénommée dans la suite de ce mémoire par l'acronyme *CDR* pour «*close direct repeat*». Arbitrairement, un Δ sera considéré comme faible lorsqu'il sera inférieur à $1kb$.

> Les *CDR* sont donc définis comme l'ensemble des répétitions directes présentant un $\Delta < 1\ kb$

Cette valeur seuil a été choisie car, pour de nombreux organismes, les *CDR* définies de la sorte possèdent des propriétés qui leur donnent un rôle prépondérant dans le modèle décrivant la dynamique des répétitions (Achaz *et coll.*, 2000, 2001, 2002).

Les duplications sont sur-représentées dans les régions codantes

Les régions codant pour des protéines représentent environ 72% du génome de la levure. Une analyse de la localisation des répétitions montre que $85,6\%$ (119/139) des *CDR* possèdent leur deux copies complètement incluses dans des gènes (Achaz *et coll.*, 2000). Cette surabondance n'est vraie que pour cette catégorie de répétitions. L'origine de ce phénomène peut principalement être expliquée de deux manières. Les sites de cassure double brin (*Double Strand Break* ou *DSB*) sont, chez la levure, principalement localisés dans les régions promotrices des gènes (Baudat et Nicolas, 1997). Ces sites étant de forts activateurs de la recombinaison, ils pourraient induire une sur-représentation des duplications à leur proximité, donc au début des gènes. L'autre explication est qu'une duplication apparue dans un gène et qui a réussi à s'y fixer est contrainte d'y rester par la pression fonctionnelle s'exerçant sur le gène, même si la duplication n'a pas elle-même de rôle fonctionnel. Son excision risquerait à

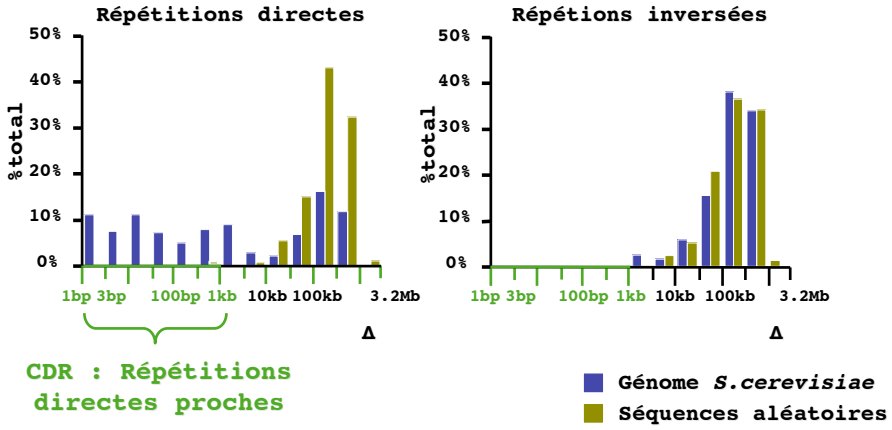


FIG. 4.2 – **Distributions de la distance entre les deux copies d'une répétition** – Les répétitions inexactes ont été identifiées dans chacun des 16 chromosomes de la levure *Saccharomyces cerevisiae*, ainsi que dans 10 jeux de séquences aléatoires de tailles identiques aux chromosomes de la levure (Achaz *et coll.*, 2000). Les deux histogrammes présentent la distribution de Δ (*voir* FIG. 4.1) pour les répétitions directes (DR) et inversées (IR).

son tour de détruire le gène. Il n'est pas simple de trancher entre les deux hypothèses. Pour le moment, des résultats préliminaires me font privilégier la seconde. Les avancées du projet EvolRep décrit dans le chapitre suivant apporteront peut-être des réponses permettant d'estimer l'influence des deux phénomènes.

Analyse des relations existant entres les paramètres qualifiant une répétition

Les paramètres mesurés pour les répétitions de la classe des *CDR* présentent des propriétés intéressantes. Pour cette catégorie de répétions, deux corrélations ont pu être observées entre d'une part Δ et le % *d'identité* (*voir* FIG. 4.3) et d'autre part entre Δ et ℓ (*voir* FIG. 4.4).

Corrélation entre % *d'identité* et Δ

Des deux corrélations, la principale est celle liant % *d'identité* et Δ (*voir* FIG. 4.3). Elle doit être mise en relation avec le phénomène de conversion génique. Dans notre contexte, la conversion peut être vue comme un mécanisme permettant l'homogénéisation des séquences. Si une séquence est dupliquée, l'apparition de mutations dans l'une ou l'autre des copies conduit les deux séquences à diverger de plus en plus au cours du temps. La conversion génique va recopier la séquence d'une des deux copies (en totalité ou en partie) en lieu et place de la seconde. Ce transfert va, selon sa direction, soit corriger les mutations présentes dans une copie grâce à la séquence d'origine présente sur l'autre copie, soit recopier la mutation à la place de la séquence d'origine. Mais, quelle que soit la direction, il en résulte qu'après un événement de conversion les deux séquences cibles sont identiques.

D'un point de vue moléculaire la conversion est un sous-produit de la recombinaison homologue. On peut donc postuler que son efficacité est liée à celle de la recombinaison. Des travaux expérimentaux ont montré que, chez les bactéries *Escherichia coli* (Lovett *et coll.*, 1994) et *Bacillus subtilis* (Chedin *et coll.*, 1994), l'efficacité de la recombinaison entre deux séquences répétées chute avec l'éloignement des deux séquences cibles.

La similarité (% *d'identité*) observée entre les deux copies d'une duplication est le résultat

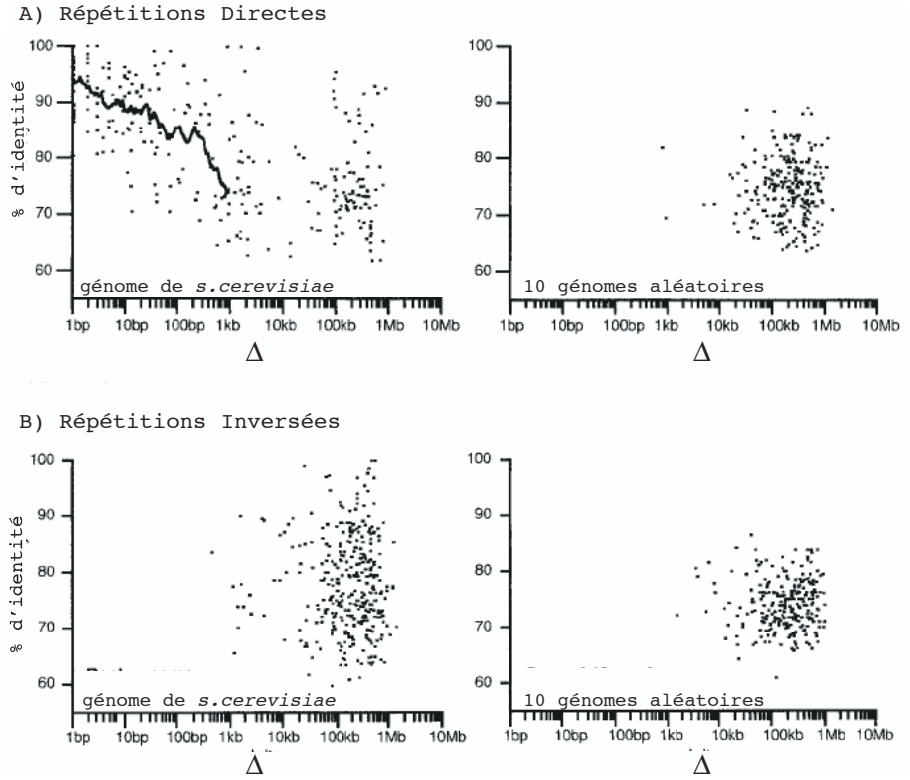


FIG. 4.3 – **Corrélation entre Δ et** % *d'identité* – Pour la sous-population des *CDR* le paramètre % *d'identité* est corrélé avec Δ. Cette corrélation est observable sur la courbe en haut à gauche. La courbe noire est la courbe moyenne de l'identité pour un Δ donné. Cette corrélation significative selon un test de Kendall-τ ($\tau = -0,36$; $p \approx 10^{-10}$, Kendall, 1938 ; Hollander et Wolfe, 1973) n'est observable ni pour les répétitions inversées, ni pour les génomes aléatoires.

de l'équilibre entre deux forces opposées : la divergence par mutation et la convergence par conversion, cette dernière étant d'autant plus forte que les séquences sont localisées proches l'une de l'autre et que la similarité de séquence est forte. La résultante que l'on peut supposer de ce double phénomène est que les deux copies d'une duplication divergent d'autant plus lentement qu'elles se trouvent proches sur le chromosome.

Corrélation entre ℓ et Δ

La seconde corrélation observée (*voir* FIG. 4.4) est elle aussi certainement en rapport avec l'efficacité des mécanismes de recombinaison. Si la distance séparant les deux copies et la similarité de leurs séquences sont deux paramètres qui influent sur l'efficacité de la recombinaison, la longueur des séquences cibles (ℓ) est un troisième paramètre non négligeable. On peut penser que la probabilité de recombinaison entre deux séquences augmente avec la longueur de celles-ci.

La résolution des structures de Holliday (Holliday, 1964; Potter et Dressler, 1976) lors d'un événement de recombinaison entre les deux copies d'une duplication directe conduit une fois sur deux à l'excision d'une des deux copies et de la région de jonction. Une duplication a donc d'autant moins de chance de se fixer qu'elle est soumise à une forte activité de recombinaison. Une répétition proche et longue est donc particulièrement instable.

Une autre possibilité d'explication est de considérer l'impact d'une délétion sous son aspect fonctionnel. Si nous considérons que plus la distance augmente entre les deux copies plus il est envisageable que la région de jonction ait une fonction cellulaire, la délétion de

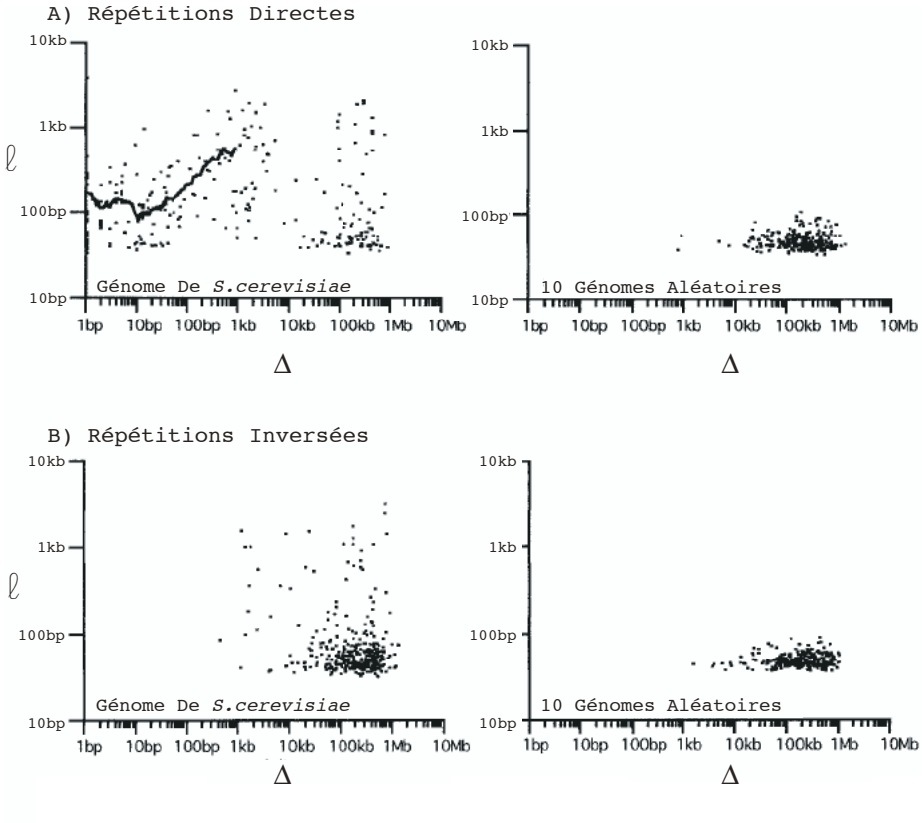


FIG. 4.4 – **Corrélation entre Δ et ℓ** – Comme pour la corrélation observable entre % *d'identité* et Δ (*voir* FIG. 4.3), la corrélation entre ℓ et Δ n'existe que pour les *CDR*. Elle est observable dans la courbe présentée en haut à gauche et est statistiquement significative d'après le test de Kendall-τ ($\tau = +0,26;\ p \approx 3 \cdot 10^{-6}$)

celle-ci risque d'être contre sélectionnée. Aucune donnée disponible actuellement ne nous permet d'estimer l'importance de cette dernière hypothèse, mais elle montre qu'une analyse plus avancée des duplications ne pourra continuer d'ignorer les aspects fonctionnels associées à celles-ci.

Le modèle

La prise en compte de ces différentes caractéristiques a permis de construire un modèle décrivant la genèse et l'évolution des duplications dans le génome de la levure (*voir* FIG. 4.5 ; Achaz *et coll.*, 2000). Ce modèle propose que les duplications apparaissent comme *CDR*. Rien ne permet, cependant, de faire des suppositions quant au mécanisme moléculaire générant ces duplications. Il n'est, par exemple, pas possible de dire si elles proviennent d'erreurs lors de la réplication ou de recombinaisons aberrantes.

Selon ce modèle, les autres duplications proviennent, au moins pour partie, des *CDR* par des remaniements chromosomiques ultérieurs. Ce postulat repose sur la plus faible similarité, en moyenne, des répétitions distantes en comparaison des *CDR* et sur l'organisation de certaines duplications qui peut être assimilée à des traces provenant de ces remaniements (Achaz *et coll.*, 2001).

Enfin, les corrélations observées s'expliquent, selon ce modèle, par l'équilibre entre les mutations se produisant sur chacune des copies d'une répétition et les mécanismes de conversion s'exerçant plus ou moins efficacement suivant la disposition des copies.

Extension du modèle aux autres organismes

Selon notre modèle, les duplications se caractérisent par les propriétés suivantes :

- une surabondance des *CDR*
- une corrélation entre % *d'identité* et Δ
- une corrélation entre ℓ et Δ
- une surabondance des répétitions dans les gènes
- la présence de traces de réarrangements

Ce modèle initialement construit à partir de l'analyse du génome de *Saccharomyces cerevisiae* est, selon les caractéristiques décrites ci-dessus, compatible avec l'analyse des répétitions réalisée sur d'autres génomes eucaryotes : *Caenorhabditis elegans*, *Drosphila melanogaster*, *Plasmodium falciparum*, *Arabidopsis thaliana*, et *Homo sapiens* (Achaz *et coll.*, 2001). Seul le génome de *Plasmodium falciparum* ne possède pas toutes les caractéristiques prédites par le modèle. L'apparente très grande plasticité de ce génome pourrait expliquer cette anomalie, en ne laissant pas le temps aux corrélations de s'établir (Achaz *et coll.*, 2001).

La validation du modèle pour les *Eubacteria* et les *Archae* n'est pas aussi claire que pour les *Eucaryota* (voir TAB. 4.1). L'analyse réalisée sur 52 génomes d'*Eubacteria* et d'*Archae* (Achaz *et coll.*, 2002) montre que toutes les caractéristiques des duplications prévues par le modèle ne sont pas présentes chez l'ensemble des organismes étudiés. L'absence de ces propriétés est parfois due à l'impossibilité de réaliser un test statistique significatif du fait d'un effectif de répétitions trop faible pour certaines espèces. Il faut aussi admettre que la diversité du monde des eubactéries et des archaebactéries est bien plus grande que celle des eucaryotes. Néanmoins, chez les espèces pour lesquelles les tests ont été réalisables, la surabondance des *CDR* existe, sauf chez *Buchnera sp.* La corrélation principale : % *d'identité versus* Δ, existe pour plus des 2/3 de ces espèces et la corrélation ℓ *versus* Δ est observée dans presque la moitié des cas (voir TAB. 4.1).

La présence quasi générale de la surabondance des *CDR* et la fréquente occurrence des corrélations laissent à penser que le modèle proposé s'applique à un très grand nombre d'organismes indépendamment du règne auquel ils appartiennent. Cela montre que les mécanismes mis en jeu doivent être très anciens. C'est la plus ou moins forte présence des

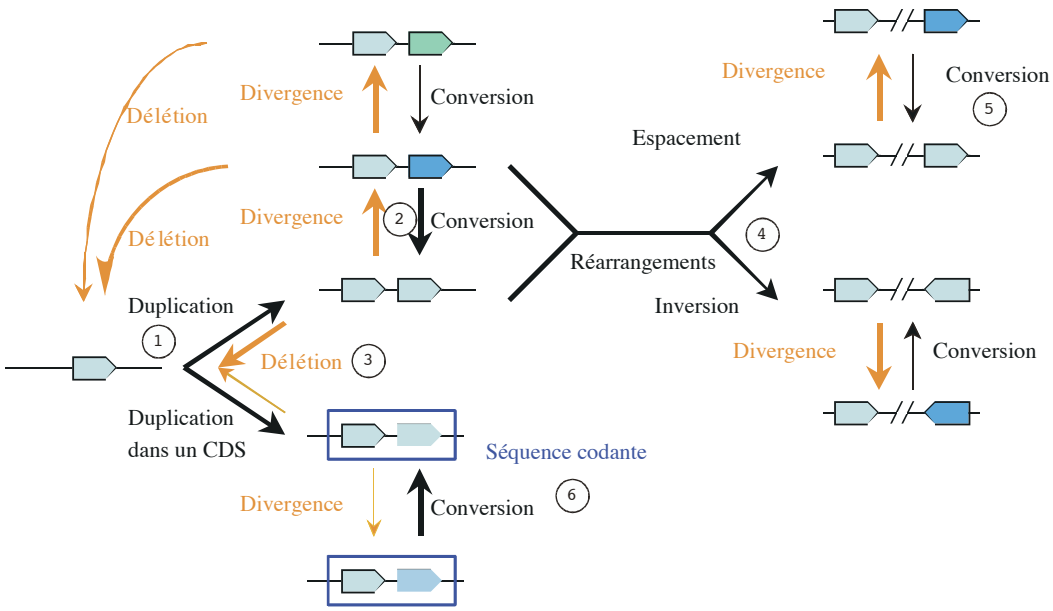


FIG. 4.5 – **Modèle de genèse et d'évolution des duplications** – Dans le schéma, les boites pleines en forme de flèches représentent une copie d'une duplication. La direction de la boite indique l'orientation de la copie. Les rectangles bleus du point (6) symbolisent les gènes incluant une répétition. Initialement les séquences subissent un événement de duplication (1) conduisant à la formation d'une *CDR*. Des mutations peuvent alors s'accumuler indépendamment dans chacune des deux copies (2). La proximité physique et la très forte similarité de séquence existant entre les deux copies permettent de limiter ce phénomène de divergence par conversion génique. De plus, les deux copies étant proches et ressemblantes, un événement de recombinaison les prenant pour cible peut permettre le retour à l'état initial (3). Des événements secondaires peuvent rendre impossible cette délétion et permettre à la délétion de se fixer : une augmentation de la divergence, un réarrangement éloignant les deux copies en les inversant éventuellement (4). Dans ce dernier cas, l'importance de la conversion chute, ce qui permet à la divergence inter copie d'augmenter (5). Enfin, les événements de duplication localisés à l'intérieur d'un gène sont piégés par la fonction du gène, même si la répétition n'influe pas sur cette fonction (6). Il devient donc plus difficile de réaliser l'opération de délétion d'où une surabondance des répétitions dans les zones codantes.

	Surabondance de CDR	Corrélation % *d'identité*/ Δ	Corrélation ℓ/ Δ	Traces de réarangement
Saccharomyces cerevisiae (Achaz *et coll.*, 2000, 2001)	+	+	+	+
Eucaryotes (Achaz *et coll.*, 2001)	5/5	4/5	4/5	+
Eubacteries (Achaz *et coll.*, 2002)	39/40	28/38	22/42	+
Archaebactéries (Achaz *et coll.*, 2002)	11/11	5/8	2/8	+

TAB. 4.1 – **Généralisation du modèles aux autres organismes :** Le tableau indique le nombre d'espèces présentant une des quatres caractéristique du modèle relativement au nombre d'espèces étudiées pour ce critère. Les article ayant servis à la réalisation de ce tableau sont référencés en tête de chaque ligne. Les tests de corrélation correspondent à des corrélations de rang selon la méthode de $kendall - \tau$

corrélations qui différencie le plus les *Eubacteria* et les *Archae* des *Eucaryota* et ce sont ces deux caractéristiques qui sont les plus dépendantes des mécanismes de recombinaison homologue. Ces mécanismes très unifiés chez les eucaryotes sont beaucoup plus divers dans leur mécanismes moléculaires et dans leur efficacité dans les deux autres règnes. Cela renforce l'idée du rôle important de la recombinaison homologue, *via* la conversion génique dans l'établissement des corrélations. Il est à noter que les différences d'effectifs observées entre *Archae* et *Eubacteria* ne sont pas significatives[1], il n'est donc pas possible de dire que le modèle s'applique moins aux *Archae* qu'aux *Eubacteria*.

[1] $p_{value} = 0,67$ pour la corrélation % *d'identité* versus Δ et $p_{value} = 0,25$ pour la corrélation ℓ *versus* Δ selon un test de *Pearson* χ^2 réalisé par simulation

Chapitre 5

Comparaison des duplications entre espèces

Pourquoi une approche «multi-génome» ?

L'analyse des duplications en tant que marqueurs des événements de remaniement des chromosomes offre l'avantage de pouvoir être menée à partir d'un seul génome. Cet argument était particulièrement valable à la fin des années 1990, car peu de génomes étaient complètement séquencés. Aujourd'hui, les séquences de plusieurs centaines de génomes sont connues. Il est donc intéressant de tirer parti de cette nouvelle source d'information pour combler les lacunes des études réalisées en suivant l'approche «mono-génome». En effet avec celles-ci, il est impossible d'introduire dans le modèle certains paramètres. Parmi ceux-ci, le temps, pourtant fondamental dans une étude évolutive, était ignoré. Une nouvelle analyse, prenant en compte simultanément plusieurs génomes, permettra de l'introduire dans le modèle.

Les duplications intragéniques

Pour extraire une information temporelle de l'analyse des duplications, il faut être capable d'identifier un même événement de duplication dans un grand nombre d'espèces. Il faut, ensuite, comparer les caractéristiques des différentes occurrences de la répétition chacune ayant suivi une histoire évolutive distincte.

La principale difficulté de ce travail réside dans l'identification de duplications orthologues. L'orthologie s'entend ici dans son acceptation évolutive c'est-à-dire : provenant d'un même ancêtre commun par un phénomène de spéciation. Un groupe de «duplications orthologues» correspond donc à un ensemble des duplications observables à ce jour dans plusieurs espèces et provenant d'un même événement de duplication ancestral. Si la définition de l'orthologie est simple, son assertion est beaucoup plus complexe. Elle l'est d'autant plus lorsque l'on ne s'intéresse pas à une orthologie entre gènes, mais à une orthologie entre séquences d'ADN quelconques. Pour simplifier l'inférence de l'orthologie entre des duplications, une solution consiste à se restreindre à l'analyse d'une classe particulière de répétitions que nous identifierons sous l'acronyme *ICDR* pour «*Intragenic Close Direct Repeats*». Les *ICDR* correspondent à une sous-classe des *CDR* précédemment définies. Elles possèdent comme caractéristique supplémentaire la localisation des deux copies de la répétition à l'intérieur du même gène. Cette limitation simplifie l'inférence de l'orthologie des duplications en la réduisant au problème plus classique de l'inférence des orthologies de gènes. Ce projet de recherche nommé EvolRep est en cours de réalisation et seulement quelques résultats préliminaires sont disponibles. Ils permettent néanmoins de se faire une idée précise de nos chances d'aboutir.

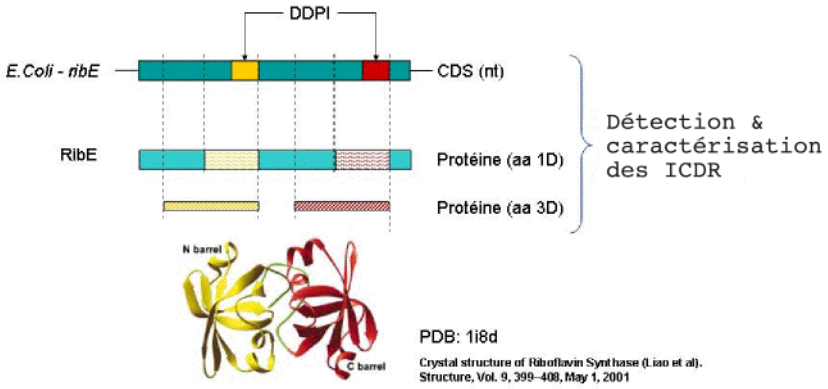


FIG. 5.1 – **Comparaison d'une duplication intragénique aux trois niveaux d'information étudiés** – Le gène *ribE* code pour la riboflavine synthase, il présente une *ICDR* observable aussi bien au niveau de la séquence nucléotidique que de la séquence en acides aminés. Pour cette enzyme dont la structure 3D est connue, la duplication se traduit par deux domaines structuraux similaires. Le report de la localisation des régions répétées sur la séquence nucléotidique (telles qu'elles ont été identifiées sur les trois niveaux) est matérialisé par les traits verticaux en pointillés. Un code couleur identifie chacune des deux copies de la duplication.

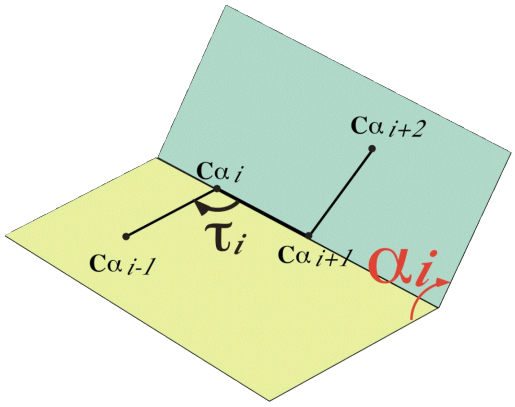
L'intégration des différents niveaux d'information

La restriction du problème aux duplications localisées à l'intérieur de gènes codant pour des protéines offre d'autres attraits. Une région codante peut-être considérée selon différents niveaux d'information. Le premier niveau correspond à la séquence nucléique, le deuxième à la séquence en acides aminés et le troisième à la structure tridimensionnelle de la protéine. Un quatrième niveau, le niveau fonctionnel peut aussi être envisagé. Ce dernier point sera abordé plus tard dans le chapitre consacré à la représentation des connaissances en biologie. Même si l'on se cantonne aux trois premiers niveaux cités, il est intéressant d'observer simultanément chacun d'eux afin d'analyser comment une duplication observable à un niveau se répercute sur les deux autres. Pour illustrer ce propos, la figure 5.1 présente une analyse préliminaire réalisée sur le gène *ribE* d'*E.coli*, ainsi que sur la séquence de la protéine *RibE* et sur sa structure tridimensionnelle. Une recherche des répétitions présentes dans ce gène a été réalisée au niveau de la séquence nucléique et de la séquence protéique grâce à une version modifiée de l'algorithme d'alignement local développé initialement par Pellegrini *et coll.* (1999) et présenté précédemment (voir page 41).

Les duplications au niveau de la structure tridimensionnelle peuvent être identifiées par des méthodes similaires à celles utilisées pour les identifier au niveau de la séquence du gène ou de la protéine. En effet, comme le montre la figure 5.2 la structure tridimensionnelle d'une protéine peut être décrite par une séquence définie sur un «alphabet angulaire» (Sagot *et coll.*, 1994). Toutes les méthodes classiques d'alignement peuvent ensuite être utilisées sur cette représentation atypique d'une structure tertiaire de protéine.

La mise en correspondances des résultats obtenus aux différents niveaux d'information étudiés montre clairement que la petite répétition, observable au niveau nucléique, est certainement issue d'un événement ancien couvrant une région beaucoup plus importante. Il va donc être possible, dans ce cas précis, de mettre en relation des séquences d'ADN orthologues ayant divergé depuis suffisamment longtemps pour que les algorithmes d'alignements classiques ne puissent plus les associer.

L'intégration de l'information «multi-génome»



$\ldots\alpha_{03}-\alpha_{21}-\alpha_{14}-\alpha_{21}-\ldots-\alpha_{17}-\alpha_{16}-\alpha_{03}-\alpha_{22}-\ldots$

FIG. 5.2 – **Les structures protéiques vues comme une séquence** – Il existe plusieurs façons de décrire la structure tridimensionnelle d'une protéine. La plus couramment utilisée consiste à donner l'ensemble des coordonnées cartésiennes des atomes de la protéine. Une autre possibilité est d'utiliser un système de coordonnées angulaires décrivant les positions relatives des atomes les un par rapport aux autres. En limitant la structure d'une protéine à son squelette de carbone α ($C\alpha$) deux systèmes de coordonnées angulaires sont utilisables : les angles ϕ, ψ d'une part et le système α, τ d'autre part. C'est ce dernier système de coordonnées internes qui est certainement le plus intéressant dans notre cas. L'angle α est l'angle formé entre deux plans passant par les carbones $(C_{\alpha,i-1}, C_{\alpha,i}, C_{\alpha,i+1})$ pour le premier et par les carbones $(C_{\alpha,i}, C_{\alpha,i+1}, C_{\alpha,i+2})$ pour le second. L'angle τ correspond à l'angle formé par trois carbones α consécutifs. Ce dernier angle a comme particularité de peu varier et d'être fortement corrélé avec l'angle α. De ce fait, il est possible de décrire le squelette carboné d'une protéine uniquement par une succession d'angle α. Ces angles α peuvent facilement être discrétisés pour former un «alphabet angulaire» représenté ici par les symboles α_{xx}. Deux symboles α_{xx} et α_{yy} correspondent à des angles d'autant plus similaires que xx est proche de yy. Dans notre exemple les facteurs $\alpha_{03} - \alpha_{21}$ et $\alpha_{03} - \alpha_{22}$ sont considérés comme répétés.

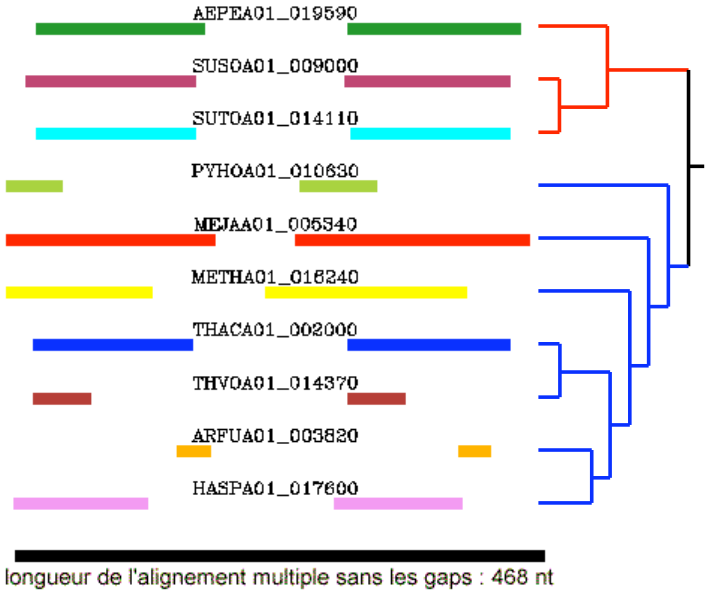


FIG. 5.3 – **Comparaison de la structure d'une répétition à l'intérieur de l'embranchement des archaebactéries** – Les gènes du facteur de transcription *TFIID* ont été identifiés dans 10 espèces d'*Archaea*. Une recherche des répétition au niveau de la séquence polypeptidique permet de mettre en évidence une *ICDR* observable dans toutes les espèces. La position des régions dupliquées est reportée après l'alignement multiple des séquences sur la séquence nucléotidique du gène (ligne noire en bas de la figure). La classification indiquée sur le côté est basée sur les ARN 16s et extraite du «*Bergey's Manual of Systematic Bacteriology*» (Garrity *et coll.*, 2001). Mnémoniques des espèces : AEPEA01, *Aeropyrum pernix K1* ; SUSOA01, *Sulfolobus solfataricus* ; SUTOA01, *Sulfolobus tokodaii* ; PYHOA01, *Pyrococcus horikoshii* ; MEJAA01, *Methanococcus jannaschii* ; METHA01, *Methanothermobacter thermautotrophicus* ; THACA01, *Thermoplasma acidophilum* ; THVOA01, *Thermoplasma volcanium* ; ARFUA01, *Archaeoglobus fulgidus* ; HASPA01, *Halobacterium sp. NRC-1*

Une seconde analyse préliminaire souligne l'intérêt de l'approche multi-génome dans l'analyse des duplications. Ce travail, réalisé dans le cadre du DEA d'Anna Jakubiec, a porté sur l'analyse des *ICDR* observables chez les archaebactéries. La famille des gènes codant pour le facteur de transcription *TFIID* est particulièrement intéressante. Comme nous le posions en hypothèse préalable à cette approche, la figure 5.3 montre que chaque version de la répétition présente dans le gène observé a suivi une histoire évolutive qui lui est propre, puisque les zones détectées comme répétées ne sont pas identiques d'une espèce à l'autre.

Dans ce cas, la duplication est observable dans toutes les espèces analysées. Elle est donc ancienne, puisque présente dès la racine des *Archae*. D'ailleurs, la similarité mesurée entre les deux copies est de 66% d'identité en acides nucléiques, ce qui est proche de la limite significative inférieure (*voir* FIG. 2.2, page 34). Une analyse plus fine du type de mutations accumulées dans chacune des deux copies de la duplication, lorsque l'on compare deux espèces proches, devrait permettre de dire si les deux copies évoluent de manière similaire et si par exemple elles ont une importance équivalente dans le rôle fonctionnel de la protéine. Une approche simple pour ce type d'analyse est de découper la séquence en différentes régions, suivant les limites des zones dupliquées et de comparer la pression de sélection s'appliquant sur chacune des zones (*voir* FIG. 5.4). Un premier regard sur les

résultats pourrait laisser croire qu'une pression de sélection plus importante s'exerce sur le segment 4 que sur le segment 2. Mais le faible nombre de sites utilisables pour mesurer les distances induit un écart type élevé. L'égalité des distances D_n (distance due aux mutations non synonymes) ou D_s (distance due aux mutations synonymes) entre ces deux régions (2 et 4) ne peut donc pas être rejetée ($P_{value} \approx 0,12$ pour les deux distances). Cette paire de gènes *TFIID* est néanmoins emblématique du problème méthodologique qui se pose. La comparaison des pressions de sélection exercées sur les différentes copies d'une répétition implique des calculs de distances évolutives à partie de courtes sous-séquences (un gène étant découpé en 5 régions). De nouvelles méthodologies de mesure du différentiel de pression de sélection entre les différentes zones d'une protéine devront donc être recherchées de manière à essayer de contourner la limite imposée par le faible nombre de sites.

Cette dernière partie du projet EvolRep met l'accent sur le quatrième niveau d'information : la fonction portée par la protéine. L'impact d'une duplication d'une séquence d'ADN sur la fonction d'une protéine est l'aboutissement final des impacts observés sur la structure primaire et tertiaire. Depuis la première étude réalisée au niveau des familles multigéniques de la levure *Saccharomyces cerevisiae*, j'ai volontairement ignoré les implications fonctionnelles des duplications. Ces approches développées dans le cadre du projet EvolRep sont sans doute une première étape dans la prise en compte de cet aspect des choses.

Les premiers résultats

Des premiers résultats quantitatifs ont été obtenus pour le projet EvolRep. Ils concernent la phase d'identification des *ICDR* au niveau des séquences des gènes et des séquences primaires des protéines qu'ils codent.

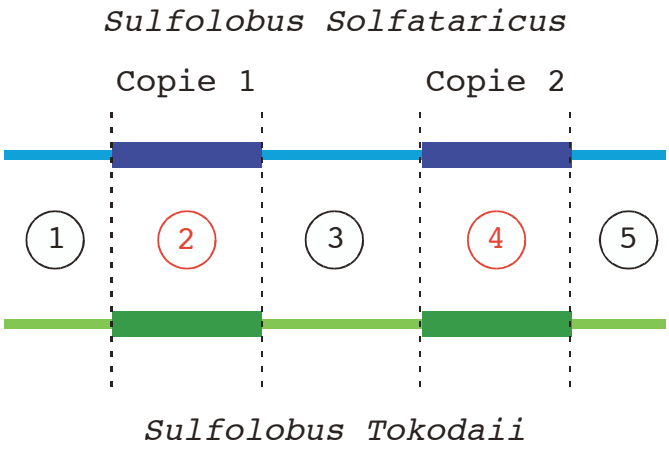
Pour chaque séquence nucléique et chaque séquence protéique, les répétitions internes ont été identifiées en alignant la séquence avec elle-même selon l'algorithme décrit précédemment (voir page 41). Un score centré réduit (Z_{score}) est calculé pour chacune des répétitions identifiées. Ce score normalisé est utilisé pour déterminer si la répétition est significative. Le Z_{score} limite a été fixé à $6,1$ en se basant sur une distribution empirique de ce score. D'après cette distribution cela correspond à un risque de 1/5000 (soit moins d'un faux positif par génome bactérien).

Une premier groupe de 75 espèces (59 *Eubacteria* et 16 *Archae*) soit 257470 gènes ont été analysés. 4% de ces gènes possèdent une *ICDR* détectable au niveau de leur séquence nucléique et 6% au niveau de leur séquence protéique (*voir* FIG. 5.5).

La plus grande sensibilité de détection des répétitions au niveau de la séquence protéique, relativement à la détection au niveau de la séquence nucléique, permet d'identifier normalement des répétitions plus anciennes. Il est donc normal d'identifier des *ICDR* au niveau protéique qui ne sont plus détectables au niveau nucléique. Il est par contre surprenant que plus d'un tiers des *ICDR* détectées au niveau nucléique n'induisent aucune répétition au niveau protéique. Bien que rien n'oblige une duplication à posséder ces deux copies dans le même cadre de lecture à l'intérieur d'un gène, l'hypothèse d'indépendance complète des phases de lecture des deux copies est difficile à imaginer car elle suppose que la même séquence soit lisible dans au moins deux cadres différentes. Une faible proportion de duplications hors phase était donc attendue.

Analyse entre espèces proches

Le projet EvolRep se limite aux *ICDR* afin de simplifier l'inférence de l'orthologie. Si cette limitation offre plusieurs avantages, il peut être intéressant de la dépasser de manière à disposer d'une vision plus générale du monde des duplications. Une solution maintenant envisageable est d'observer les duplications présentes dans des groupes de génomes très proches. La multiplication des projets de séquençage permet aujourd'hui de disposer de plusieurs séquences de génomes phylogénétiquement très proches (plusieurs souches d'une



	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5
% d'identité	78%	76%	78%	81%	80%
D_s	$0.70\ ^{0,351}_{13,2}$	$0,64\ ^{0,104}_{52,4}$	$0,55\ ^{0,080}_{83,6}$	$0,48\ ^{0,095}_{74,6}$	$0.36\ ^{0,120}_{24,7}$
D_n	$0\ ^{--}_{13,8}$	$0,07\ ^{0,038}_{55,6}$	$0,04\ ^{0,026}_{81,3}$	$0,02\ ^{0,020}_{75,4}$	$0,12\ ^{0,066}_{26,3}$
D_n/D_s	–	0,10	0,08	0,04	0,34

FIG. 5.4 – **Analyse comparative des différentes parties du gène *TFIID*** – Un gène possédant une *ICDR* peut être découpé en cinq segments d'après les limites des deux copies de la répétition. La partie haute de la figure schématise ce découpage pour le gène *TFIID* de deux *Archae* du genre *Sulfolobus*. Le tableau indique, pour chacune des régions ainsi délimitée, le pourcentage d'identité en acides nucléiques et les distances évolutives dues aux substitutions synonymes et non synonymes calculées selon un principe analogue à celui décrit par Nei et Gojobori (1986). La différence entre cette méthode et celle utilisée dans le tableau vient de la définition d'une substitution synonyme. Ici une substitution est comptée comme synonyme même si elle provoque une changement d'acides aminés, à condition que la paire d'acide aminé substituée possède un score positif dans une matrice de substitutions type PAM (Dayhoff *et coll.*, 1978) ou BLOSUM (Henikoff et Henikoff, 1992). La matrice utilisée ici est BLOSUM80. À chaque distance sont associées deux valeurs : en exposant l'écart type et en indice le nombre de sites utilisables pour calculer la distance.

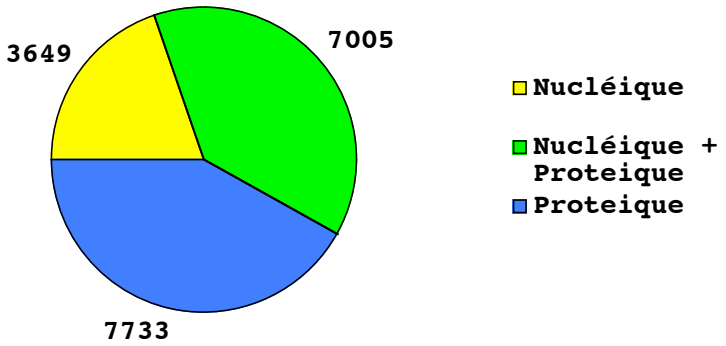


FIG. 5.5 – **Détections des *ICDR* résultats préliminaires** – Nombre de gènes porteur d'une *ICDR* détectables au moins à l'un des deux niveaux informationnels étudiés (nucléique ou protéique). L'analyse porte sur 75 espèces (59 *Eubacteria* et 16 *Archae*) soit 257470 gènes. Il a été identifié 18387 gènes porteurs d'une *ICDR*. 10654 *ICDR* étaient visibles au niveau de la séquence nucléique et 14738 au niveau de la séquence protéique. Il est à remarquer que 34% des *ICDR* detectables au niveau de la séquence nucléique n'induisent pas de répétition détectable au niveau de la séquence protéique.

même espèce bactérienne). Cette proximité se traduit par des génomes fortement colinéaires (synténiques). Cette propriété peut être utilisée avantageusement pour inférer l'orthologie entre les duplications. Une fois réalisée, l'identification des paires de gènes orthologues entre les deux génomes proches analysés : si deux paires de gènes orthologues sont contigus et orientés de la même façon dans les deux génomes, alors la zone intergénique les séparant peut être considérée comme orthologue. Cette possibilité d'étendre la notion d'orthologie hors des zones codantes permet l'étude de l'évolution comparée des répétitions entre plusieurs espèces de manière analogue à celle réalisée dans le cadre du projet EvolRep en assouplissant la limitation aux duplications intragéniques.

Troisième partie

L'intégration des connaissances en biologie

Chapitre 6

La gestion des connaissances biologiques

L'étude simultanée d'un événement de duplication dans de nombreuses espèces, telle qu'elle est présentée dans le chapitre précédent, devrait permettre :

- d'intégrer la notion de temps dans le modèle de dynamique des répétitions de manière plus explicite et
- d'analyser les imbrications des différents niveaux d'information : séquence nucléique, séquence protéique et structure tridimensionnelle.

Comme je l'ai aussi évoqué plus tôt, il est possible de prendre en compte un quatrième niveau d'information : la fonction. Ce niveau supplémentaire ne peut pas être abordé comme les trois premiers par des méthodes relevant de l'analyse de séquences. Pour être capable d'associer un éventuel effet des duplications et plus généralement des remaniements chromosomiques sur une fonction cellulaire, il faut pouvoir mettre en relation des informations de type annotation fonctionnelle avec les régions chromosomiques porteuses des répétitions étudiées. C'est dans ce but que je travaille depuis trois ans maintenant à la mise en place d'une base de connaissances intégrant les informations relatives, bien évidemment, à l'organisation des génomes microbiens mais aussi des informations fonctionnelles liées aux protéomes et au réseau métabolique.

La nécessité de gérer les connaissances associées à des données de séquences est concomitante à l'avènement des techniques de séquençage. La première version de l'«*Atlas of Protein Sequence and Structure*» par Dayhoff *et coll.* (1965) suit d'un an la présentation de la technique de séquençage automatique de protéines par Edman et Begg en 1964 (Edman et Begg, 1967). Il en a été de même pour le séquençage des acides nucléiques (Maxam et Gilbert, 1977; Sanger *et coll.*, 1977) et le développement des deux grandes banques de séquences : Genbank (Dayhoff *et coll.*, 1981) et EMBL (Kneale et Kennard, 1984). Ainsi, dès les années 1970, des banques de séquences se sont organisées pour gérer l'ensemble des séquences obtenues par les différents laboratoires de recherche. Le terme «banque» exprime le fait que les séquences sont déposées à la demande et sous la responsabilité des chercheurs les ayant produites. En 1986, Amos Bairoch a développé ce que nous pouvons considérer comme la première «base» de données en biologie : Swiss-Prot (Bairoch et Boeckmann, 1991). Contrairement aux principes suivis par Genbank et l'EMBL, les données sont incorporées dans *Swiss-Prot* sous le contrôle d'une équipe d'annotateurs qui corrige et complète manuellement chacune des entrées. Cette différence fondamentale a pour principal avantage d'assurer la qualité et la cohérence des données représentées.

La multiplication des données

La multiplication des données biologiques est de deux types : d'abord quantitative puis plus récemment qualitative. Pour illustrer l'aspect quantitatif, la banque EMBL contenait, en 1982, 508 séquences. Elle en comptait plus de 54 millions en juin 2005. Cette évolution due à l'amélioration des techniques de séquençage et, par là même, au développement des projets de séquençage de génomes complets, est à mettre en parallèle avec une diversification des types de données manipulées par les biologistes. Cet aspect qualitatif a conduit au développement d'un nombre toujours croissant de bases de données. Une première catégorie de ces nouvelles bases résulte d'un travail d'analyse sur les données des grandes banques généralistes précédemment citées. Ces travaux ont conduit notamment au développement de bases de données de motifs protéiques comme ProDom (Corpet *et coll.*, 1998) ou à la création de bases de données spécialisées par organisme comme Hovergen (Duret *et coll.*, 1994) pour les vertébrés ou TAIR (Rhee *et coll.*, 2003) pour *Arabidopsis thialiana*. Une deuxième catégorie se rapporte aux données expérimentales telles que la banque PDB[1] pour les données cristallographiques de macromolécules, la base DIP[2] pour des données d'interactions protéine-protéine ou encore la base Malaria IDC[3] présentant des données issues des nouvelles technologies liées à l'analyse des génomes (transcriptome, protéome). Enfin le dernier groupe tend à représenter des concepts biologiques plus théoriques, comme le métabolisme dans KEGG ligand (Kanehisa, 2002) ou la classification des enzymes dans la base de données Enzyme (Bairoch, 2000).

Les relations entre les bases

Si la plus grande partie des connaissances biologiques se retrouvent aujourd'hui dispersées dans la littérature sous forme de texte écrit en langage naturel, une part croissante de celles-ci est décrite dans de nombreuses bases de données. Il est rapidement apparu nécessaire aux auteurs de ces différentes bases et indispensable aux biologistes les utilisant de mettre en relations les données contenues dans chacune d'elles. Cette volonté est devenue réellement cruciale avec l'avènement de la génomique. La mise en relation de plusieurs bases de données pose certes des problèmes informatiques, mais encore une fois, les principaux problèmes sont d'ordre biologique car liés à la sémantique des informations contenues.

Le cas le plus simple à prendre en compte est la mise en relations de plusieurs banques gérant le même type de données. À ce titre, le cas des trois grandes banques de données dédiées aux séquences nucléiques, Genbank, EMBL et DDBJ (Uchida, 1986), est exemplaire. Alors, qu'elles furent créées de manière indépendante, la volonté d'enrichir chacune d'elles des données contenues dans les deux autres a incité leurs auteurs à mettre en place une procédure d'échange. Ainsi, aujourd'hui, elles partagent sensiblement les mêmes données et constituent de ce fait trois points d'entrée d'une seule et même banque mondiale. Le prérequis à ce travail a été de définir un format d'échange de données permettant un transfert simple des informations. Dans ce cas, le seul problème à résoudre fût de décrire un protocole d'échange efficace permettant la propagation d'une mise à jour d'une base à l'autre.

Malheureusement, les mises en relation les plus intéressantes sont sans doute celles qui cherchent à rapprocher des bases gérant des données de natures différentes. Par exemple, pour conserver l'origine des informations résultant du travail d'expertise effectué par le groupe Swiss-Prot, il est nécessaire de lier des références bibliographiques aux séquences protéiques et aux annotations décrites dans la base. Les données bibliographiques sont gérées par ailleurs dans la base de données Medline. Plusieurs solutions sont envisageables pour associer ces deux types de données. Le choix du groupe Swiss-Prot consiste à recopier, dans chaque entrée de la base, les références bibliographiques fournies par Medline. Ce choix permet un accès simple à l'information référencée et procure une certaine indépendance de Swiss-Prot vis-à-vis de Medline. Cependant, cette option complique les opérations de propa-

[1] http ://www.rcsb.org/pdb
[2] http ://dip.doe-mbi.ucla.edu/
[3] http ://malaria.ucsf.edu/

gation des mises à jour de Medline vers Swiss-Prot. De plus, si plusieurs entrées Swiss-Prot référencent le même article, il sera nécessaire de corriger chacune d'elles afin de maintenir la cohérence de la base.

Hormis ces problèmes assez facilement résolus par des techniques informatiques classiques, la mise en relation d'objets différents décrits dans différentes bases de données implique que l'on mette en relation les concepts décrits dans chacune d'elles. Elle repose donc sur la sémantique associée à chaque élément décrit. Il ne suffit pas de considérer que deux éléments sont nommés de manière identique pour conclure qu'ils recouvrent la même signification. Prenons l'exemple des séquences codant pour des protéines dans le génome, généralement nommées CDS (pour coding sequence). Suivant les bases de données, le codon stop terminant une région codante est, ou n'est pas, inclus dans la CDS. Il faut donc associer à chaque concept décrit dans une base de donnée une définition explicite qui permet de prendre en compte les différences sémantiques lors du processus de mise en correspondance.

Dans d'autres bases, un CDS sera décrit sous le terme : *gène*. Un tel glissement sémantique peut, de proche en proche, conduire à des associations aberrantes lorsque l'on prend en considération l'ensemble des définitions d'un gène, qui va de la vision mendélienne à son support moléculaire. Pour mettre en relation des concepts de nature différente, il ne suffit pas de définir un format de données, comme pour résoudre l'échange entre les banques Genbank, EMBL et DDBJ. Il est nécessaire de décrire un modèle de données définissant chacun des concepts représentés et leurs relations. On parle alors d'ontologies.

Format de données versus modèle de données

Le contenu des grandes bases de données comme l'EMBL ou Swiss-Prot est décrit par un format de données et un manuel d'utilisation. Le format peut être assimilé à une grammaire permettant à un programme informatique la lecture et l'écriture des entrées de la base. Le manuel, destiné aux utilisateurs, décrit l'ensemble des informations présentes dans la base. Cette description dans un langage naturel a deux limites. La principale est qu'elle est souvent ambiguë, le langage naturel n'obligeant pas l'auteur à expliciter de manière univoque les concepts manipulés. La seconde, très liée à l'ambiguïté décrite précédemment, est qu'il est difficile de s'en servir comme référence pour l'élaboration de programmes visant à exploiter les informations contenues dans la base. La définition d'un modèle de données explicite vise donc à pallier ces limites.

Si les modèles de données ont une aussi grande importance, il peut être légitime de se demander pourquoi ils n'ont pas été pris en compte lors de la conception de ces bases. La meilleure explication est, à mon avis, historique. Lors de leur création, les grandes banques de données avaient peu à interagir avec des sources de données externes. Aussi, elles ont été conçues comme des collections d'entrées stockées dans de simples fichiers textes qui devaient avoir comme première qualité d'être lisibles par l'homme. À aucun moment elles n'ont été conçues dans l'idée de mettre en relation les informations contenues dans plusieurs centaines voir plusieurs milliers d'entrées, provenant éventuellement de bases différentes.

Une réflexion est maintenant engagée pour décrire des modèles permettant de restructurer l'information au niveau des organismes gérant les grandes banques de données (EMBL, *Swiss-Prot*). La définition d'un modèle de données, puis la restructuration des données suivant ce modèle impliquent d'extraire des anciennes entrées de nombreuses connaissances exprimées en langage naturel. Cette phase d'extraction des connaissances rend quasiment impossible une transcription automatique. Malheureusement, le volume actuel de ces bases rend leur transcription manuelle laborieuse. Concernant les bases de données généralistes, la difficulté principale consiste dans la définition d'un modèle prenant en compte la diversité des utilisateurs. En effet, toute modélisation est conditionnée par l'utilisation que l'on souhaite faire du modèle. Ainsi pour un généticien, un biochimiste ou une personne s'intéressant à la phylogénie des espèces, le modèle idéal de description d'une protéine ne sera sans doute pas identique.

La prise en compte simultanée de nombreux points de vue induit rapidement une complexité du modèle le rendant inutilisable. Il est impossible de concevoir un modèle universel réconciliant tous les points de vue. C'est donc par l'intermédiaire des bases de données spécialisées qu'il faut espérer structurer finement l'information. Afin d'éviter les écueils décrits précédemment, les personnes produisant de nouveaux types de données essayent de mettre en place, dès maintenant, les modèles d'échange consensuels qui seront utilisés par les nouvelles bases. Sans aller jusqu'à une réelle modélisation des connaissances, ces «super-formats» s'appuient sur une description explicite des informations stockées. Ils devraient donc au moins permettre une meilleure exploitation automatique des données. Par contre, leur caractère consensuel rend leur structure complexe et donc souvent compliquée à utiliser. Ce type de travail est en cours notamment pour les données de transcriptome avec le projet MIAME pour «*Minimum Information About a Microarray Experiment*» (Brazma *et coll.*, 2001) et pour les données de protéome avec le projet PSI acronyme de «*the Proteomics Standards Initiative*» (Taylor *et coll.*, 2003).

Chapitre 7

MicrOBI : une approche pragmatique de l'intégration de données en biologie

Si l'établissement d'un modèle de données est d'une complexité quasi rédhibitoire pour les grandes banques généralistes, il doit être un préalable impératif à la mise en place de bases de données spécialisées. Ces modèles correctement définis permettent d'envisager avec sérénité l'intégration de données provenant de nombreuses sources hétérogènes. Deux grandes familles d'approches peuvent être envisagées pour associer différentes bases de données.

L'approche fédérative définit un ensemble d'interfaces permettant de spécifier les liens unissant les concepts décrits dans chacune des bases. Cette vision des choses offre comme avantage principal que chacune des sources de données reste indépendante et donc libre de faire évoluer son contenu. On peut donc ainsi espérer que chaque élément de la base fédérative soit maintenu le plus à jour possible par les meilleurs spécialistes des sous-domaines mis en relation et que l'ensemble bénéficie donc de la somme des compétences de chacun.

L'autre approche, dite intégrative, vise à construire un entrepôt de données (*data warehouse*) possédant son propre modèle et s'alimentant des informations contenues dans plusieurs autres bases. Cette approche implique une recopie de l'information, donc une relative indépendance de l'entrepôt vis-à-vis de ces sources. Elle permet aussi, par une redéfinition du modèle de données, d'apporter un point de vue peut-être plus pertinent aux connaissances ainsi associées. Son principal inconvénient est sans doute la duplication physique de l'information et de ce fait la nécessité de mettre en place une politique de synchronisation des mises à jour de données réalisées au niveau de chacune des sources utilisées.

Les objectifs de *MicrOBI*

Mettre en relation les données taxonomiques, génomiques et fonctionnelles des micro-organismes possédant leur génome complètement séquencé est nécessaire à la poursuite de l'étude que je mène sur la fluidité des génomes. Plusieurs bases de données doivent être mises en relation pour pouvoir poser des requêtes intégrant ces différents types d'informations. *MicrOBI* a été développée afin de réaliser cette intégration. Les problèmes d'une telle intégration sont divers. Ils tiennent à la fois : de la volumétrie des données, de leur diversité en types et en origines et de leur mise à jour asynchrone. Ce dernier point est particulièrement problématique pour le maintien des liens existants. Enfin *MicrOBI* a été conçue de façon à pouvoir facilement y ajouter de nouveaux type de données publiques ou privées.

C'est une approche intégrative qui a été choisie. *MicrOBI* a été développée dans un système de bases de données relationnelles. Le premier objectif est que la base doit garantir

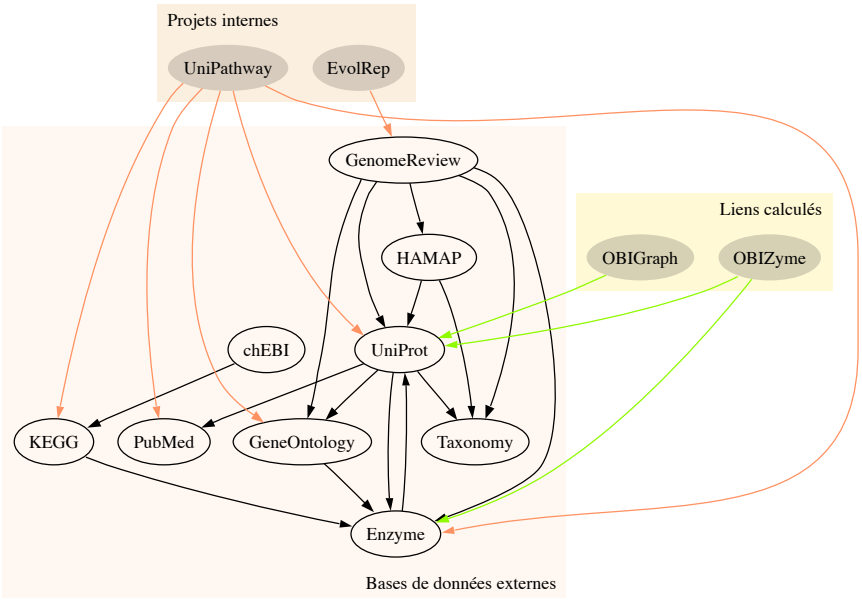


FIG. 7.1 – **Schéma global de la base de données MicrOBI** – La base de données *MicrOBI* contient à la fois des données publiques, des données calculées et des données rattachées aux projets de recherche utilisant cette base. Sur ce schéma simplifié, chaque source de données est identifiée par un ovale d'où partent des flèches pointant les autres sources référées par celle-ci. La cohérence des liens ainsi marqués est assurée par la base.

la cohérence des données. Pour ce faire, elle doit être capable de propager automatiquement les corrections de l'information réalisées par la mise à jour d'une des sources de données aux autres données liées.

Aujourd'hui la base de données *MicrOBI* fonctionne et est utilisée non seulement pour l'analyse des duplications dans un contexte fonctionnel, mais aussi dans le cadre d'une collaboration avec le SIB (*Swiss institut of bioinformatics*) à Genève pour la réannotation des données métaboliques de la base de données *Swiss-Prot*.

Une vue générale du modèle

Les données présentes dans *MicrOBI* sont de deux types : des données publiques reformatées selon nos besoins, des données privées que nous souhaitons croiser avec les données publiques précédemment citées (*voir* FIG. 7.1).

Les données publiques

Les données publiques intégrées dans *MicrOBI* couvrent différents domaines de connaissance. Pour chacun de ces domaines, une ou plusieurs banques de données considérées comme références sont utilisées.

Pour la taxonomie, c'est la banque du *NCBI* qui est utilisée (Benson *et coll.*, 2000; Wheeler *et coll.*, 2000). Elle décrivait plus de 264000 taxons au mois de juin 2005 et est utilisée comme base de données taxonomiques de référence par de nombreuses autres bases de données biologiques.

Les données de génomique (séquences des chromosomes et annotations) proviennent du projet *Genome Review* de l'EBI. *Genome Review* est une des parties du projet d'intégration de données de l'EBI «*Integr8*» (Kersey *et coll.*, 2005). L'avantage que nous avons à utiliser cette source de données est que les annotations des génomes sont corrigées par l'EBI en fonction des informations présentes dans la base de données *Swiss-Prot* et le projet *HAMAP* qui sont aussi utilisés par MicrOBI. Ce lien fort est important pour *MicrOBI* car cela augmente le nombre de liens existant entre les données et simplifie leur gestion puisque les mises à jours de *Genome Review* sont réalisées conjointement à celles de *Swiss-Prot*.

Les données de protéomiques sont issues de la base de données *Uniprot-Swiss-Prot* (Boeckmann *et coll.*, 2003) et du projet d'annotation des protéomes complets *HAMAP* (Gattiker *et coll.*, 2003). Du fait du projet de réannotation des données métaboliques des entrées de *Swiss-Prot*, ces données jouent un rôle important dans *MicrOBI*.

Les données métaboliques proviennent principalement de KEGG (Kanehisa, 2002). L'intérêt de cette base de données outre sa qualité est son aspect multi-génomes. Ce parti pris est important dans le cadre de *MicrOBI* qui cherche à mettre en relation des informations relatives à plusieurs organismes. Les données de KEGG sont complétées par chEBI, la base de données des métabolites de l'EBI (Brooksbank *et coll.*, 2005).

La dernière classe de données publiques présentes dans *MicrOBI* correspond à des classifications fonctionnelles. Deux classifications sont gérées actuellement : GeneOntology (Ashburner *et coll.*, 2000) et la classification enzymatique telle qu'elle est fournie par le SIB-Genève (Bairoch, 2000).

Les données privées

Les données privées de *MicrOBI* se divisent en deux catégories : d'une part, les données liées à des projets de recherche comme EvolRep (voir page 57, Comparaison des duplications entre espèces) ou UniPathway le projet de réannotation des données métaboliques de *Swiss-Prot*, d'autres part des données calculées.

Les données publiques associées au sein de *MicrOBI* se référencent entre elles par l'intermédiaire de liens croisés décrits dans leurs entrées (annotation du type «*dbxref*»). D'autres liens sont ajoutés par la base de données *MicrOBI* en s'appuyant sur des calculs. Deux sous-parties de *MicrOBI* correspondent à ce type de données : l'OBIGraph et OBIZyme.

L'OBIGraph partitionne *Swiss-Prot* en groupes de protéines très similaires. À la base, l'OBIGraph est un graphe où les noeuds sont les protéines complètes, longues de plus de 60 acides aminés, de *Swiss-Prot*/*HAMAP* et les arrêtes décrivent une relation de similarité entre ces protéines. La relation de similarité est calculée par le programme BLAST avec comme contrainte que le hit BLAST trouvé doit couvrir au moins la totalité de la plus grande séquence moins 30 acides aminés. Sur ce graphe, un groupe de protéines est définie comme une composante connexe. L'avantage de cette définition est que si le premier calcul de l'OBIGraph a nécessité une semaine de calcul, sa mise à jour à chaque version de *Swiss-Prot* est réalisée en quelques heures. Cette rapidité nous assure une cohérence permanente entre nos classes de protéines et les données de *Swiss-Prot* présentes dans *MicrOBI*.

OBIZyme est une collection de profils PSI-Blast (Altschul *et coll.*, 1997) permettant d'affecter des liens entre la classification enzymatique et les séquences contenues dans *Swiss-Prot*. Construit sur un principe similaire à PRIAM (Claudel-Renard *et coll.*, 2003), la gestion en interne de ces profils nous permet de les maintenir à jour à chaque version de *Swiss-Prot*.

Données particulières – Types particuliers

La mise en place d'une base de données contenant des informations biologiques amène à représenter des informations non traditionnellement représentées dans une base de données relationnelle. Pour représenter ces types de données complexes atypiques, deux solutions sont envisageables : les représenter à partir des types de données standards comme les chaînes

Type	Données représentées
Gestion des séquence	
bioseq	permet de stocker efficacement à l'intérieur d'un *blob* «*binary large object*» de grandes séquences nucléiques comme des chromosomes complets
bioseqid	gestion des identifiants internes des grandes séquences nucléiques (bioseq)
gloc	décrit une région d'intérêt sur une *bioseq* par une série d'intervalles $[debut, fin]$. Ce type permet donc la gestion de la localisation des éléments génétiques sur un chromosome.
Gestion de la machinerie de traduction	
geneticcode	Permet la représentation de toutes les informations liées à un code génétique (codons classiques, codons start...) L'intérêt d'un tel type est de faciliter l'implémentation de fonctions de traduction *in silico*
trnafold	Permet de représenter la structure secondaire d'un ARN_t. Cette structure est utile pour déterminer l'anticodon, donc l'acide aminé correspondant à l'ARN_t
mcov	Permet de représenter un modèle de covariance tel qu'il est défini par le programme tRNAScan (Lowe et Eddy, 1997)
Gestion de la similarité	
seqidx	Représente une table de hash des mots d'une taille donnée d'une séquence. Ce type permet de réaliser des comparaisons simples de séquences par des algorithmes de type FASTP (Lipman et Pearson, 1985)

TAB. 7.1 – **Les OBITypes :** Les OBITypes permettent de représenter de manière simple, différentes informations biologiques par le biais de types spécialisés. Ils sont développés en langage C et étendent le jeu des types standards de PostgreSQL.

de caractères, les entiers ou les nombres réels, ou enrichir le gestionnaire de base de données d'un jeu de nouveaux types dédiés aux éléments biologiques.

Les OBITypes répondent à ce dernier objectif. Le développement de types spécialisés permet d'une part de simplifier le schéma des bases de données et d'autre part, de spécifier des contraintes et des méthodes spécifiques à ces objets. Actuellement, les OBITypes ajoutent à PostgreSQL des types permettant de représenter des grandes séquences nucléiques, des données liées aux mécanismes de traduction et des informations permettant de facilité l'inférence de la similarité entre séquences (voir Tab. 7.1). A chacun de ces types est associé un ensemble de fonctions permettant de réaliser des «opérations biologiques» sur les objets représentés. Ces fonctions permettent d'intégrer à une requête SQL des opérations de calcul relevant de l'analyse de séquences.

Si l'on prend le cas du type *gloc* permettant de décrire une région sur une grande séquence, il est possible de l'utiliser pour décrire les positions des régions codantes (CDS) d'un chromosome. Il aurait été certe possible de représenter cette même information de position par deux entiers l'un indiquant le début de la CDS, l'autre la fin. Éventuellement, un attribut booléen supplémentaire aurait pû être utilisé pour indiquer sur quel brin de l'ADN se trouve le CDS. L'utilisation du type *gloc* simplifie le modèle en réduisant ces trois attributs à un seul. Elle garantit, de par les propriétés du type, que les bornes indiquées pour le fragment sont cohérentes avec la séquence du chromosome. Enfin grâce a des fonctions spécifiques utilisant ce type en paramètre, elle rend par exemple possible une requête SQL comme celle présentée dans la figure 7.2.

Si notre objectif est de simplifier la gestion des données biologiques et les requêtes sur ces données, l'effort de développement des OBITypes me semble important. Il se poursuit actuellement par l'ajout de nouveaux types et de nouvelles fonctionnalités associées à ces types. La prochaine étape vise à mieux représenter les données du métabolisme par des types adaptés à la description des structures des métabolites. Ils permettront notamment d'intégrer aux requêtes SQL des critères structuraux.

Gestion de la cohérence

Le dernier problème, cité en introduction de ce chapitre, correspond au maintien de références cohérentes entre des données d'origines diverses mises à jour à des fréquences différentes. La taxonomie est gérée par le NCBI et une nouvelle version est disponible chaque jour. Les données provenant de *Swiss-Prot* sont mises à jour toutes les deux semaines. KEGG évolue un peu plus lentement encore. Chacune de ces banques référence les autres par l'intermédiaire de liens croisés décrits dans leurs entrées. Ces liens sont repris pour mettre en relation les différentes tables de *MicrOBI*.

Un des objectifs de *MicrOBI* est d'assurer automatiquement la cohérence de ces liens malgré la gestion indépendante des sources de données. Le faire automatiquement est important car *MicrOBI* peut potentiellement être consultée et mise à jour par plusieurs programmes. Il n'est donc pas raisonnable de faire reposer la cohérence sur ceux-ci. Il a donc été défini, pour chaque type de liens, une stratégie de maintien de la cohérence qui est déclenchée automatiquement par la base lors de l'insertion de nouvelles données ou lors de leur correction. La table 7.2 montre l'action de ces mécanismes tels qu'ils sont transcrits dans les journaux de la base de données.

a) table CDS

Variable	Type	Description
mnemo	text	Mnémonique du CDS
location	gloc	Position du CDS
repliconid	text	identifiant du réplicon

b) Requête SQL

```
select bioseq_fasta(mnemo,
     gloc_getseq(
       gloc_5primeloc(location,-20,15)
    ) ) as fasta
     from cds
   where repliconid='escolAa01';
```

c) fonctions OBITypes

Fonction	Paramètres	Valeur retournée
bioseq_fasta	(label, text sequence)	une séquence au format Fasta
gloc_getseq	(gloc)	texte de la séquence
gloc_5primeloc	(gloc,position,length)	gloc

d) Résultat de la requête

```
                    fasta
---------------------------------------
>CDS_escolAa01_043170
tatgtgaaagaggaa
>CDS_escolAa01_043160
accaggtaaggtaaa
...
```

FIG. 7.2 – **Exemple de requête utilisant les OBITypes** – Si la table CDS (a) décrit l'ensemble des CDS présentes sur différents chromosomes bactériens, l'extraction des sous-séquences de 15 nucléotides de long localisées 20 nucléotides en amont de chacun des CDS d' *E.coli* peut être exprimée par la requête (b). Cette requête utilise trois fonctions apportées par le module OBITypes (c) et retourne un ensemble de séquences au format FASTA (d).

Message d'avertissement : les données sont corrigées automatiquement		
table	entrée	message
uniprot.enzyme_link	Q8ZP07	EC Link 4.2.1.13 is transferred to 4.3.1.17
uniprot.enzyme_link	Q8ZP27	EC Link 1.18.99.1 is transferred to 1.12.7.2
uniprot.enzyme_link	Q8ZS18	EC Link 4.2.99.2 is transferred to 4.2.3.1
kegg.reaction_enzyme	R01116	EC Link 1.14.13.45 is transferred to 1.14.18.2
geneontology.enzyme_link	0047767	EC Link 1.2.3.10 is transferred to 1.2.2.4
geneontology.enzyme_link	0004198	EC Link 3.4.22.17 is transferred to 3.4.22.52
geneontology.enzyme_link	0047767	EC Link 1.2.3.10 is transferred to 1.2.2.4
uniprot.go_link	Q7MMZ0	GO link 0008698 is transferred to 0050515

Message d'erreur : les liens sont perdus		
table	entrée	message
uniprot.enzyme_link	Q8ZHB1	EC Link 1.13.12.10 is a deleted entry from enzyme database (enzyme.deleted)
Uniprot	Q7MAM3	Uniprot keyword Flagella is unknown
kegg.pathway_component	map00710	R00762 / C05378 link not found in kegg.compound_reaction table

TAB. 7.2 – **Gestion automatique des incohérences :** La mise à jour asynchrone des différentes sources de données conduit à introduire des incohérences dans les données de *MicrOBI*. Un ensemble de règles associées aux événements de mise à jour des tables de *MicrOBI* assure lorsque cela est possible la correction des données (tableau du haut) obsolètes. Certaines incohérences fatales (tableau du bas) ne peuvent être corrigées et provoquent la perte de l'information. L'ensemble de ces corrections automatiques est répertorié dans des journaux liés à la base permettant un suivi des modifications.

Quatrième partie

Conclusion et perspectives

La biologie à l'échelle des génomes complets permet depuis les années 1990 l'émergence d'une nouvelle période dans l'histoire de cette science. Après une ère de la biologie moléculaire disséquant un à un les mécanismes moléculaires mis en jeu dans les cellules, les nouvelles techniques de la génomique en offrant une vision globale, force la réintégration des connaissances. Elles permettent aussi d'aborder à nouveau des questions anciennes, mais cette fois avec une masse de données nettement plus importante.

Parmi ces questions, l'organisation des génomes et leur évolution m'intéressent particulièrement. Si j'aborde ce sujet sous l'aspect humoristique de la revanche de l'ADN sur les protéines, je trouve plus sérieusement un intérêt dans l'impact croisé des contraintes présentes à chacun des niveaux d'informations (ADN, protéines, fonctions). Si un événement de duplication peut-être à l'origine d'une nouvelle fonction, il a avant tout une chance non négligeable d'en perturber une existante. Une fois l'évènement fixé, la présence des deux séquences met en jeu des mécanismes propre à l'ADN comme la recombinaison et la conversion génique qui vont perturber régulièrement la séquence des gènes où sont localisées ces duplications. Quels sont les degrés de liberté réellement disponibles pour permettre une conservation de la structure protéique et donc de la fonction ?

Pour comprendre les mécanismes sous-jacents aux remaniements des génomes, j'ai entrepris une analyse des duplications et plus généralement des répétitions présentes le long des chromosomes. Cette étude, basée principalement sur des technique d'analyse de séquence, m'a permis de proposer un modèle pour la genèse et l'évolution des duplications dans les génomes. Le caractère relativement universel de ce modèle montre, selon moi, qu'il repose sur des mécanismes partagés par tous les règnes du monde vivant donc certainement liés à des fonctions intrinsèquement rattachées à la gestion de l'ADN par les cellules. Si ce modèle doit encore être affiné pour mieux comprendre les mécanismes moléculaires expliquant les faits observés, la nécessité de commencer à prendre en compte l'impact fonctionnel des duplications devient évident.

Après un premier saut méthodologique durant ma thèse m'ayant fait troquer les outils du biologiste «expérimentateur» pour ceux du bio-informaticien, la prise en compte des aspects fonctionnels m'a conduit à élargir mon champs de compétences en intégrant des techniques informatiques de représentation des connaissances. Ce nouvel outil, que j'ai acquis durant mon séjour à l'INRIA dans le projet Hélix, m'a permis de créer la base de données *MicrOBI*. Cette base me permettra dorénavant de mettre en relations des informations provenant de l'analyse de séquences avec des données fonctionnelles, notamment liées au métabolisme. J'espère profiter de cette possibilité pour obtenir mes premiers résultats liant remaniements chromosomiques et fonction cellulaire. En cela, le projet EvolRep est une première étape prometteuse.

Enfin comme il m'est arrivé par le passé d'adapter des algorithmes pour répondre à mes questions biologiques, ce séjour dans le monde de la représentation des connaissances m'a permis d'appréhender les limites techniques et conceptuelles des systèmes de représentations des connaissances lorsqu'ils sont utilisés pour représenter des données biologiques. Si je n'ai pas la prétention d'être devenu un informaticien du domaine, il existe quand même dans un petit coin de ma tête quelques idées d'adaptation, ne serait-ce que technique, de ces systèmes pour les rendre plus adaptés à ma problématique.

Bibliographie

Abouelhoda,M.I., Kurtz,S. et Ohlebusch,E. (2002) The enhanced suffix array and its applications to genome analysis. In *WABI '02 : Proceedings of the Second International Workshop on Algorithms in Bioinformatics* pp. 449–463 Springer-Verlag, London, UK.

Achaz,G., Boyer,F., Rocha,E.P., Viari,A. et Coissac,E. Extracting approximate repeats from large DNA sequences. in preps.

Achaz,G., Coissac,E., Netter,P. et Rocha,E.P.C. (2003) Associations between inverted repeats and the structural evolution of bacterial genomes. *Genetics,* **164** (4), 1279–1289.

Achaz,G., Coissac,E., Viari,A. et Netter,P. (2000) Analysis of intrachromosomal duplications in yeast Saccharomyces cerevisiae : a possible model for their origin. *Mol Biol Evol,* **17** (8), 1268–1275.

Achaz,G., Netter,P. et Coissac,E. (2001) Study of intrachromosomal duplications among the eukaryote genomes. *Mol Biol Evol,* **18** (12), 2280–2288.

Achaz,G., Rocha,E.P.C., Netter,P. et Coissac,E. (2002) Origin and fate of repeats in bacteria. *Nucleic Acids Res,* **30** (13), 2987–2994.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. et Lipman,D.J. (1997) Gapped blast and psi-blast : a new generation of protein database search programs. *Nucleic Acids Res,* **25** (17), 3389–402.

Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T., Harris,M.A., Hill,D.P., Issel-Tarver,L., Kasarskis,A., Lewis,S., Matese,J.C., Richardson,J.E., Ringwald,M., Rubin,G.M. et Sherlock,G. (2000) Gene ontology : tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* **25** (1), 25–29.

Avery,0. T.,C.M.M. et McCarty,M. (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J. Exp. Med.,* **79**, 137–158.

Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res,* **28** (1), 304–305.

Bairoch,A. et Boeckmann,B. (1991) The swiss-prot protein sequence data bank. *Nucleic Acids Res,* **19 Suppl**, 2247–9.

Baudat,F. et Nicolas,A. (1997) Clustering of meiotic double-strand breaks on yeast chromosome iii. *Proc Natl Acad Sci U S A,* **94** (10), 5213–8.

Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. et Wheeler,D.L. (2000) Genbank. *Nucleic Acids Res,* **28** (1), 15–8.

Benson,G. (1997) Sequence alignment with tandem duplication. *J Comput Biol,* **4** (3), 351–67. 1066-5277.

Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. et Schneider,M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res,* **31** (1), 365–370.

Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A., Causton,H.C., Gaasterland,T., Glenisson,P., Holstege,F.C., Kim,I.F., Markowitz,V., Matese,J.C., Parkinson,H., Robinson,A., Sarkans,U., Schulze-Kremer,S., Stewart,J., Taylor,R., Vilo,J. et Vingron,M. (2001) Minimum information about a microarray experiment (miame)-toward standards for microarray data. *Nat Genet,* **29** (4), 365–71.

Bridges,C. (1936) The bar "gene" a duplication. *Science,* **83**, 210–11.

Brooksbank,C., Cameron,G. et Thornton,J. (2005) The european bioinformatics institute's data resources : towards systems biology. *Nucleic Acids Res,* **33** (Database issue), D46–53.

Casaregola,S., Nguyen,H.V., Lepingle,A., Brignon,P., Gendre,F. et Gaillardin,C. (1998) A family of laboratory strains of saccharomyces cerevisiae carry rearrangements involving chromosomes i and iii. *Yeast,* **14** (6), 551–64.

Chase,M. et Hershey,A. (1952) Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J. Gen. Physiol.,* **36**, 39–56.

Chedin,F., Dervyn,E., Dervyn,R., Ehrlich,S.D. et Noirot,P. (1994) Frequency of deletion formation decreases exponentially with distance between short direct repeats. *Mol Microbiol,* **12** (4), 561–9.

Claudel-Renard,C., Chevalet,C., Faraut,T. et Kahn,D. (2003) Enzyme-specific profiles for genome annotation : priam. *Nucleic Acids Res,* **31** (22), 6633–9.

Coissac,E., Maillier,E. et Netter,P. (1997) A comparative study of duplications in bacteria and eukaryotes : the importance of telomeres. *Mol Biol Evol,* **14** (10), 1062–1074.

Coissac,E., Maillier,E., Robineau,S. et Netter,P. (1996) Sequence of a 39,411 bp DNA fragment covering the left end of chromosome VII of Saccharomyces cerevisiae. *Yeast,* **12** (15), 1555–1562.

Corpet,F., Gouzy,J. et Kahn,D. (1998) The ProDom database of protein domain families. *Nucleic Acids Res,* **26** (1), 323–6.

Dayhoff,M., Eck,R., M.A.,C. et Sochard,M. (1965) *Atlas of Protein Sequence and Structure*, vol. 1,. National Biomedical Research Foundation, Silver Spring.

Dayhoff,M.O., Schwartz,R.M., Chen,H.R., Barker,W.C., Hunt,L.T. et Orcutt,B.C. (1981) Nucleic acid sequence database. *DNA,* **1** (1), 51–8.

Dayhoff,M.O., Schwartz,R.M. et Orcutt,B.C. (1978) *Atlas of protein sequece and structure* vol. supplement 3,. Washington, DC : Dayhoff, M. O. National Biomedical Research Foundation edition, pp. 345–352.

Delgrange,O. et Rivals,E. (2004) Star : an algorithm to search for tandem approximate repeats. *Bioinformatics,* **20** (16), 2812–20. 1367-4803 Evaluation Studies Journal Article Validation Studies.

Downie,J.A., Stewart,J.W., Brockman,N., Schweingruber,A.M. et Sherman,F. (1977) Structural gene for yeast iso-2-cytochrome c. *J Mol Biol,* **113** (2), 369–384.

Duret,L., Mouchiroud,D. et Gouy,M. (1994) Hovergen : a database of homologous vertebrate genes. *Nucleic Acids Res,* **22** (12), 2360–5.

Edman,P. et Begg,G. (1967) A protein sequenator. *Eur J Biochem,* **1** (1), 80–91.

Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M. et et al. (1995) Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science,* **269** (5223), 496–512.

Garrity,G., Winters,M. et Searles,D. (2001) *Bergey's Manual of Systematic Bacteriology* vol. 1,. New York : Bergey's Manual springer edition, pp. 1–39.

Gattiker,A., Michoud,K., Rivoire,C., Auchincloss,A.H., Coudert,E., Lima,T., Kersey,P., Pagni,M., Sigrist,C.J.A., Lachaize,C., Veuthey,A.L., Gasteiger,E. et Bairoch,A. (2003) Automated annotation of microbial proteomes in SWISS-PROT. *Comput Biol Chem,* **27** (1), 49–58.

Gilbert,W. et Maxam,A. (1973) The nucleotide sequence of the lac operator. *Proc Natl Acad Sci U S A,* **70** (12), 3581–4.

Henikoff,S. et Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A,* **89** (22), 10915–9.

Hollander,M. et Wolfe,D.A. (1973) *Nonparametric statistical inference*. New York : John Wiley & Sons pp. 185–194.

Holliday,R. (1964) The induction of mitotic recombination by mitomycin c in ustaligo and saccharomyces. *Genetics,* **50** (0016-6731), 323–35.

Jacob,F. (1981) *Le jeu des possibles - Essai sur la diversité du vivant.* Fayard, Paris.

Kalnins,A., Otto,K., Ruther,U. et Muller-Hill,B. (1983) Sequence of the lacz gene of escherichia coli. *Embo J,* **2** (4), 593–7.

Kanehisa,M. (2002) The KEGG database. *Novartis Found Symp,* **247**, 91–101 ; discussion 101–3, 119–28, 244–52. 1528-2511 Journal Article Review Review, Tutorial.

Karin,M., Najarian,R., Haslinger,A., Valenzuela,P., Welch,J. et Fogel,S. (1984) Primary structure and transcription of an amplified genetic locus : the CUP1 locus of yeast. *Proc Natl Acad Sci U S A,* **81** (2), 337–341.

Kärkkäinen,J. et Sanders,P. (2003) Simple linear work suffix array construction. In *Proc. 13th International Conference on Automata, Languages and Programming* Springer.

Karlin,S. et Ost,F. (1985) Maximal segmental match length among random sequences from a finite alphabet. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, (Cam,L.M.L. et Olshen,R.A., eds), vol. 1, pp. 225–243 Wadsworth.

Karp,R.M., Miller,R.E. et Rosenberg,A.L. (1972) Rapid identification of repeated patterns in strings, trees and arrays. In *Conference record, Fourth Annual ACM Symposium on Theory of Computing : papers presented at the symposium, Denver, Colorado, May 1, 2, 3, 1972*, (ACM, ed.), pp. 125–136 ACM Press, New York, NY 10036, USA.

Kellis,M., Birren,B.W. et Lander,E.S. (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature,* **428** (6983), 617–624.

Kendall,M.G. (1938) A new measure of rank correlation. *Biometrika,* **30** (1–2), 81–93.

Kersey,P., Bower,L., Morris,L., Horne,A., Petryszak,R., Kanz,C., Kanapin,A., Das,U., Michoud,K., Phan,I., Gattiker,A., Kulikova,T., Faruque,N., Duggan,K., Mclaren,P., Reimholz,B., Duret,L., Penel,S., Reuter,I. et Apweiler,R. (2005) Integr8 and genome reviews : integrated views of complete genomes and proteomes. *Nucleic Acids Res,* **33** (Database issue), D297–302.

Kneale,G.G. et Kennard,O. (1984) The EMBL nucleotide sequence data library. *Biochem Soc Trans,* **12** (6), 1011–4.

Kolpakov,R., Bana,G. et Kucherov,G. (2003) MREPS : efficient and flexible detection of tandem repeats in dna. *Nucleic Acids Res,* **31** (13), 3672–8. 1362-4962.

Kurtz,S., Ohlebusch,E., Schleiermacher,C., Stoye,J. et Giegerich,R. (2000) Computation and visualization of degenerate repeats in complete genomes. *Proc Int Conf Intell Syst Mol Biol,* **8** (1553-0833), 228–38.

Kurtz,S., Phillippy,A., Delcher,A.L., Smoot,M., Shumway,M., Antonescu,C. et Salzberg,S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol,* **5** (2), R12.

Kurtz,S. et Schleiermacher,C. (1999) REPuter : Fast Computation of Maximal Repeats in Complete Genomes. *Bioinformatics,* **15** (5), 426–427.

Landau,G.M., Schmidt,J.P. et Sokol,D. (2001) An algorithm for approximate tandem repeats. *J Comput Biol,* **8** (1), 1–18. 1066-5277.

Lefebvre,A., Lecroq,T., Dauchel,H. et Alexandre,J. (2003) Forrepeats : detects repeats on entire chromosomes and between genomes. *Bioinformatics,* **19** (3), 319–26. 1367-4803 Evaluation Studies Journal Article Validation Studies.

Lipman,D.J. et Pearson,W.R. (1985) Rapid and sensitive protein similarity searches. *Science,* **227** (4693), 1435–41.

Lovett,S.T., Gluckman,T.J., Simon,P.J., Sutera,V.A.J. et Drapkin,P.T. (1994) Recombination between repeats in escherichia coli by a reca-independent, proximity-sensitive mechanism. *Mol Gen Genet,* **245** (3), 294–300.

Lowe,T.M. et Eddy,S.R. (1997) tRNAscan-SE : a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res,* **25** (5), 955–64.

Manber,U. et Myers,G. (1990) Suffix arrays : a new method for on-line string searches. In *SODA '90 : Proceedings of the first annual ACM-SIAM symposium on Discrete algorithms* pp. 319–327 Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Maxam,A.M. et Gilbert,W. (1977) A new method for sequencing DNA. *Proc Natl Acad Sci U S A,* **74** (2), 560–4.

McCreight,E.M. (1976) A space-economical suffix tree construction algorithm. *J. ACM,* **23** (2), 262–272.

Mendel,G. (1865) Versuche über plflanzenhybriden. In *Verhandlungen des naturforschenden Vereines in Brünn* pp. 3–47, Brünn.

Muller,H., Prokofyeva-Belgovskaya,A. et Kossikov,K. (1936) Unequal crossing-over in the bar mutant as a result of duplication of a minute chromosome section. *CR (Doklady) Acad. Sci. URSS,* **1**, 87–88.

Needleman,S.B. et Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol,* **48** (3), 443–53.

Nei,M. et Gojobori,T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol,* **3** (5), 418–26.

Pellegrini,M., Marcotte,E.M. et Yeates,T.O. (1999) A fast algorithm for genome-wide analysis of proteins with repeated sequences. *Proteins,* **35** (4), 440–6.

Potter,H. et Dressler,D. (1976) On the mechanism of genetic recombination : electron microscopic observation of recombination intermediates. *Proc Natl Acad Sci U S A,* **73** (9), 3000–4.

Quesneville,H., Nouaud,D. et Anxolabehere,D. (2003) Detection of new transposable element families in drosophila melanogaster and anopheles gambiae genomes. *J Mol Evol,* **57 Suppl 1** (0022-2844), S50–9.

Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G., Montoya,M., Miller,N., Mueller,L.A., Mundodi,S., Reiser,L., Tacklind,J., Weems,D.C., Wu,Y., Xu,I., Yoo,D., Yoon,J. et Zhang,P. (2003) The arabidopsis information resource (tair) : a model organism database providing a centralized, curated gateway to arabidopsis biology, research materials and community. *Nucleic Acids Res,* **31** (1), 224–8. 1362-4962 Journal Article.

Rocha,E.P., Danchin,A. et Viari,A. (1999) Analysis of long repeats in bacterial genomes reveals alternative evolutionary mechanisms in bacillus subtilis and other competent prokaryotes. *Mol Biol Evol,* **16** (9), 1219–30. 0737-4038.

Sagot,M., Viari,A., Pothier,J. et Soldano,H. (1994) Finding flexible patterns in a text - an application to 3d molecular matching. *Comput. Application in the Biosciences,* **11**, 59–70.

Sanger,F., Coulson,A.R., Friedmann,T., Air,G.M., Barrell,B.G., Brown,N.L., Fiddes,J.C., Hutchison,C. A.,r., Slocombe,P.M. et Smith,M. (1978) The nucleotide sequence of bacteriophage phix174. *J Mol Biol,* **125** (2), 225–46.

Sanger,F., Nicklen,S. et Coulson,A.R. (1977) Dna sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A,* **74** (12), 5463–7.

Sanger,F., Thompson,E.O. et Kitai,R. (1955) The amide groups of insulin. *Biochem J,* **59** (3), 509–18.

Smit,A., Hubley,R. et Green,P. RepeatMasker Open-3.0. *http ://www.repeatmasker.org.*

Smith,T.F. et Waterman,M.S. (1981) Identification of common molecular subsequences. *J Mol Biol,* **147** (1), 195–197.

Soldano,H., Viari,A. et Champesme,M. (1995) Searching for flexible repeated patterns using a non transitive similarity relation. *Pattern Recognition Letters,* **16**, 233–245. JAIJAI.

Sutton,E. (1943) Bar eye in drosophila melanogaster : a cytological analysis of some mutations and reverse mutations. *Genetics,* **28**, 97–107.

Sverdlov,E.D., Monastyrskaya,G.S., Chestukhin,A.V. et Budowsky,E.I. (1973) The primary structure of oligonucleotides. Partial apurinization as a method to determine the positions of purine and pyrimidine residues. *FEBS Lett,* **33** (1), 15–7.

Taylor,C.F., Paton,N.W., Garwood,K.L., Kirby,P.D., Stead,D.A., Yin,Z., Deutsch,E.W., Selway,L., Walker,J., Riba-Garcia,I., Mohammed,S., Deery,M.J., Howard,J.A., Dunkley,T., Aebersold,R., Kell,D.B., Lilley,K.S., Roepstorff,P., Yates,J. R.,r., Brass,A., Brown,A.J., Cash,P., Gaskell,S.J., Hubbard,S.J. et Oliver,S.G. (2003) A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nat Biotechnol,* **21** (3), 247–54.

Tice,S.C. (1914) A new sex-linked character in drosophila. *Biological Bulletin,* **26**, 221–230.

Tsubota,S.I., Rosenberg,D., Szostak,H., Rubin,D. et Schedl,P. (1989) The cloning of the bar region and the b breakpoint in drosophila melanogaster : evidence for a transposon-induced rearrangement. *Genetics,* **122** (4), 881–90.

Uchida,H. (1986) [dna data bank of japan]. *Tanpakushitsu Kakusan Koso,* **29 Suppl**, 159–62.

Ukkonen,E. (1992) Constructing suffix trees on-line in linear time. In *Proceedings of the IFIP 12th World Computer Congress on Algorithms, Software, Architecture - Information Processing '92, Volume 1* pp. 484–492 North-Holland.

Venema,J. et Tollervey,D. (1999) Ribosome synthesis in Saccharomyces cerevisiae. *Annu Rev Genet,* **33**, 261–311.

Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A., Gocayne,J.D., Amanatides,P., Ballew,R.M., Huson,D.H., Wortman,J.R., Zhang,Q., Kodira,C.D., Zheng,X.H., Chen,L., Skupski,M., Subramanian,G., Thomas,P.D., Zhang,J., Gabor Miklos,G.L., Nelson,C., Broder,S., Clark,A.G., Nadeau,J., McKusick,V.A., Zinder,N., Levine,A.J., Roberts,R.J., Simon,M., Slayman,C., Hunkapiller,M., Bolanos,R., Delcher,A., Dew,I., Fasulo,D., Flanigan,M., Florea,L., Halpern,A., Hannenhalli,S., Kravitz,S., Levy,S., Mobarry,C., Reinert,K., Remington,K., Abu-Threideh,J., Beasley,E., Biddick,K., Bonazzi,V., Brandon,R., Cargill,M., Chandramouliswaran,I., Charlab,R., Chaturvedi,K., Deng,Z., Di Francesco,V., Dunn,P., Eilbeck,K., Evangelista,C., Gabrielian,A.E., Gan,W., Ge,W., Gong,F., Gu,Z., Guan,P., Heiman,T.J., Higgins,M.E., Ji,R.R., Ke,Z., Ketchum,K.A., Lai,Z., Lei,Y., Li,Z., Li,J., Liang,Y., Lin,X., Lu,F., Merkulov,G.V., Milshina,N., Moore,H.M., Naik,A.K., Narayan,V.A., Neelam,B., Nusskern,D., Rusch,D.B., Salzberg,S., Shao,W., Shue,B., Sun,J., Wang,Z., Wang,A., Wang,X., Wang,J., Wei,M., Wides,R., Xiao,C., Yan,C., Yao,A., Ye,J., Zhan,M., Zhang,W., Zhang,H., Zhao,Q., Zheng,L., Zhong,F., Zhong,W., Zhu,S., Zhao,S., Gilbert,D., Baumhueter,S., Spier,G., Carter,C., Cravchik,A., Woodage,T., Ali,F., An,H., Awe,A., Baldwin,D., Baden,H., Barnstead,M., Barrow,I., Beeson,K., Busam,D., Carver,A., Center,A., Cheng,M.L., Curry,L., Danaher,S., Davenport,L., Desilets,R., Dietz,S., Dodson,K., Doup,L., Ferriera,S., Garg,N., Gluecksmann,A., Hart,B., Haynes,J., Haynes,C., Heiner,C., Hladun,S., Hostin,D., Houck,J., Howland,T., Ibegwam,C., Johnson,J., Kalush,F., Kline,L., Koduru,S., Love,A., Mann,F., May,D., McCawley,S., McIntosh,T., McMullen,I., Moy,M., Moy,L., Murphy,B., Nelson,K., Pfannkoch,C., Pratts,E., Puri,V., Qureshi,H., Reardon,M., Rodriguez,R., Rogers,Y.H., Romblad,D., Ruhfel,B., Scott,R., Sitter,C., Smallwood,M., Stewart,E., Strong,R., Suh,E., Thomas,R., Tint,N.N., Tse,S., Vech,C., Wang,G., Wetter,J., Williams,S., Williams,M., Windsor,S., Winn-Deen,E., Wolfe,K., Zaveri,J., Zaveri,K., Abril,J.F., Guigo,R., Campbell,M.J., Sjolander,K.V., Karlak,B., Kejariwal,A., Mi,H., Lazareva,B., Hatton,T., Narechania,A., Diemer,K., Muruganujan,A., Guo,N., Sato,S., Bafna,V., Istrail,S., Lippert,R., Schwartz,R., Walenz,B., Yooseph,S., Allen,D., Basu,A., Baxendale,J., Blick,L., Caminha,M., Carnes-Stine,J., Caulk,P., Chiang,Y.H., Coyne,M., Dahlke,C., Mays,A., Dombroski,M., Donnelly,M., Ely,D., Esparham,S., Fosler,C., Gire,H., Glanowski,S., Glasser,K., Glodek,A., Gorokhov,M., Graham,K., Gropman,B., Harris,M., Heil,J., Henderson,S., Hoover,J., Jennings,D., Jordan,C., Jordan,J., Kasha,J., Kagan,L., Kraft,C., Levitsky,A., Lewis,M., Liu,X., Lopez,J., Ma,D., Majoros,W., McDaniel,J., Murphy,S., Newman,M., Nguyen,T., Nguyen,N., Nodell,M., Pan,S., Peck,J., Peterson,M., Rowe,W., Sanders,R., Scott,J., Simpson,M., Smith,T., Sprague,A., Stockwell,T., Turner,R., Venter,E., Wang,M., Wen,M., Wu,D., Wu,M., Xia,A., Zandieh,A. et Zhu,X. (2001) The sequence of the human genome. *Science,* **291** (5507), 1304–51.

Viswanathan,M., Muthukumar,G., Cong,Y.S. et Lenard,J. (1994) Seripauperins of Saccharomyces cerevisiae : a new multigene family encoding serine-poor relatives of serine-rich proteins. *Gene,* **148** (1), 149–153.

Waterman,M.S. et Eggert,M. (1987) A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *J Mol Biol,* **197** (4), 723–8.

Watson,J.D. et Crick,F.H. (1953) Molecular structure of nucleic acids ; a structure for deoxyribose nucleic acid. *Nature,* **171** (4356), 737–8.

Weiner,P. (1973) Linear pattern matching algorithms. In *FOCS* pp. 1–11.

Wheeler,D.L., Chappey,C., Lash,A.E., Leipe,D.D., Madden,T.L., Schuler,G.D., Tatusova,T.A. et Rapp,B.A. (2000) Database resources of the national center for biotechnology information. *Nucleic Acids Res,* **28** (1), 10–4.

Wolfe,K.H. et Shields,D.C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature,* **387** (6634), 708–713.

Table des figures

Cinquième partie

Annexes

Sequence of a 39,411 bp DNA fragment covering the left end of chromosome VII of Saccharomyces cerevisiae

Coissac, E and Maillier, E and Robineau, S and Netter, P

Nous avons séquencé un fragment d'ADN de 39411 pb incluant les régions télomérique et subtélomérique gauches du chromosome VII de *Saccharomyces cerevisiae*. Nous avons identifié 19 phases ouvertes de lecture (ORF) : 6 correspondent à des gènes connus de la levure (*ADH4*, *FZF1*, *HKB*, *RTG2*, *HFM1* et *PDE1*), neuf présentent une similarité avec d'autres gènes connus et quatre ne possèdent aucune similarité avec aucun autre gène connu. La taille moyenne de ces ORF semble corrélée avec leur position, les huit ORF les plus proches de l'extrémité télomérique sont plus courtes que les onze autres. Les deux groupes sont séparés par une région de 4,5 kb dépourvue de toute ORF significative. Une ORF nommée *NRF120* est un nouveau membre de la famille des séripaupérine dont un membre est présent dans chacune des régions subtélomériques de la levure séquencées à ce jour. Cette séquence à été déposée à l'EMBL sous le numéro d'accession X94357.

YEAST VOL. 12: 1555–1562 (1996)

VII Yeast Sequencing Reports

Sequence of a 39 411 bp DNA Fragment Covering the Left End of Chromosome VII of *Saccharomyces cerevisiae*

ERIC COISSAC*, EVELYNE MAILLIER, SYLVIANE ROBINEAU AND PIERRE NETTER

Centre de Génétique Moléculaire, Avenue de la Terrasse, F91198 Gif sur Yvette, France
Laboratoire propre du CNRS associé à l'Université Pierre et Marie Curie

Received 11 April 1996; accepted 4 July 1996

We have sequenced a DNA fragment of 39 411 bp which includes part of the left telomere of chromosome VII of *Saccharomyces cerevisiae*. We have identified 19 open reading frames (ORFs); six correspond to known yeast genes (*ADH4*, *FZF1*, *HKB*, *RTG2*, *HFM1* and *PDE1*), nine have similarity with other genes and four exhibit no significant similarity with any known gene. The average size of these ORFs seems to be related to their location, the eight ORFs nearest the telomere being shorter than the 11 others. These two groups of genes are separated by a region of 4·5 kb devoid of significant ORFs. One ORF, NRF120, is a new member of the seripauperine family, represented once in all sequenced yeast chromosomes, in a subtelomeric location. This sequence has been entered in the EMBL data library under accession number X94357.

KEY WORDS — chromosome sequencing; *Saccharomyces cerevisiae*; *ADH4*; *FZF1*; *HKB*; *RTG2*; *HFM1*; *PDE1*

INTRODUCTION

In the framework of the European yeast genome sequencing project, we have sequenced a 39 411 bp region which includes the left telomere of chromosome VII of *Saccharomyces cerevisiae*. The centromeric side of this fragment is carried by the cosmid pEGH420, and the telomeric side by the plasmid pEL164. Nineteen open reading frames (ORFs) were defined from the sequence data. This 39 411 bp sequence includes the genes *ADH4*, *FZF1*, *HKB*, *RTG2*, *HFM1* and *PDE1*, which have been described previously. A small ORF of 363 bp corresponds to a new member of the seripauperine family (Viswanathan *et al.*, 1994).

*Corresponding author.

CCC 0749-503X/96/151555–08

MATERIALS AND METHODS

Bacterial strains and vector

All the subcloning and sequencing work was done in the *Escherichia coli* strain *XLI-Blue* (F′::Tn*10* *proA*$^+$*B*$^+$ *lacI*^{q} Δ(*lacZ*)*M15*/*recA1* *endA1* *gyrA96* (Nal^r) *thi* *hsdR17*(r_k^- m_k^+) *supE44* *relA1* *lac*).

The cosmid pEGH420 and its *Eco*RI restriction map were provided by H. Tettelin. It was isolated from a cosmid library constructed from the strain FY1679 (Thierry *et al.*, 1995). This cosmid is located 5·5 kb from the beginning of the telomeric CA repeats in the left arm of chromosome VII. The gap between the CA repeats and the cosmid pEGH420 was filled by the plasmid pEL164 provide by Ed Louis (Louis and Borts, 1995) (Figure 1). Culture, subcloning, *E. coli* transformation and all basic molecular biology manipulations

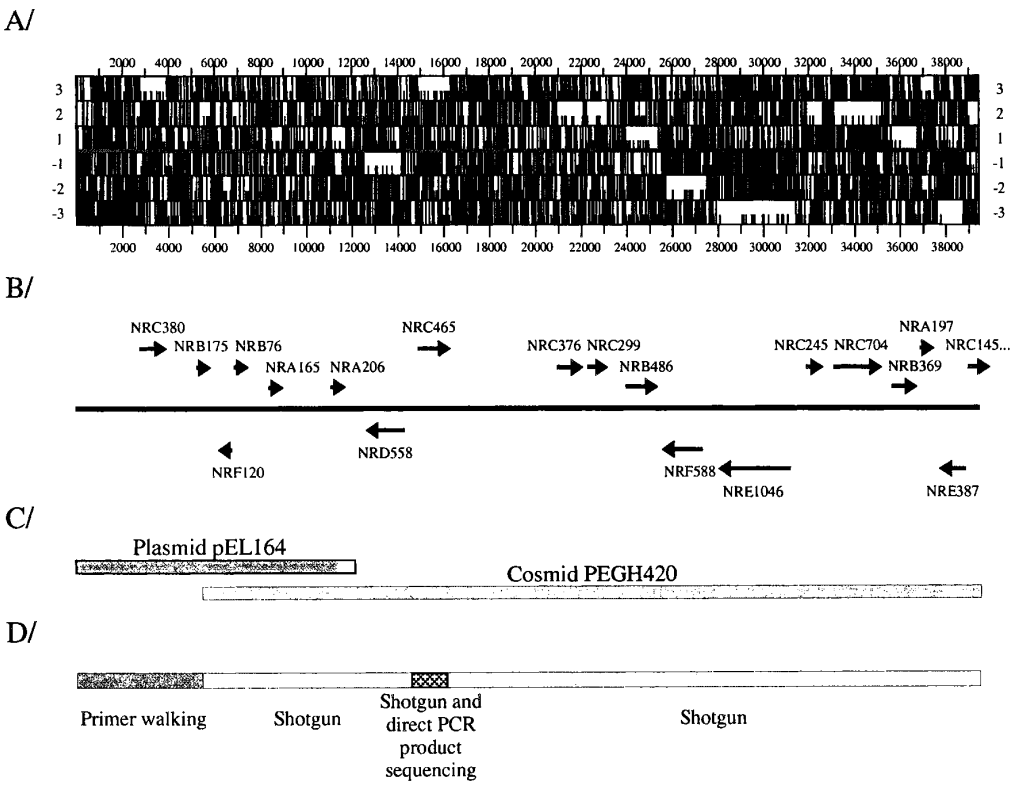


Figure 1. Physical map of the left telomeric and subtelomeric region of chromosome VII. The left telomere of chromosome VII is located on the left side of the figure. (A) Map of the Start and Stop codon in the six reading frames created by the Strider program (Marck, 1988). The open reading frames appear as white boxes. (B) Map and name of the ORFs retained from sequence data. (C) Relative position of the plasmid pEL164 and of the cosmid pEGH420. The centromeric extremity of the plasmid pEL164 is not precisely known. (D) Sequencing strategy used.

were performed using standard protocols (Maniatis *et al.*, 1982).

Sequencing strategy

Fragments from the entire cosmid pEGH420 were subcloned, after sonication and repair, into the plasmid vector pBSIISK$^+$ (Stratagene, La Jolla, CA, USA) in the *Eco*RV restriction site of the polylinker. The sonication was performed on 50 µg of cosmid DNA in 10 mM-Tris, 1 mM-EDTA pH 8·0 buffer. After repair by Mung Bean Nuclease and T4 DNA Polymerase, the fragments of 0·5 to 2 kb were isolated from an agarose gel. 366 clones from this library were sequenced. A region of 1·5 kb on the 5′ side of the gene *ADH4* was not represented in this library. This gap was filled by sequencing a small library made by subcloning *Alu*I, *Hae*III or *Rsa*I total digests of several independent polymerase chain reaction amplifications of this region. These small restriction fragments were cloned into the *Eco*RV site of pBSIISK$^+$. 32 subclones of this library were used to close the gap. In addition, 50 sequence-specific oligonucleotides were used to close small gaps or to complete the sequencing of the second strand.

The 5·5 kb fragment of pEL164 corresponding to the region between the telomeric side of pEGH420 and the telomeric CA repeats was sequenced directly on this plasmid by a primer walking strategy. For this purpose, 47 new oligonucleotides were synthesized.

The sequence was established on double-stranded DNA templates prepared either by alkaline lysis and purified on an anionic exchange

column (Quiagen or Pharmacia EasyPrep System) or by ultracentrifugation on a CsCl density gradient. The sequence was established using KS and SK 5′ fluorescein end-labelled primers. The sequences obtained from the sequence-specific primers were established with a fluorescein-15-dATP internal labelling system. The electrophoresis was performed on an automated sequencer (ALF Pharmacia). An average output of 300 bp was read for each clone. In total, 228 220 bp were read, resulting in an average redundancy of 5·8 per base.

Computer analysis

The completed sequence was assembled using the Staden package of programs (Dear and Staden, 1991). The sequence analysis was performed using GDE, FASTA (Pearson and Lipman, 1988), BLAST (Altschul *et al.*, 1990) and home-made programs on a Sun SPARCstation 20.

RESULTS AND DISCUSSION

Chromosome VII is one of the largest chromosomes of the yeast *S. cerevisiae* (about 1·15 Mb long). The physical map of this chromosome was established by H. Tettelin from a cosmid library constructed by Thierry *et al.* (1995). pEGH420 is the leftmost cosmid isolated from this library. Prior to sequencing, a restriction map of the cosmid pEGH420 was constructed using the enzymes *Eco*RI, *Bam*HI, *Bsu*36I, *Sal*I and *Not*I. The left telomeric extremity of the map was completed using the plasmid pEL164 supplied by Ed Louis (Louis and Borts, 1995). The total length of the sequenced region is 39 411 bp and the first 36 bp are CA repeat motifs characteristic of yeast telomeres. From the information of the yeast genetic map (Mortimer *et al.*, 1992), this region was expected to contain the genes *ADH4* and *HXK2*.

Sequencing reveals a GC content of 37%. Eighteen ORFs of more than 100 codons (according to the general conventions of the Yeast Genome Sequencing Project) are present in this region. A shorter ORF of 76 codons (NRB76) may also be significant (Figure 3). The ORF NRC145, located at the centromeric extremity of the fragment, is truncated and continues on the cosmid pEGH175 (Vandenbol, M. *et al.*, 1995; accession number EMBL: Z49149).

The distribution and the size of the ORFs seem to define three domains. The first one (20 kb), on the centromeric side, contains 11 ORFs; in this area, the average size of the ORFs is 1·4 kb and the average 'intergenic' region is 0·4 kb. The second domain (16 kb), nearest to the telomere, contains eight ORFs, the average size of these ORFs is 0·8 kb and five of them are smaller than 618 bp. The average size of the 'intergenic' region is 1·0 kb. These two domains are separated by a third one of 4·5 kb, where no significant ORFs are present (Figure 1A). The characteristics of the ORF distribution of the first region agree with those described for the sequences of chromosome II (Feldmann *et al.*, 1994), III (Oliver *et al.*, 1992), VIII (Johnston *et al.*, 1994) and XI (Dujon *et al.*, 1994). The lower density of ORFs in the second region has already been observed in the other subtelomeric regions (Louis, 1995).

Six ORFs correspond to previously known genes: *ADH4* (Williamson and Paquin, 1987), *FZF1* (Breitwieser *et al.*, 1993), *HXK2* (Stachelek *et al.*, 1986), *RTG2* (Liao and Butow, 1993), *HFM1* (West *et al.*, 1995, accession number gb: U22156) and *PDE1* (Nikawa *et al.*, 1987).

Analysis of putative ORF products

The main genetic elements present on the 39 411 bp contig are described in Table 1. Results of a more precise analysis of each ORF are also given in this table. For some ORFs, additional information is presented in the following paragraphs.

ORF NRC465

This ORF corresponds to the gene *ADH4*. It encodes a putative peptide of 465 amino acids. The determination of the initiation transcription site (Williamson and Paquin, 1987) seems to indicate that the real start codon is the codon M_{84}. We observe 22 differences with the sequence previously published in EMBL (accession number X05992). From the peptide sequence ('long' ORF version) described in PIR (accession number S07614), five of these differences lead to changes in the amino acid sequence, the others correspond to silent changes. Amongst the modifications, three correspond to replacement by similar amino acids: the Q_{171} in NRC465 becomes an E, E_{422} becomes a D and E_{431} becomes a D. The two other substitutions are: D_{143} changed to G and F_{393} changed to C. It should be noted that the majority of changes are silent or lead to conservative amino acid replacements.

Table 1. Main genetic elements present on the left telomeric and subtelomeric region of chromosome VII, and quick analysis of the ORFs present.

Position	Type	Strand	Observation	Codon bias	Molecular weight (kDa)	Isoelectric point	Length (aa)
506–531	Misc. feature		ABF1/OBF1-consensus 'Abf1p binding site'				
733–743	Misc. feature		ARS-consensus				
863–1079	LTR		Ty5-LTR				
1080–1135	LTR		Ty5-LTR remnant				
2379–2389	Misc. feature		ARS-consensus				
2790–3929	ORF	W	Name: NRC380 Exhibits high similarity with ORF *YBR302c YFL062w YHL048w YIR049c YJR161c YKL219w*	0·01	44·8	8·6	380
5312–5836	ORF	W	Name: NRB175 No significant similarity was found in Swissprot and PIR databases	−0·12	20·1	6·1	175
6293–6652	ORF	C	Name: NRF120 Member of the seripauperine family (Figure 2)	0·75	12·8	5·1	120
6860–7087	ORF	W	Name: NRB76 Probably a pseudo-gene member of the family constituted by the ORF *YBL108w, YCR103c, YIR040c, YKL223w* (Figure 3)	−0·15	9·0	8·7	76
8470–8964	ORF	W	Name: NRA165 Similar to the ORF *YIR039c* accession number: SP: P40583 Hypothetical aspartyl proteinase precursor	0·04	17·9	4·7	165
11110–11727	ORF	W	Name: NRA206 No significant similarity was found in Swissprot and PIR databases	0·49	22·1	4·3	206
12484–14157	ORF	C	Name: NRD558 Similar to the gene *MNN1* (ORF *YER001w*) accession number: SP: P39106 and to the ORF *YIL014w* accession number: SP: P40549	−0·05	64·8	5·4	558
14910–16304	ORF	W	Name: NRC465 Alcohol dehydrogenase IV Corresponds to gene *ADH4* accession number: EMBL: X05992 PIR: S07614	0·53	50·6	8·6	465
20978–22105	ORF	W	Name: NRC376 Match with the Swissprot entry YFZF_YEAST accession number: SP: P32804 Hypothetical protein of unknown function	0·35	41·5	5·5	376
22304–23200	ORF	W	Name: NRC299 Zinc finger protein Corresponds to gene *FZF1* accession number: EMBL: X67787 SP: P32805	0·04	34·0	7·9	299

Continued

Table 1. *Continued.*

Position	Type	Strand	Observation	Codon bias	Molecular weight (kDa)	Isoelectric point	Length (aa)
23935–25392	ORF	W	Name: NRB486 Hexokinase B Corresponds to gene *HXK2* accession number: EMBL: M11181 PIR: S33656	0·75	53·9	5·2	486
25721–27484	ORF	C	Name: NRF588 Corresponds to gene *RTG2* accession number: EMBL: M97691 PIR: B44344	0·12	65·6	8·2	588
27924–31061	ORF	C	Name: NRE1046 DNA/RNA helicase Corresponds to gene *HFM1* accession number: EMBL: U22156 PIR: S59815	−0·01	118·8	7·5	1046
31898–32632	ORF	W	Name: NRC245 No significant similarity was found in Swissprot and PIR databases	0·11	28·1	4·3	245
33098–35209	ORF	W	Name: NRC704 Match with one EST of *S. cerevisiae* (98% identical on 263 bp) accession number: EMBL: T17530 (Burns *et al.*, 1994)	0·01	82·8	8·5	704
35567–35577	Misc. feature		ARS-consensus				
35653–36759	ORF	W	Name: NRB369 Phosphodiesterase low affinity Corresponds to gene *PDE1* accession number: EMBL: M17781 PIR: S05879	0·08	42·0	5·7	369
36933–37523	ORF	W	Name: NRA197 Match with one EST of *S. cerevisiae* (100% identical on 277 bp): accession number: EMBL: T38166 Similar to the ORF *YHR036w* accession number: SP: P38770	−0·04	22·8	8·4	197
37620–38780	ORF	C	Name: NRE387 No significant similarity was found in Swissprot and PIR databases	0·05	44·5	6·1	387
38975–39409	ORF	W	Name: NRC145 Partial sequence of the ORF, continues on the cosmid pEGH175 accession number: EMBL: Z49149				

Coordinates are given from the telomeric extremity (position 1) according to the EMBL entry (accession number X94357). All elements (ORF, LTR, Motifs) were identified by Karl Kleine at Martinsried Institute for Protein Sequences. The calculated molecular weights and isoelectric points were determined on the EXPASY server of the Geneva University Hospital and the University of Geneva, Geneva, Switzerland. The codon bias was calculated by the Bennetzen formula (Bennetzen and Hall, 1982). The W (Watson) strand corresponds to the strand submitted to EMBL.

```
    1          11         21         31         41         51         61         70
    |          |          |          |          |          |          |          |
 1  MVKLTSIAAG VAAIAATAS- A---TTTLAQ SDERVNLVEL GVYVSDIRAH LAQYYSFQVA HPTETYPVEI 66  II/R       (YBR301w)
 1  MVKLTSIAAG VAAIAAGI-A AAPATTTLSP SDERVNLVEL GVYVSDIRAH LAQYYLFQAA HPTETYPVEI 69  III/R      (YCR104w)
 1  MVKLTSIAAG VAAIAATASA ----TTTLAQ SDERVNLVEL GVYVSDIRAH LAQYYMFQAA HPTETYPVEV 66  IV/R
 1  MVKLTSIAAG VAAIAATASA ----TTTLAQ SDERVNLVEL GVYVSDIRAH LAQYYSFQAA HPTETYPIEV 66  V/L        (YEL049w)
 1  MVKLTSIAAG VAAIAAGASA AA--TTTLSQ SDERVNLVEL GVYVSDIRAH LAEYYSFQAA HPTETYPVEI 68  VI/L       (YFL020c)
 1  MVKLTSIAAG VAAIAATASA ----TTTLAQ SDERVNLVEL GVYVSDIRAH LAQYYMFQAA HPTETYPVEV 66  VII/L      (NRF120)
 1  MVKLTSIAAG VAAIAATASA ----TTTLAQ SDERVNLVEL GVYVSDIRAH LAQYYMFQAA HPTETYPVEV 66  VIII/L     (YHL046c)
 1  MVKLTSIAAG VAAIAATASA ----TTTLAQ SDERVNLVEL GVYVSDIRAH LAQYYMFQAA HPTETYPVEV 66  IX/L       (YIL176c)
 1  MVKLTSIAAG VAAIAAGV-A AAPATTTLSP SDERVNLVEL GVYVSDIRAH LAQYYLFQAA HPSETYPVEI 69  IX/R       (YIR041w)
 1  MVKLTSIAAG VAAIAATASA ----TTTLAQ SDERVNLVEL GVYVSDIRAH LAQYYMFQAA HPTETYPVEV 66  X/L        (YJL223c)
 1  MVKLTSIAAG VAAIAAGV-A AAPATTTLSP SDERVNLVEL GVYVSDIRAH LAQYYLFQAA HPSETYPVEI 69  XI/L       (YKL224c)
 1  MVKLTSIAAG VAAIAATASA ----TTTLAQ SDERVNLVEL GVYVSDIRAH LAQYYMFQAA HPTETYPVEV 66  XII/L
 1  MVKLTSIAAG VAAIAATASA ----TTTIAQ SDERVNLVEL GVYVSDIRAH LAQYYMFQAA HPTETYPVEV 66  XII/R
 1  MVKLTSIAAG VAAIAAGV-A AAPATTTLSP SDERVNLVEL GVYVSDIRAH LAQYYLFQAA HPTETYPVEI 69  XIII/R
 1  MVKLTSIAAG VAAIAATASA ----TTTLAQ SDERVNLVEL GVYVSDIRAH LAQYYMFQAA HPTETYPVEV 66  XIV/R
 1  MVKLTSIAAG VAAIAAGV-A AAPATTTLSP SDERVNLVEL GVYVSDIRAH LAQYYLFQAA HPTETYPVEI 69  XV/R
 1  MVKLTSIAAG VAAIAAGASA AA--TTTLSQ SDERVNLVEL GVYVSDIRAH LAEYYSF*AA HPTETYPVEI 68  Pseudo-I/R
 1  MVKLTSIAAG VAAIAATA-S A---TTTLAQ SDERVNLVEL GVYVSDIRAH LAQYYMFQAA HPTETYPVEV 67  Pseudo-II/L

    71         81         91         101        111        121
    |          |          |          |          |          |
67  AEAVFNYGDF TTMLTGIAPD QVTRMITGVP WYSSRLKPAI SSALSKDGIY TIAN                    120 II/R       (YBR301w)
70  AEAVFNYGDF TTMLTGIPAE QVTRVITGVP WYSTRLRPAI SSALSKDGIY TAIPK                   124 III/R      (YCR104w)
67  AEAVFNYGDF TTMLTGIAPD QVTRMITGVP WYSSRLKPAI SSALSKVGIY TIAN                    120 IV/R
67  AEAVFNYGDF TTMLTGIAPD QVTRMITGVP WYSSRLKPAI SSALSKDGIY TIAN                    120 V/L        (YEL049w)
69  AEAVFNYGDF TTMLTGIPAD QVTRVITGVP WYSSRLKPAI SSALSADGIY TIAN                    122 VI/L       (YFL020c)
67  AEAVFNYGDF TTMLTGIAPD QVTRMITGVP WYSSRLKPAI SSALSKDGIY TIAN                    120 VII/L      (NRF120)
67  AEAVFNYGDF TTMLTGIAPD QVTRMITGVP WYSSRLKPAI SSALSKDGIY TITN                    120 VIII/L     (YHL046c)
67  AEAVFNYGDF TTMLTGISPD QVTRMITGVP WYSSRLKPAI SSALSKDGIY TIAN                    120 IX/L       (YIL176c)
70  AEAVFNYGDF TTMLTGIPAE QVTRVITGVP WYSTRLRPAI SSALSKDGIY TAIPK                   124 IX/R       (YIR041w)
67  AEAVFNYGDF TTMLTGISPD QVTRMITGVP WYSSRLKPAI SSALSKDGIY TIAN                    120 X/L        (YJL223c)
70  AEAVFNYGDF TTMLTGIPAE QVTRVITGVP WYSTRLRPAI SSALSKDGIY TIAN                    123 XI/L       (YKL224c)
67  AEAVFNYGDF TTMLTGIAPD QVTRMITGVP WYSTRLKPAI SKALSKDGIY TIAN                    120 XII/L
67  AEAVFNYGDF TTMLTGIAPD QVTRMITGVP WYSSRLKPAI SKALSKDGIY TIAN                    120 XII/R
70  AEAVFNYGDF TTMLTGIPAE QVTRVITGVP WYSTRLRPAI SSALSKDGIY TAIPK                   124 XIII/R
67  AEAVFNYGDF TTMLTGIAPD QVTRMITGVP WYSTRLKPAI SKALSKDGIY TIAN                    120 XIV/R
70  AEAVFNYGDF TTMLTGIPAE QVTRVITGVP WYSTRLRPAI SSALSKDGIY TAIPK                   124 XV/R
69  AEAVFNYGDF TTMLTGIPAD QVTRVITGVP WYSSRLKPAI SSALSVDGIY TIAN                    123 Pseudo-I/R
68  AEAVFNYGDF TTMLTGIAPD QVTRMITGVP WYSSRLKPAI SSALSKDGIY TIAN                    121 Pseudo-II/L
```

Figure 2. Alignment of the seripauperine ORF family. The seripauperine ORF family has at least one member in one subtelomeric region of each sequenced yeast chromosome. Three of them exhibit different features. (i) The chromosome I copy (ORF YAR020c) is interrupted by a STOP codon at position 56 (the sequence Pseudo-I/R corresponds to this ORF). It clearly extends after this STOP codon and another STOP codon is present at the usual position, giving rise to a protein whose length corresponds to that of the other members of the family. This indicates that this pseudogene is probably a recent one. (ii) The copy present on chromosome V is not located in the subtelomeric region of this chromosome, but in the middle of the short arm of the chromosome, at 61 kb from the telomere. (iii) On chromosome XVI, no seripauperine has been described, but the sequence of the telomeric regions of this chromosome are not yet available. Finally, chromosomes IX and XII have one seripauperine at each telomere. A second copy is also present on the left telomere of the chromosome II (Pseudo-II/L), but in a pseudogene state as it is interrupted by a frame shift after codon 18.

ORF NRC299

The ORF NRC299 corresponds to the gene *FZF1*, a zinc finger protein potentially acting as a transcription factor. We have observed that the amino acid T_{235} in NRC299 is a W in the sequence described in Swissprot (accession number P32805). This single amino acid replacement corresponds to the only nucleotide substitution.

ORF NRB486

This ORF corresponds to the gene *HXK2* and encodes hexokinase B. The protein NRB486 is 239 amino acids longer than the protein previously described in PIR (accession number S33656) and completely identical to the one described in Swissprot (accession number P04807).

ORF NRF588

This ORF corresponds to the gene *RTG2* and encodes a protein involved in a pathway of inter-organelle communication between mitochondria, peroxisomes and the nucleus. The protein encoded by the ORF NRF588 is 194 amino acids longer at the C-terminal than the protein described in PIR (accession number B44344); the 389 N-terminal amino acids are identical in both sequences. Many differences are also observed with the nucleic sequence described in EMBL (accession number M97691).

```
     1          11         21         31         41         51         60
     |          |          |          |          |          |          |
   1 MEMLLFLNES YIFHRLRMWS IVLWHSCVFV CAECGNANYR VP-RCLIKPF SVPVTFPFSV 59          NRB76
   1 MEMTLFLNES YIFHRLRMWS IVLWHSCVFV CAECGNANYR VP-RCLIKPF SVPVTFPFSV 59          YIR040c
   1 MEMLLFLNES YIFHRLRMWS TVLWHSCVFV CVECENANYR VP-RCLIKPF SVPVTFPFSV 59          YKL223w
   1 MEMLLFLNES YIFHRFRMWS IVLWHSCVFV CAECGNAYYR GAGGCLEKPF CAPVKFPFSV 60          YBL108w
   1 MEMLLFLNES YIFHRFRMWS IVLWHSCVFV CAECGNANYR GAGVPCKTLL RAPVKFPLSV 60          YCR103c

     61         71         81         91         101        111        120
     |          |          |          |          |          |          |
  60 KKNIRILDLD PRTEAYC*SP YSVCSKRLPC KKYFYLLNSY NIKRVLGVVY CWRLQICEFS 118         NRB76
  60 KKNIRILDLD PRTEAYCLSP YSVCSKRLPC KKYFYLLNSY NIKRVLGVVY C*-------- 110         YIR040c
  60 KKNIRILDLD PRTEAYCLSP YSVCSKRLPC KKYFYLLNSY NIKRVLGVVY C*-------- 110         YKL223w
  61 KKNIRILDLD PRSEAYCLSH HLVCPKRFPC KATSLLL--- -------IPE G*-------- 101         YBL108w
  61 KKNIRILDLD PRSEAYCLSL NSVCFKRLPR KKYFHLLNSY NIKRVLGVVY C*-------- 111         YCR103c

     121        131        141        151
     |          |          |          |
 119 ANALVQLRNC SRRRMVLREV DVQRIVYDKK KKL*                              152         NRB76
 111 ---------- ---------- ---------- ----                              111         YIR040c
 111 ---------- ---------- ---------- ----                              111         YKL223w
 102 ---------- ---------- ---------- ----                              102         YBL108w
 112 ---------- ---------- ---------- ----                              112         YCR103c
```

Figure 3. Alignment of the NRB76 family.

ORF NRE1046

This ORF corresponds to the gene *HFM1* and encodes a putative RNA helicase. The sequence is indentical to the published one, with the exception of the addition of six nucleotides. This results in the insertion of two alanines in position 412 of the *HFM1* gene (accession number S59815).

ORF NRF120

This ORF encodes a small protein of 120 amino acids. This protein is similar to the protein encoded by the gene *PAU1*. It is also similar to the ORFs YAR020c, YBR301w, YCR104w, YEL049w, YFL020c, YHL046c, YIL176c, YIR041w, YJL223c, YKL224c. Other members of this family are present in the telomeric region of chromosome IV right arm (accession number U43834), chromosome XII left and right arms (accession numbers Z47973 and U22383), chromosome XIII right arm (accession number Z454141), chromosome XIV right arm (accession number X86790) and chromosome XV right arm (accession number U23472). This family, previously described as the seripauperine family (Viswanathan *et al.*, 1994), is composed now of at least 17 ORFs. No function is already associated to this familly. All these ORFs exhibit 80% or more amino acid identity (Figure 2).

ORF NRB76

The ORF NRB76 belongs to an ORF family including the ORFs YIR040c, YKL223w, YBL108w and YCR103c. It differs from the other members of the family by a stop codon (UAA) at position 77, after which the amino acid sequence, deduced from the nucleotide sequence, is still similar to the consensus sequence. The next stop codon (UAA) is not in the same position as in the other ORFs of the family but 38 codons downstream (Figure 3). Several truncated ORFs of this family have already been reported (Levesque *et al.*, 1996; Pryde *et al.*, 1995).

ACKNOWLEDGEMENTS

This work was supported by the Commission of European Communities and by the French Groupement de Recherche et d'Etude sur les Génomes. E. C. was financed by a fellowship of the Direction des Recherches Etudes et Techniques. We would like to thank Ed Louis and Hervé Tettelin for personal communications, Denise Menay for the synthesis of all the oligonucleotides used for this work and Christopher Herbert for critical reading of the manuscript.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410.

Bennetzen, J. L. and Hall, B. D. (1982). Codon selection in yeast. *J. Biol. Chem.* **257,** 3026–3031.

Breitwieser, W., Price, C. and Schuster, T. (1993). Identification of a gene encoding a novel zinc finger protein in *Saccharomyces cerevisiae*. *Yeast* **9,** 551–556.

Burns, N., Grimwade, B., Ross, M. P., *et al.* (1994). Large-scale analysis of gene expression, protein localization, and gene disruption in *Saccharomyces cerevisiae*. *Genes Dev.* **8,** 1087–1105.

Dear, S. and Staden, R. (1991). A sequence assembly and editing program for efficient management of large projects. *Nucl. Acids Res.* **19,** 3907–3911.

Dujon, B., Alexandraki, D., Andre, B., *et al.* (1994). Complete DNA sequence of yeast chromosome XI. *Nature* **369,** 371–378.

Feldmann, H., Aigle, M., Aljinovic, G., *et al.* (1994). Complete DNA sequence of yeast chromosome II. *EMBO J.* **13,** 5795–5809.

Johnston, M., Andrews, S., Brinkman, R., *et al.* (1994). Complete nucleotide sequence of *Saccharomyces cerevisiae* chromosome VIII. *Science* **265,** 2077–2082.

Levesque, H., Lepingle, A., Nicaud, J.-M. and Gaillardin, C. (1996). Sequencing of a 9·2 kb telomeric fragment from the right arm of *Saccharomyces cerevisiae* chromosome XIV. *Yeast* **12,** 289–295.

Liao, X. and Butow, R. A. (1993). RTG1 and RTG2: two yeast genes required for a novel path of communication from mitochondria to the nucleus. *Cell* **72,** 61–71.

Louis, E. J. (1995). The chromosome ends of *Saccharomyces cerevisiae. Yeast* **11,** 1553–1573.

Louis, E. J. and Borts, R. H. (1995). A complete set of marked telomeres in *Saccharomyces cerevisiae* for physical mapping and cloning. *Genetics* **139,** 125–136.

Maniatis, T., Fritsh, E. F. and Sambrook, J. (Eds) (1982). *Molecular Cloning. A Laboratory Manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Marck, C. (1988). 'DNA Strider': a 'C' program for the fast analysis of DNA and protein sequences on the Apple Macintosh family of computers. *Nucl. Acids Res.* **16,** 1829–1836.

Mortimer, R. K., Contopoulou, C. R. and King, J. S. (1992). Genetic and physical maps of *Saccharomyces cerevisiae*, Edition 11. *Yeast* **8,** 817–902.

Nikawa, J., Sass, P. and Wigler, M. (1987). Cloning and characterization of the low-affinity cyclic AMP phosphodiesterase gene of *Saccharomyces cerevisiae. Mol. Cell. Biol.* **7,** 3629–3636.

Oliver, S. G., van der Aart, Qj, Agostoni, C. M., *et al.* (1992). The complete DNA sequence of yeast chromosome III. *Nature* **357,** 38–46.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA* **85,** 2444–2448.

Pryde, F. E., Huckle, T. C. and Louis, E. J. (1995). Sequence analysis of the right end of chromosome XV in *Saccharomyces cerevisiae*: an insight into the structural and functional significance of sub-telomeric repeat sequences. *Yeast* **11,** 371–382.

Stachelek, C., Stachelek, J., Swan, J., Botstein, D. and Konigsberg, W. (1986). Identification, cloning and sequence determination of the genes specifying hexokinase A and B from yeast. *Nucl. Acids Res.* **14,** 945–963.

Thierry, A., Gaillon, L., Galibert, F. and Dujon, B. (1995). Construction of a complete genomic library of *Saccharomyces cerevisiae* and physical mapping of chromosome XI at 3·7 kb resolution. *Yeast* **11,** 121–135.

Viswanathan, M., Muthukumar, G., Cong, Y. S. and Lenard, J. (1994). Seripauperins of *Saccharomyces cerevisiae*: a new multigene family encoding serine-poor relatives of serine-rich proteins. *Gene* **148,** 149–153.

Williamson, V. M. and Paquin, C. E. (1987). Homology of *Saccharomyces cerevisiae* ADH4 to an iron-activated alcohol dehydrogenase from *Zymomonas mobilis. Mol. Gen. Genet.* **209,** 374–381.

A comparative study of duplications in bacteria and eukaryotes : the importance of telomeres

Coissac, E and Maillier, E and Netter, P

Les génomes de trois bactéries (*Haemophilus influenzae*, *Mycoplasma genitalium* et *Escherichia coli*) et de deux eucaryotes (*Saccharomyces cerevisiae* et *Caenorapbditis elegans*) sont comparés. La distribution de leurs phases ouvertes de lecture putatives (ORF) a été étudiée et plusieurs conclusions ressortent de cette analyse :

- Tous ces génomes, même le plus petit, présentent une proportion significative (de 7 à 30%) d'ORF dupliquées. Cette proportion est une fonction de la taille du génome et semble ne pas être liée à la division eucaryote/bactérie.
- Certaines de ces ORF forment des familles de vingt membres ou plus.
- Le niveau de similarité observable au sein d'une famille est très variable et leur distribution est différente chez les bactéries et chez les eucaryotes.
- Chez la levure, il y a un lien topologique entre les membres d'une même famille. Les ORF appareillées sont fréquemment dans la même orientation relativement au télomère du bras chromosomique où elles sont localisées et à une distance comparable de celui-ci.

A Comparative Study of Duplications in Bacteria and Eukaryotes: The Importance of Telomeres

Eric Coissac, Evelyne Maillier, and Pierre Netter
Institut Jacques Monod, Paris, France

The genomes of three bacteria (*Haemophilus influenzae*, *Mycoplasma genitalium*, and *Escherichia coli*) and two eukaryotes (*Saccharomyces cerevisiae* and *Caenorhabditis elegans*) were compared. The distribution of their putative open reading frames (ORFs) was studied, and several conclusions were drawn: (1) All of these genomes, even the smallest, exhibit a significant proportion (7%–30%) of duplicated ORFs. This proportion is a function of genome size and appears unrelated to the bacteria/eukaryote division. (2) Some of these ORFs constitute families of up 20 or more members. (3) The levels of sequence similarity within these families are highly variable and their distribution is different among bacteria and eukaryotes. (4) In yeast, there are topological relationships between members of the same family. The paired ORFs are frequently in the same orientation with regard to their respective telomeres and located at comparable distances from them.

Introduction

The role of gene duplications in evolution has been studied for a long time. Up to now, a large part of this work was devoted to the relationship between duplications and the acquisition of new functions. In the neutral theory of molecular evolution (Kimura 1983), the duplication of a gene relaxes the selective constraints exerted on one of the two copies, then allowing the accumulation of mutations leading to the emergence of a new function. Several theoretical models have been developed to estimate the importance of this phenomenon. Ohta (1988) confirmed the importance of constraint relaxation on one of the two copies for the acquisition of a new function. In a more recent model, Walsh (1995) reinforced this idea, considering that, in large populations, a duplicated gene is more likely to give rise to a new functional gene than to a pseudogene.

Several projects to sequence entire genomes have been undertaken recently. Three of them have been completed: two bacterial genomes, those of *Haemophilus influenzae* (Fleischmann et al. 1995) and *Mycoplasma genitalium* (a very small genome of 0.58 Mb) (Fraser et al. 1995), and one eukaryote genome, that of budding yeast *Saccharomyces cerevisiae* (Goffeau et al. 1997). Sequence information is also available for the genomes of *Escherichia coli* and *Caenorhabditis elegans* (vide infra). A sufficient amount of sequence data from various organisms is thus available for investigating the structural organization of these genomes.

The first demonstration of the existence of duplicated regions in yeast came from the work of Lalo et al. (1993), which showed, unambiguously, a large duplication between chromosome III and XIV covering the two centromeric regions. A rough estimation, based on these results and on the previously sequenced yeast chromosomes (Oliver et al. 1992; Dujon et al. 1994; Feldmann et al. 1994), predicted that ca. 40% of the genome was duplicated, this peculiarity possibly being specific to the budding yeast. It was thus tempting to use budding yeast to study genome duplication and to extend this analysis to other organisms, both eukaryotes and bacteria.

In the present work, we concentrated on the open reading frames (ORFs) revealed by these sequencing projects, using them as milestones on the genomes. The sequence similarities among the ORFs revealed that many of them are related and that they can frequently be organized into families. We have shown that a large extent of genome duplication is not only characteristic of yeast, but is a property shared by all the organisms studied (eukaryotes or bacteria), including *M. genitalium*, initially chosen for its small genome thought to possess a minimal gene set (Fraser et al. 1995). Assuming that the topological relationships between the members of these families correspond to the imprinting of physical rearrangements during chromoid or chromosome evolution, we demonstrate the importance of telomeres for duplication mechanisms in yeast.

Materials and Methods

Construction of ORF Databases

All the sequences analyzed were obtained from systematic sequencing projects in order to avoid bias in the sequence selection.

Mycoplasma genitalium

The sequence of the complete *M. genitalium* chromoid (0.58 Mb) revealed 470 ORFs (Fraser et al. 1995).

Haemophilus influenzae

The sequence of the complete *H. influenzae* chromoid (1.83 Mb) revealed 1,680 ORFs (Fleischmann et al. 1995).

Escherichia coli

The *E. coli* sequence analyzed corresponds to a contig of 1.6 Mb localized on the genetic map between positions 4.1 min and 67.4 min. It corresponds to one third of the complete genome. This contig results from the assembly of the following EMBL database entries:

Abbreviation: ORF, open reading frame.

Key words: genome evolution, duplication, telomere, *Saccharomyces cerevisiae*, bacteria.

Address for correspondence and reprints: Pierre Netter, Institut Jacques Monod, 2 place Jussieu–Tour 43, 75251 Paris Cedex 05, France. E-mail: netter@ijm.jussieu.fr.

Mol. Biol. Evol. 14(10):1062–1074. 1997
 ISSN: 0737-4038

U18997, U00039, L10328, L19201, U00006, U28379, and U14003 (Daniels et al. 1992; Blattner et al. 1993; Burland et al. 1993; Plunkett et al. 1993; Sofia et al. 1994). In this region, 1,296 ORFs have been described.

Saccharomyces cerevisiae

The *S. cerevisiae* sequence analyzed corresponds to the complete genome. The set of ORFs was described by the Martinsried Institute for Protein Sequences (MIPS) April 24, 1996 (all of these sequences are available on the MIPS Web server http://speedy.mips.biochem.mpg.de/mips/yeast). This set of data was filtered in order to eliminate all the overlapping ORFs. When two ORFs were shown to overlap, the shorter one was eliminated. The ORFs corresponding to the genes *TYA* and *TYB* of the yeast transposable element Ty have also been eliminated. Five thousand six hundred ninety-seven ORFs covering 11.2 Mb were conserved. The data do not include the rDNA cluster localized on chromosome XII (Oliver et al. 1992; Dujon et al. 1994; Feldmann et al. 1994; Johnston et al. 1994; Bussey et al. 1995; Murakami et al. 1995; Galibert et al. 1996).

Caenorhabditis elegans

There is no large contig available from the *C. elegans* genome project, but a large number of cosmids have been retrieved from the FTP anonymous server of the Sanger Institute (ftp://ftp.sanger.ac.uk/pub/databases/C.elegans sequences). We have used 346 cosmids covering 10.5 Mb (10% of the total genome size) within 1,903 ORFs that have been described.

Construction of Duplication Databases

The duplication databases were constructed in three steps. All of these steps were accomplished automatically on a SPARC 20 SUN workstation. For this purpose, several small programs were written in PERL language.

First Step: ORF Pairing

A BLAST formatted database was constructed from the set of all ORFs. Each ORF was compared to the whole set by the program BLASTP version 1.4 (Altschul et al. 1990) using the similarity matrix PAM250 (Dayhoff, Barker, and Hunt 1983). All of the ORFs exhibiting a P score smaller than 0.01 were conserved. Obviously, this first step was not fully selective, and many ORFs forming nonsignificant pairs were retained.

Second Step: Statistical Validation of the Pairs

To select only paired ORFs having comparable lengths and exhibiting homology over their entire sequences, we used a program developed in our laboratory that is based on a simplified version of the Needleman and Wunsch (1970) algorithm. In this version, only the score matrix was calculated, and the alignment was based on a strict identity between amino acids. The absence of reference to any similarity matrix allowed an important gain in the calculation time. First, the score, X, of the alignment of a given couple of ORFs (seqA and seqB) was calculated. Then, two sets of random peptides (ranA and ranB) were constructed from the composition in amino acids of seqA and seqB. The score, Z, was calculated for each pair of ORFs. This score is defined as $Z = (\bar{X} - X)/\sigma_x$, ($X$ being the score of the alignment between seqA and seqB, $\bar{X}$ the average of the scores of the alignment between seqA and the peptides of ranB and between seqB and the peptides of ranA, σ_X the standard deviation of these scores). The score Z was shown to give an estimation of the statistical significance of the alignment (Needleman and Wunsch 1970). Five parameters may be introduced: the total number N of random peptides ($N/2$ peptides generated from seqA and $N/2$ from seqB), the match bonus M, the mismatch penalty P, the gap penalty G, and the gap extension penalty E. For our set of data, we have established that a score $Z \geq 10$, calculated with the parameters $N = 100$, $M = 5$, $P = 5$, $G = 10$, and $E = 10$, resulted almost exclusively in retention of pairs of ORFs having similar lengths and exhibiting a homology distributed along their entire amino acid sequences. By this type of analysis, an ORF either is unrelated to any other ORF or is related to one or several other ORFs by one or several relationships, here called "links."

Third Step: Calculation of the Levels of Identity and Similarity

After validation of the pairs of ORFs, their levels of identity and similarity were calculated using the program CLUSTAL V and the similarity matrix PAM250 (Higgins 1994). The percentages of identity and similarity were calculated with regard to the whole length of the sequence alignment.

Results

Using the methodology described above, we undertook a systematic study of the genomes of two eukaryotes (*S. cerevisiae* and *C. elegans*) and three bacteria (*H. influenzae, M. genitalium,* and *E. coli*) and addressed several questions:

1. Are there significant differences among the organisms studied, notably between eukaryotes and bacteria?
2. How many ORFs can be grouped into families of proteins exhibiting significant levels of similarity?
3. What are the sizes of these families, if they exist (duplications, "triplications," or more), and do they constitute coherent sets?
4. What are the levels of similarity within these families?
5. Can one detect groups of ORFs duplicated simultaneously?
6. How are members of the same family distributed along the genome?

All Organisms Tested Exhibit a Significant Level of Genomic Duplication

The algorithms described in *Materials and Methods* allowed us to classify the ORFs of each organism into families composed of one, two, three, or more members. The results are presented in table 1.

Table 1
Comparison of the Sizes of the ORF Families in Various Organisms

	Number of ORFs[a]				
Number of Links	*S. cerevisiae* 11.2 Mb 5,697	*C. elegans* 10.5 Mb 1,903	*H. influenzae* 1.85 Mb 1,680	*E. coli* 1.6 Mb 1,296	*M. genitalium* 0.58 Mb 470
0[b]	3,962	1,571	1,493	1,097	438
1	1,027	176	138	142	28
2	255	48	23	33	4
3	126	21	13	14	
4	84	19	7	8	
5	64	18	4	1	
6	37	5	1	1	
7	28	2	1		
8	8	11			
9	8	7			
10	5	5			
11	16	5			
12	0	10			
13	0	2			
14	4	1			
15	3	2			
16	14				
17	2				
18	16				
19	23				
20	7				
21	5				
22	3				

[a] The number of ORFs for which 0, 1, 2, . . . , *n* links (see *Materials and Methods*) were established with other ORFs in the same genome is indicated. The total or partial (*C. elegans* and *E. coli*) genome size and the number of ORFs analyzed for each species are indicated under its name.

[b] The absence of link (0) correspond to ORFs which, following our criteria, have not been duplicated.

At first sight, one would expect that the number of ORFs in a given class would be a multiple of the family size (for example, 5, 10, 15, . . . , ORFs in the family of size 5). This is not always the case, and this apparent discrepancy has to be explained. In fact, our methodology consists of establishing relations (or "links") between a given ORF used as a query and all other ORFs. It means that a single pair of ORFs generates two links, A→B and B→A. The number and the nature of these links create a "family." For example, if two ORFs (B and C) exhibit, by the criteria described above, a significant similarity with ORF A (see fig. 1, links A→C and A→B), they can, in turn be used as queries. Two kinds of results can then be obtained: (1) The latter tests do not reveal new ORFs (fig. 1*a*). Therefore, the family grouping the ORFs A, B, and C is closed and defined by the links A→B, B→A, A→C, C→A, B→C, and C→B. (2) New ORFs are revealed (like D, fig. 1*b*). We call the group formed by A, B, C, and D a "superfamily." In this example, when D is in turn used as a query, there is no new ORF revealed, and therefore the superfamily ABCD is closed and restricted to the links A→B, B→A, A→C, C→A, B→C, C→B, C→D, and D→C. It should be noted that the result depends on the ORF used as query: in the example of the figure 1*b*, D reveals a family of only two members (D+C); A or B,

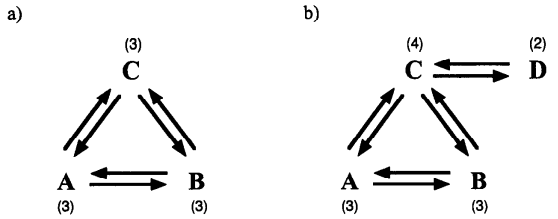


FIG. 1.—Families and superfamilies. A, B, C, and D are hypothetical ORFs. Each arrow represents a "link" (see text). *a*, An example of a group of three ORFs, each one (A, B, or C) belonging to the same closed family of three members. *b*, An example of a group of four members, forming a closed superfamily in which some ORFs belong to groups of two (D) three (A, B) or four (C) members.

a family of three members (A+B+C); and C, a family of four members (A+B+C+D).

The determination of the number of superfamilies allowed us to estimate the consistency of our classification. For the reasons previously described, it is apparent that, in the superfamilies harboring many ORFs, not all ORFs reveal a number of links corresponding to the number of members (table 1). However, all the ORFs presented in this table are grouped in closed families or superfamilies which are not interrelated step by step.

Comparison Among the Different Organisms

The definition of families being established, the first conclusion which can be drawn from the data presented in table 1 is that all the organisms tested exhibit a significant proportion of duplicated or triplicated (or more) ORFs, even when the genome is as small as it is in the case of *M. genitalium* (0.58 Mb). The distribution of the sizes of ORF family indicates that the number of duplicated ORFs increases with the size of the genome. However, some analyses were done on complete genomes (*S. cerevisiae, H. influenzae, M. genitalium*), whereas some others are based on only partial data (*C. elegans, E. coli*), and this difference obviously introduces a bias in the results. In order to circumvent this problem and compare the levels of duplication independently of the genome size, we have generated, for each species, several subsets of an increasing number of ORFs. In each one of these subsets, the number of duplicated ORFs is determined by the same method as for the complete data set. The features of the curve relating the numbers of duplicated ORFs as a function of the size of the subset can be used to compare the levels of duplication independently of the genome size (fig. 2). It appears that the differences between organisms are very small. If we consider the slope of the curve as an index of duplication independent of the genome size, it is impossible to distinguish eukaryotes from bacteria. Increasing indices of duplication gives the order: *H. influenzae, S. cerevisiae, M. genitalium, C. elegans,* and *E. coli.* Even though these curves do not strictly reflect a single distribution, an average curve could provide a rough estimate of the level of duplication of any given genome based only on its size and independent of its origin, eukaryotic or bacterial. Similar results have been obtained on *E. coli* and *H. influenzae* (Labedan and

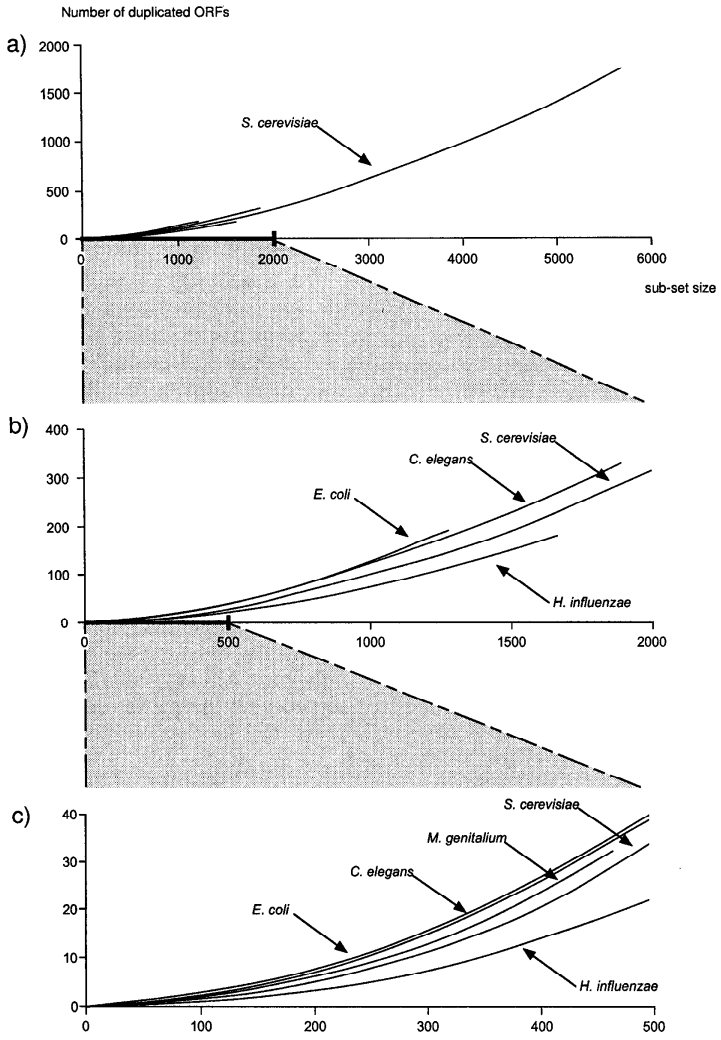


FIG. 2.—Comparison of the levels of duplication in various species. For each species, a large number of ORF subsets have been randomly generated from the total ORF set of the organism. The sizes of these subsets vary from five to the size of the total set by steps of five. For each size, two subsets have been generated. In each one of these subsets, the number of duplicated ORFs was established. The resulting curves for all tested organisms are plotted, showing the variation of the number of duplications as a function of the subset size. To take into account the differences between the genome sizes, the same plot has been enlarged to different scales: (*a*) 0–6,000 ORFs, (*b*) 0–2,000 ORFs, and (*c*) 0–500 ORFs.

Riley 1995; Tatusov et al. 1996) in spite of the fact that the study was focused on the functional implication of the duplications.

Using the preceding criteria, it seems that there is no qualitative difference between the eukaryotes and the bacteria. However, the comparison of these two groups shows quantitative differences which have to be analyzed further. If one considers the distribution of the sizes of the various families (table 1), it seems that the number of large families is greater for eukaryotes than for bacteria. To quantify this observation, we divided the families into three classes (small families of two members, medium families of three or four members, and large families) and tested the fit of the resulting distributions to two mutually exclusive hypotheses. The first hypothesis assumes a unique distribution common to eukaryotes and bacteria. The second hypothesis is based on distinct distributions for the two phyla. The results are presented in table 2 and show that there are at least two types of distribution, one for eukaryotes and one for eubacteria. The difference between these two distributions is essentially based on the number of large families, which is greater in eukaryotes than in bacteria.

Distribution of the Percentage of Identity

Another point of interest is the analysis of the level of identity observed among members of the same family. Figure 3 shows the distribution of the percentage of amino acid identity calculated between such paired ORFs. The filter applied to define these pairs is based

Table 2
Statistical Analysis of the Distribution of Family Sizes

	EUKARYOTES		BACTERIA	
DISTRIBUTION	*S. cerevisiae*	*C. elegans*	*E. coli*	*H. influenzae*
General		**59.4**		
		6 degrees		
Bacterial	581	268	**2.07**	**2.04**
	2 degrees	2 degrees	**2 degrees**	**2 degrees**
Eukaryotic...	**1.53**	8.00	28.6	24.3
	2 degrees	2 degrees	2 degrees	2 degrees

NOTES.—To test if the eukaryotes exhibit significantly larger families than the bacteria, we divided the duplicated ORFs into three classes: a poorly duplicated ORF class (ORFs with one link), a highly duplicated ORF class (more than three links), and an intermediate one (two or three links). A classical conformity test (χ^2) was performed for the three possible distribution types: the "general" distribution model, where the distribution of family sizes is the same for all the organisms (for this first hypothesis, a global χ^2 was calculated); the "bacterial" distribution model, which is defined by the addition of *E. coli* and *H. influenzae* ORFs; and the "eukaryotic" distribution model, which is defined by the addition of *S. cerevisiae* and *C. elegans* ORFs. For the last two models, the χ^2 was calculated for each species. This table summarizes the results for all thse tests by indicating the χ^2 values and the number of degrees of freedom. Values in bold correspond to nonrejected hypotheses, with a risk of 5%.

on the *Z* score (see *Materials and Methods*). We used a minimal *Z* score of 10, which corresponds here to a threshold of amino acid identity between 15% and 25%. Therefore, it is not surprising that features common to all species are observed, particularly the decreasing number of ORF pairs with an identity below 30% and their absence under 15%.

All the species exhibit distributions situated primarily between 15% and 45% with a mean value around 25%–30%. In addition, an excess of pairs with a high level of identity is present in eukaryotes, especially in *S. cerevisiae*, in which one fourth of duplicated ORFs present a level of identity over 65%.

Topology of Duplicated ORFs

We analyzed the topological relationships between the two members of pairs of duplicated ORFs (characterized as described in the previous paragraph). Because this analysis preferentially requires the comparison of complete chromosomes, this question is more easily addressed when the complete sequence of the organism is available, i.e., for *S. cerevisiae, H. influenzae,* and *M. genitalium.* The method used consists simply of drawing a two-dimensional graph in which each axis represents a chromosome. A pair of duplicated ORFs is revealed by a dot located at the intersection of the positions of each ORF on its chromosome.

For the bacteria, the two axes obviously represent the same unique chromosome (Fig. 4). We have shown the results only for the two complete genomes of *H. influenzae* and *M. genitalium*, but a similar result is observed with the partial data from *E. coli.* It clearly appears that the dots are dispersed almost randomly over

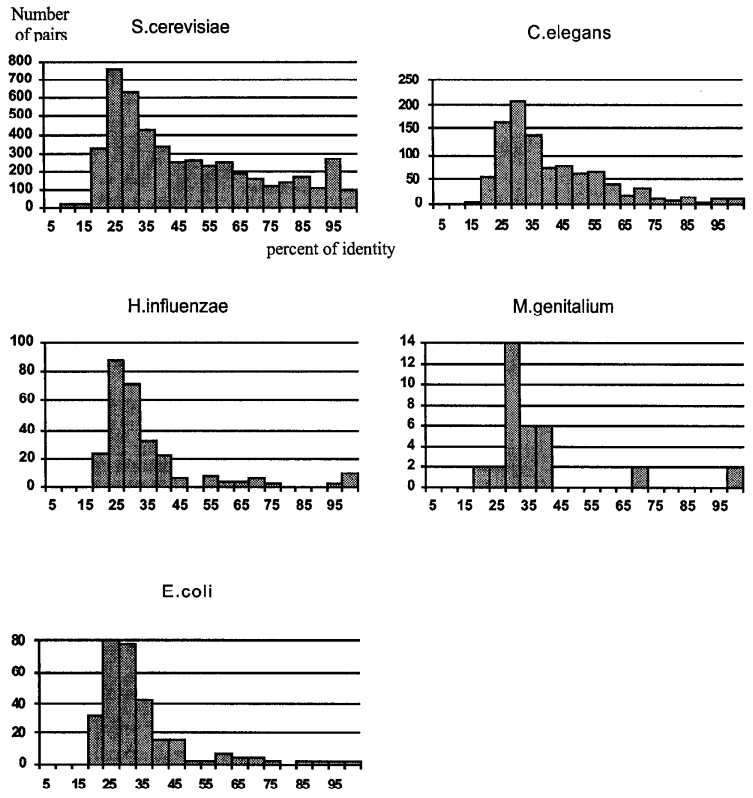


FIG. 3.—Levels of identity among duplications. These histograms show, for the five organisms analyzed, the distribution of the level of amino acid sequence identity between duplicated ORFs (i.e., ORFs exhibiting at least one link with another ORF in the genome).

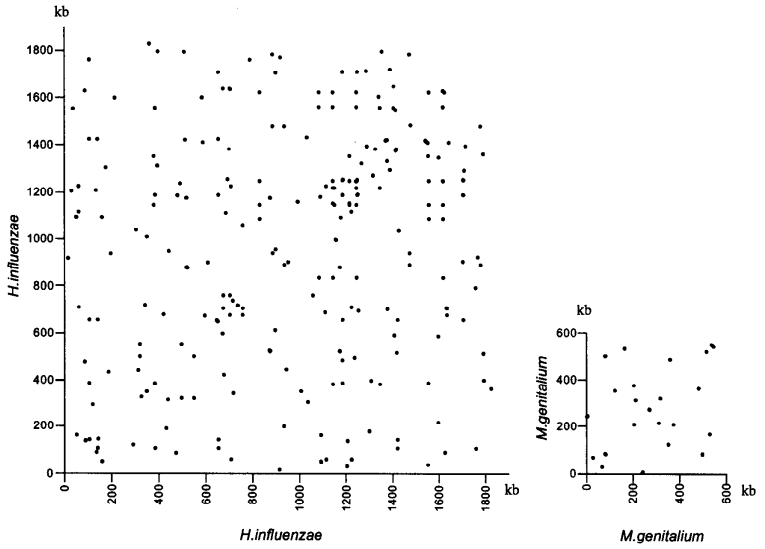


FIG. 4.—Topological organization of duplications in bacteria. Each plot represents, at the same scale, the complete genome of a bacterium (*H. influenzae* or *M. genitalium*) compared with itself. The axes correspond to linear maps of the genomes. For each duplication, a dot is plotted at the intersection of the positions of the two members. Dots seem to be distributed randomly all over the graph. Dots on the diagonal correspond to tandem duplications. The density of dots is equivalent on the two graphs, suggesting a similar level of duplication in the two species *H. influenzae* and *M. genitalium*.

the whole graph area. The points on (or near) the diagonal correspond, in these bacterial species, to tandemly duplicated ORFs. With these exceptions, the localization of duplications does not seem to follow any clear topological rules. In particular, the duplication of several adjacent ORFs cannot be detected.

In the same analysis performed on *S. cerevisiae,* the two axes of the graphs correspond to any one of the 16 chromosomes. Figure 5 shows examples obtained for some combinations, namely the comparison of chromosomes X and XI, III and XIV, and VIII and XIV.

It is apparent from these examples that one can find several groups of dots which are roughly aligned (gray boxes), this result is indicative of the coduplication of several ORFs in a row. We have undertaken a systematic search for such coduplications based on two selective criteria: (1) at least four ORFs must be clearly aligned on the graph, and (2) these ORFs must maintain their relative orientation on each chromosome.

A schematic example is given in figure 6, corresponding to a coduplication between chromosomes II and V. One can notice that the pairs of duplicated ORFs are frequently interspersed with nonduplicated ones (forming "loops"). Assuming the initial event must be the coduplication of several adjacent ORFs, this dislocation could be, a priori, the consequence of three nonexclusive phenomena: (1) the divergence of initially duplicated ORFs to such an extent that they are no longer detected as duplicated by our criteria, (2) the differential deletion of one member of a previously existing duplicated ORF pair, or (3) the insertion of previously absent sequences.

The analysis of the complete yeast genome revealed 46 groups of four or more cooriented duplica-

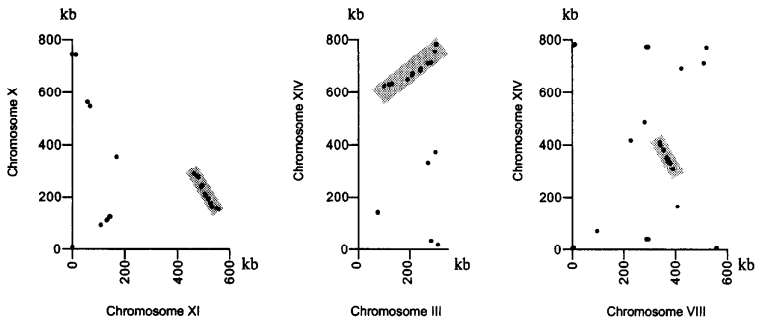


FIG. 5.—Topological organization of duplications in *S. cerevisiae*. Same representation as in figure 4. On these graphs, each axis corresponds to a given chromosome. Three pairs of chromosomes are shown: X and XI, III and XIV, and VIII and XIV. Some dots are aligned (gray boxes), this organization being indicative of the coduplication of several ORFs in a row. The duplication between chromosomes III and XIV includes all of the right arm of chromosome III.

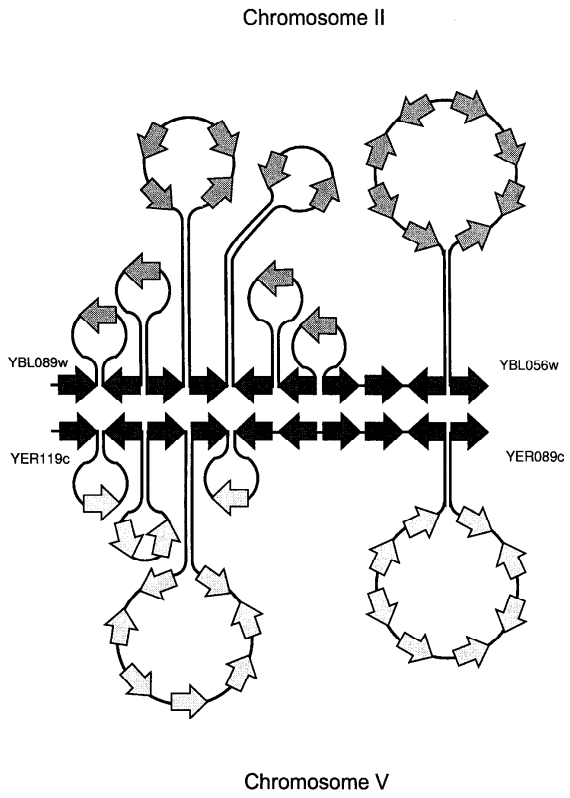


FIG. 6.—Detailed structure of the duplication between *S. cerevisiae* chromosomes II and V. This figure shows the duplication between chromosomes II (upper part, ORFs YBL056w to YBL089w) and V (lower part, ORFs YER089c to YER119c). Each arrow corresponds to an ORF and indicates its orientation. The dark gray arrows correspond to the paired ORFs forming the central core of the duplicated blocks. It is important to stress that these ORFs are colinear and in the same orientation on the two chromosomes. This set of ORFs must be part of the original duplication. The ORFs localized in "loops" (light gray arrows) may be the consequence of insertion, deletion, or divergence during the evolution of the two copies. The percentage of identity between paired ORFs is very variable from one pair to another in the same chromosomal duplication, thus being useless for dating of the initial duplication event.

tions (table 3). These duplicated regions are found on all 16 chromosomes, extend over 5.7 Mb (half of the whole genome), and include 2,726 ORFs (among which only 670 [= 335 × 2] are really duplicated). Furthermore, when several independent coduplications affect the same chromosome, it is possible to have an overview of their topological distribution. An example is given in figure 7, which shows that one can identify blocks of duplications all along chromosome IV, over roughly one half of its total length. Similar studies have been performed on DNA sequences by the MIPS group (Mewes et al. 1997) and on protein sequences by Wolfe and Shields (1997). These two studies also show large duplicated blocks, the latter study proposes the hypothesis of an ancient duplication of the entire yeast genome. As mentioned above in the analysis of the whole genome, 392 ORFs out of the 724 present on the chromosome are involved in the duplications, but of these, only 93 are duplicated. This apparently low proportion will be discussed later.

The total number of pairs of ORFs (table 2, right column) in the 46 blocks fulfilling the previously described criteria of selection is 335 (i.e., 670 ORFs and 335 × 2 links). This number has to be compared to the total number of (1,027 + 255 + ... + 3) = 1,935 ORFs, members of a family of two or more members and {(1,027 × 1) + (255 × 2) + (126 × 3) + ... + (3 × 22)} = 4,746 links corresponding to 2,373 pairs (see table 1). The difference, due to the stringency of our initial screening, includes duplicated ORFs for which the coduplication of neighbor sequences cannot be established, or involves less than four pairs (vide supra). These are then considered as isolated ORFs, but it remains that, on their respective chromosomes, the topological relationships between the members of these 2,038 (2,373 − 335) pairs can be analyzed.

For this purpose, we have investigated the frequency of coorientation of these ORFs on each chromosome arm with regard to their respective telomeres. In order to avoid artifacts due to some genomic peculiarity of yeast, we have compared two genomes, the first one being the "natural" genome of *S. cerevisiae,* the second one, used as a control, being a genome (in fact, two independently generated ones) in which the positions of the ORFs have been conserved but their orientations with regard to the telomere have been randomly redistributed. The results are unambiguous: The "control" genome gives, as expected, a value of coorientation of the ORFs of 50%, consistent with its random original construction. The "natural" genome exhibits a significantly different organization due to a strong bias in favor of the coorientation of duplicated ORFs on their respective chromosomal arms. Among the 2,038 pairs, it appears that 1,285 (63%) exhibit ORFs in the same orientation, this value being incompatible with a random distribution (χ^2 = 138.8). This result, which will be commented on more detail in the *Discussion,* demonstrates that whatever the mechanism of duplication, the locations of the concerned ORFs on the chromosome are important.

It should also be noted that some pairs of ORFs are located on the same chromosome. Several observations can be drawn from the study of these pairs: (1) The number of ORFs in these pairs (79) is very close to the number (76.5) predicted in the case of a random distribution of the ORFs on the 16 chromosomes. (2) Fifty-six pairs (71%) exhibit cooriented ORFs, 33 (out of 47) on the same chromosomal arm, 23 (out of 32) on two different arms. (3) Among the 47 ORFs located on the same arm, 25 are separated by less than 10 kb, 23 of these 25 being cooriented and the remaining two being in opposite orientation and separated by 828 nucleotides.

Discussion

We have undertaken a systematic analysis of the duplications in several organisms, bacteria and eukaryotes. These duplications are the result of primitive rearrangements affecting the nucleic acid, followed by

Table 3
List of Coduplications in *S. cerevisiae*

Chromosomes[a]		Orientation[b]	Positions of Duplicated Regions[c]		ORFs on Chromosome[d]		Number of Pairs[e]
			Chromosome A	Chromosome B			
A	B		Begin End	Begin End	A	B	
I	VIII	D	203380–228928	525387–555959	12	7	7
I	XV	I	1810–108542	916020–1083200	47	67	10
II	IV.....	D	238901–394810	450263–564897	75	53	13
II	V.....	I	49413–115123	333934–398827	28	30	10
II	VII	I	645504–662203	385194–401284	9	9	4
II	VII	D	800476–811432	1070289–1082725	5	6	5
II	XVI	D	579106–607090	352861–407531	17	26	6
II	XVI	D	449621–483320	673746–703966	16	21	4
III	IV.....	I	42140–81566	1452626–1475568	22	13	7
III	X.....	I	128073–195916	60843–109453	32	29	4
III	XIV	D	120553–306895	629622–782272	91	73	11
IV	V.....	D	581950–654462	448552–544397	33	38	6
IV	VIII	D	1062136–1181980	218230–330059	53	52	12
IV	X.....	I	1802–18566	727097–743691	6	7	6
IV	XII	I	322225–368255	871694–931750	24	31	4
IV	XII	D	669056–832509	294094–500271	73	87	11
IV	XIII	I	298419–314381	546124–565858	9	14	5
IV	XIII	I	1325072–1355913	220138–241536	16	13	7
IV	XIV	D	44066–76546	252058–295506	14	23	4
IV	XV	D	376800–414925	200369–224173	16	13	4
IV	XV	D	994280–1015737	392417–417680	13	14	4
IV	XV	D	1204155–1274047	551111–624369	36	30	5
IV	XVI	I	1456721–1521966	283–42866	33	18	5
V	IX.....	I	262054–322077	221081–272773	29	28	9
V	X.....	I	65388–115797	486277–531051	25	19	5
VI	X.....	I	5066–14763	727097–744953	7	8	4
VI	XIV	D	2–15431	374–17248	10	8	8
VII	VIII	D	937119–1006095	394458–472421	34	35	7
VII	XII	D	502933–631264	779213–866778	62	40	11
VII	XVI	I	344792–374079	168090–194141	17	15	5
VII	XVI	D	668183–777499	747303–859737	51	42	12
VIII	XI.....	D	70272–108112	362265–446365	19	40	5
VIII	XIV	I	340114–391694	308959–409905	34	53	9
IX	X.....	D	486–21217	469–28593	7	10	5
IX	XI.....	I	48091-69525	617636–634812	9	5	4
IX	XIV	D	89233–202040	478568–578769	54	47	11
X	XI.....	D	90783–122645	109274–145380	18	16	5
X	XI.....	I	151414–290470	463600–558590	55	46	15
XI	XIII	I	179672–357489	303238–502733	91	99	16
XII	XIII	I	1002550–1035754	115737–164089	17	21	8
XII	XV	D	681185–730297	657129–692026	27	17	4
XIII	XV	D	715640–761887	844989–911432	24	34	7
XIII	XVI	I	616565–664280	64980–128087	23	31	5
XIV	XV	D	56449–106693	78354–115807	23	23	7
XIV	XV	I	419016–465943	483221–530242	20	22	8
XV	XVI	I	722908–783292	278397–326263	32	25	11
Total			Number of duplicated blocks: 46		1,368	1,358	335

[a] This table summarizes the 46 coduplications involving at least 4 ORFs in each block. Chromosomes A and B are the two chromosomes involved in the duplication.

[b] The relative orientations of the two copies are presented (the chromosomes are oriented, by convention, from the left to the right telomere). The duplications are direct (D) if, after alignment of the ORFs, the two chromosomes are in the same orientation and inverted (I) in the other cases.

[c] Coordinates of the limits of the duplicated blocks are given in base pairs from the left telomere of each chromosome. Underlined coordinates correspond to coduplicated regions which are both very close to their respective telomeres (less than 30 kb). For comparison, the sizes of the 16 chromosomes are, respectively, I: 230,195; II: 813,137; III: 315,354; IV: 1,522,191; V: 574,860; VI: 270,148; VII: 1,090,936; VIII: 562,638; IX: 439,885; X: 745,443; XI: 666,448; XII: 1,078,171; XIII: 924,430; XIV: 784,328; XV: 1,091,282; and XVI: 948,061 nucleotides.

[d] For chromosomes A and B, respectively, the number of ORFs included within these limits.

[e] The numbers of paired ORFs between the two blocks.

their maintenance in the genome. It should be recalled that all the genomes tested have very dense genetic information (one gene every 1.1, 1.2, 1.1, and 2.1 kb for *H. influenzae, M. genitalium, E. coli,* and *S. cerevisiae,* respectively) and therefore display short intergenic regions. In this context, we have analyzed ORFs as milestones in the sequence. The use of amino acid sequences facilitated our analysis. We have not considered the role of functional constraints in the process of divergence. While these constraints may influence the level of divergence of individual ORF pairs, they should not affect the topology of the duplications.

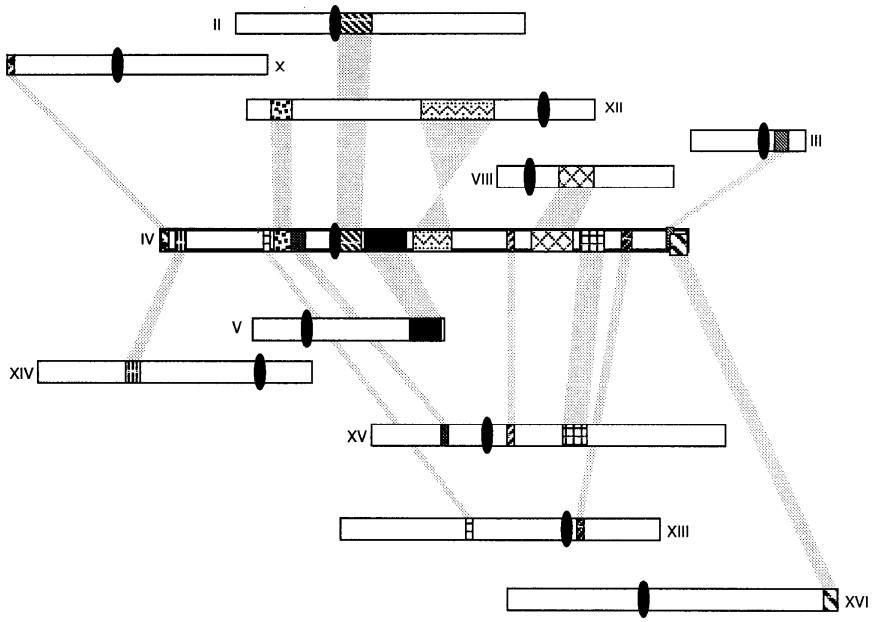


FIG. 7.—Mosaic organization of chromosome IV of *S. cerevisiae*. Chromosome IV is related to 10 other chromosomes by blocks of duplications (see table 2). This figure shows a graphic representation of these relations. The name of each chromosome is indicated, and the centromeres are represented by black ovals. It should be stressed that the duplicated blocks are in the same orientation on their respective chromosomal arms.

Duplications are Widely Represented in All Tested Organisms

Several conclusions can be drawn from the comparison of the various genomes studied (keeping in mind the restriction that our analysis concerns only a limited sample of species):

1. Duplications of genes have taken place in both phyla studied (eukaryotes and bacteria), and therefore they are neither the consequence of the physical organization of the genome (circular and unique chromoid *vs.* linear and multiple chromosomes) nor correlated with their replication processes (one replicon vs. mitosis and meiosis).
2. Duplications are not a prerogative of large genomes, although the number of duplicated ORFs is more or less a function of genome size (see fig. 2).
3. Among the eukaryotes studied (*S. cerevisiae* and *C. elegans*), there is no apparent difference between unicellular and multicellular organization.
4. The main distinction between eukaryotes and bacteria concerns the proportion of ORFs exhibiting a high level of similarity, which is significantly increased for the two eukaryotes (fig. 3). At the present stage of our analysis, it seems difficult to suggest any convincing interpretation for this observation, whatever the model: different rates for the molecular clocks in both phyla, a more ancient origin of duplications in bacteria, more active homeologous recombination and homogenization in eukaryotes, etc.

In Yeast, the Positions of the Duplicated ORFs and Their Orientations are Correlated with the Telomeres

The mechanisms by which duplications are built are still a matter of speculation. Furthermore, if recombination and/or conversion must take place at some time during the initial process, what we observe now is also the result of subsequent events leading to chromosome shuffling. However, although it would be naive to hope to find clear-cut and untouched consequences of the initial events, we think that it is still possible to identify their traces in the present genomes.

The topology of each chromosome is essentially defined by the landmarks which are its centromere and telomeres. We asked a simple question: Is there any indication that these structures have played some role in the initial duplicative process, and can one find a significant trace of this role?

In order to answer this question, we systematically analyzed the respective positions of the ORFs engaged in the 4,746 links previously described. The relative distances either to their telomeres or to their centromeres were compared, and a graphic representation is shown in figure 8. A control genome was also built by a randomly generated relocalization of the paired ORFs anywhere on the 16 chromosomes. It clearly appears that a correlation (correlation index $r = 0.71$) is observed if one takes into account the distances to the telomere (fig. 8*a*), whereas the correlation is weak ($r = 0.19$) for the distances to the centromere (fig. 8*b*). In fact, visual inspection of the graph in figure 8*a* reveals a higher den-

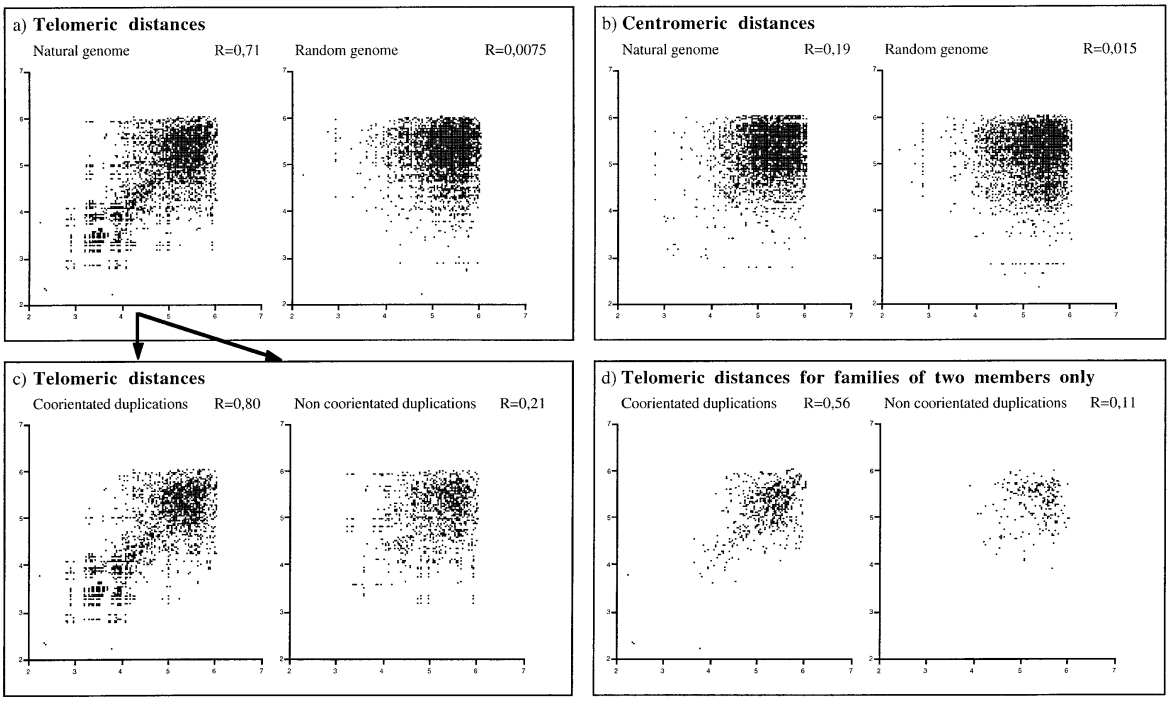


FIG. 8.—Correlation between the distances of paired ORFs to their respective telomeres in *S. cerevisiae*. See the *Discussion*. All plots and correlation indexes are established on the logarithmic values of distances.

1072 Coissac et al.

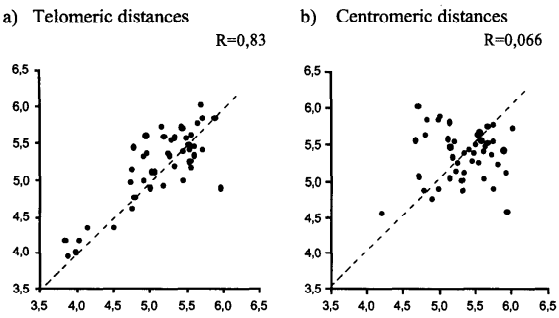


FIG. 9.—Correlation between the localizations of the 46 coduplicated blocks in *S. cerevisiae*. The central coordinates of each of the 2 × 46 groups were used to test the hypothesis of a correlation with their distance to the telomere or the centromere. All plots and correlation indexes (R) are established on the logarithmic values of distances.

sity of dots along the diagonal. In both cases, the random genome gave very low correlation indices (r = 0.0075 and r = 0.015), this background being the consequence of topological constraints associated with the various lengths of the 16 chromosomes.

The correlation with telomeric distances can be further studied if one subdivides the pairs of ORFs into two groups (fig. 8*c*). One can see that the global correlation (r = 0.71) was mainly due to the subgroup of cooriented ORFs (r = 0.80), whereas the contribution of the noncoorientated ones is much lower (r = 0.21).

However, there could be a bias in favor of the superfamilies composed of many members. For example, a family of 20 members corresponds to 20 × 20 = 400 links and therefore has an abnormally high weight among the 4,746 links. This bias can be visualized on the lower left part of the diagonal (fig. 8*a* and *c*). To eliminate this problem, we performed the same type of analysis on the links issued from small families only. Figure 8*d* shows the results obtained with the 1,027 links restricted to families of two members. There is still a good correlation with telomeric distances (r = 0.48) which is mainly due to the subgroup of 73.5% of cooriented ORFs (r = 0.56). The statistical robustnesses of the correlation indices were tested and are highly significant (data not shown).

Finally, we analyzed whether the same phenomenon is apparent for the 46 groups of four or more cooriented ORFs (see table 2). Figure 9 clearly shows a highly significant telomeric correlation indice (r = 0.83). This result has to be compared with the relative orientations of the groups of paired ORFs on their respective chromosomes. In 44 out of 46 cases (except for duplication between chromosomes II and XVI and between chromosomes VII and XVI), the groups are identically oriented with regard to the telomere, and the same is true for 626 ORFs among 670 (93.4%). Furthermore, these apparent exceptions, concerning duplications between chromosomes II, VII, and XVI, can easily be reintroduced in the general scheme if one imagines a subsequent inversion of a single fragment of chromosome XVI (fig. 10).

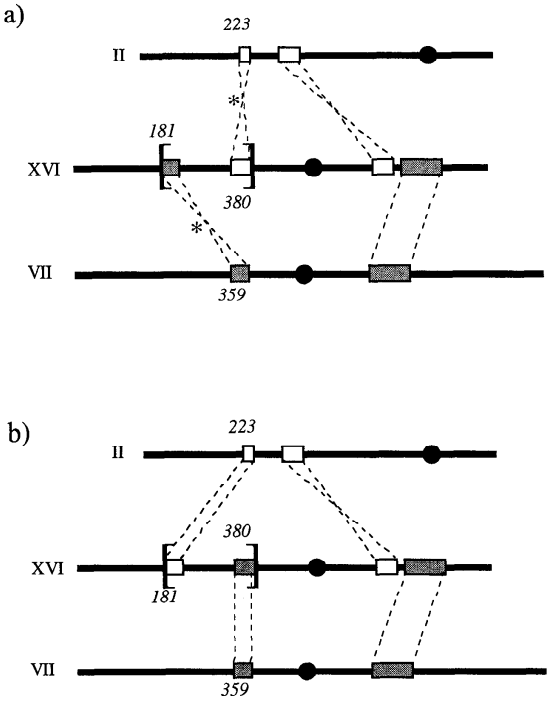


FIG. 10.—A possible mechanism to explain inverted blocks between chromosomes VII and XVI in *S. cerevisiae*. Topological organization of coduplications between chromosomes II and XVI (white boxes) and VII and XVI (gray boxes). Black circles represent centromeres. The distances of some duplications to the telomere are indicated in italics (kb). *a*, The coduplications between chromosomes II, VII, and XVI in the present status: Two pairs of coduplicated blocks are in inverted orientation (shown by asterisks) with regard to their respective telomeres. *b*, A simple inversion of a segment of chromosome XVI (indicated by brackets) restores both the coorientation and the distances of the coduplicated blocks to their telomeres. We propose that this organization could be the ancient one of chromosome XVI.

Conclusion and Perspectives

Taken together, these data unambiguously demonstrate the importance of the telomeres in at least some of the processes leading to duplication events. Little is known about the nuclear organization except during mitotic or meiotic divisions. *In situ* immunofluorescence experiments have shown that the 64 telomeres of a yeast diploid cell are clustered into 5–8 foci (for a review, see Louis 1995). This type of organization brings together nonhomologous chromatids and could allow chromosomal fragments located at comparable distances from their respective telomeres to preferentially interact. Whatever the mechanisms, these events have involved fragments that are widely variable in size and have been frequently followed by subsequent reshuffling. The fragments observed in the present genome represent the subset which has been conserved by selection. It is still unclear if these events have taken place in haploid and/or diploid cells, or neither if particular phases of the

cellular life are more favorable for these exchanges. Furthermore, this telomere-related process is not inconsistent with other mechanisms frequently evoked, such as illegitimate recombination or slippage during replication.

It should be remembered that the importance of the telomere has been described in maize by McClintock (1951) in the Breakage–Fusion–Bridge (BFB) cycle mechanism. This model, leading to the accumulation of inverted repeats, was more recently confirmed by fluorescent in situ hybridization (FISH) in the case of adenylate deaminase 2 (AMPD2) gene amplification in Chinese hamster cells (Toledo *et al.* 1992). However, the resulting organization corresponds to head-to-head duplication, and therefore this mechanism seems not to account for the results presented here.

We are currently developing experimental models, particularly in yeast, to identify the main rules governing this telomeric-related exchange. Furthermore, this process should be observable in all eukaryotes, as it relates to the chromosomal organization of the nucleus. We can therefore predict that it should be detected in the next complete eukaryotic genome to be published (*C. elegans, A. thaliana*), in spite of the existence in these organisms of introns making studies more difficult.

Acknowledgments

We would like to thank Christophe Cullin and Jean-Marie Rouillard for their helpful discussions and Lon Aggerbeck, Christopher J. Herbert, and François Michel for critical reading of the manuscript. E.C., E.M., and P.N. are members of the Université Pierre et Marie Curie, Paris.

LITERATURE CITED

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS, and D. J. LIPMAN. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

BLATTNER, F. R., V. BURLAND, G. PLUNKETT, H. J. SOFIA, and D. L. DANIELS. 1993. Analysis of the Escherichia coli genome. IV. DNA sequence of the region from 89.2 to 92.8 minutes. Nucleic Acids Res. **21**:5408–5417.

BURLAND, V., G. PLUNKETT, D. L. DANIELS, and F. R. BLATTNER. 1993. DNA sequence and analysis of 136 kilobases of the Escherichia coli genome: organizational symmetry around the origin of replication. Genomics **16**:551–561.

BURLAND, V., G. PLUNKETT, H. J. SOFIA, D. L. DANIELS, and F. R. BLATTNER. 1995. Analysis of the Escherichia coli genome VI: DNA sequence of the region from 92.8 through 100 minutes. Nucleic Acids Res. **23**:2105–2119.

BUSSEY, H., D. B. KABACK, W. ZHONG et al. (13 co-authors). 1995. The nucleotide sequence of chromosome I from Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. USA **92**: 3809–3813.

DANIELS, D. L., G. PLUNKETT, V. BURLAND, and F. R. BLATTNER. 1992. Analysis of the Escherichia coli genome: DNA sequence of the region from 84.5 to 86.5 minutes. Science **257**:771–778.

DAYHOFF, M. O., W. C. BARKER, and L. T. HUNT. 1983. Establishing homologies in protein sequences. Methods Enzymol. **91**:524–545.

DUJON, B., D. ALEXANDRAKI, B. ANDRE et al. (108 co-authors). 1994. Complete DNA sequence of yeast chromosome XI. Nature **369**:371–378.

FELDMANN, H., M. AIGLE, G. ALJINOVIC et al. (96 co-authors). 1994. Complete DNA sequence of yeast chromosome II. EMBO J. **13**:5795–5809.

FLEISCHMANN, R. D., M. D. ADAMS, O. WHITE (11 co-authors). 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science **269**:496–512.

FRASER, C. M., J. D. GOCAYNE, O. WHITE et al. (29 co-authors). 1995. The minimal gene complement of Mycoplasma genitalium. Science **270**:397–403.

GALIBERT, F., D. ALEXANDRAKI, A. BAUR et al. (57 co-authors). 1996. Complete nucleotide sequence of Saccharomyces cerevisiae chromosome X. EMBO J. **15**:2031–2049.

GOFFEAU, A., R. AERT, M. L. AGOSTINI-CARBONE et al. (633 co-authors). 1997. Overview of the yeast genome. Nature **387**(Suppl.):1–105.

HIGGINS, D. G. 1994. CLUSTAL V: multiple alignment of DNA and protein sequences. Methods Mol. Biol. **25**:307–318.

JOHNSTON, M., S. ANDREWS, R. BRINKMAN et al. (35 co-authors). 1994. Complete nucleotide sequence of Saccharomyces cerevisiae chromosome VIII. Science **265**:2077–2082.

KIMURA, M. 1983. The neutral theory of molecular evolution. Cambrige University Press, Cambridge.

LABEDAN, B., and M. RILEY. 1995. Gene products of *Escherichia coli*: sequence comparisons and common ancestries. Mol. Biol. Evol. **12**:980–987.

LALO, D., S. STETTLER, S. MARIOTTE, P. P. SLONIMSKI, and P. THURIAUX. 1993. Two yeast chromosomes are related by a fossil duplication of their centromeric regions. C. R. Acad. Sci. III **316**:367–373.

LOUIS, E. J. 1995. The chromosome ends of Saccharomyces cerevisiae. Yeast **11**:1553–1573.

MCCLINTOCK, B. 1951. Chromosome organization and genic expression. Cold Spring Harb. Symp. Quant. Biol. **16**:13–47.

MEWES, H. W., M. ALBERMANN, M. BÄHR et al. (12 co-authors). 1997. Overview of the yeast genome. Nature **387**(Suppl.):7–65.

MURAKAMI, Y., M. NAITOU, H. HAGIWARA et al. (13 co-authors). 1995. Analysis of the nucleotide sequence of chromosome VI from Saccharomyces cerevisiae. Nat. Genet. **10**: 261–268.

NEEDLEMAN, S. B., and C. D. WUNSCH. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48**: 443–453.

OHTA, T. 1988. Evolution by gene duplication and compensatory advantageous mutations. Genetics **120**:841–847.

OLIVER, S. G., Q. J. M. VAN DER AART, M. L. AGOSTONI-CARBONE et al. (147 co-authors). 1992. The complete DNA sequence of yeast chromosome III. Nature **357**:38–46.

PLUNKETT, G., V. BURLAND, D. L. DANIELS, and F. R. BLATTNER. 1993. Analysis of the Escherichia coli genome. III. DNA sequence of the region from 87.2 to 89.2 minutes Nucleic Acids Res. **21**:3391–3398.

SMITH, M. M. 1987. Molecular evolution of the Saccharomyces cerevisiae histone gene loci. J. Mol. Evol. **24**:252–259.

SOFIA, H. J., V. BURLAND, D. L. DANIELS, G. PLUNKETT, and F. R. BLATTNER. 1994. Analysis of the Escherichia coli genome. V. DNA sequence of the region from 76.0 to 81.5 minutes. Nucleic Acids Res. **22**:2576–2586.

TATUSOV, R. L., A. R. MUSHEGIAN, P. BORK, N. P. BROWN, W. S. HAYES, M. BORODOVSKY, K. E. RUDD, and E. V. KOONIN. 1996. Metabolism and evolution of *Haemophilus influenza* deduced from a whole-genome comparison with *Escherichia coli.* Curr. Biol. **6**:279–291.

TOLEDO, F., R. D. LE, G. BUTTIN, and M. DEBATISSE. 1992. Co-amplified markers alternate in megabase long chromosomal inverted repeats and cluster independently in interphase nuclei at early steps of mammalian gene amplification. EMBO J. **11**:2665–2673.

WALSH, J. B. 1995. How often do duplicated genes evolve new functions? Genetics **139**:421–428.

WOLFE, K. H., and D. C. SHIELDS. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387**:708–713.

MANOLO GOUY, reviewing editor

Accepted July 7, 1997

Analysis of intrachromosomal duplications in yeast Saccharomyces cerevisiae : a possible model for their origin

Achaz, G and Coissac, E and Viari, A and Netter, P

La séquence complète du génome de la levure *Saccharomyces cerevisiae* a été analysée pour rechercher les répétitions intrachromosomiques directement au niveau de la séquence nucléique. L'analyse a été réalisée en recherchant les grandes répétitions approchées (de 30 à 3885 pb) présentes sur chacun des chromosomes. Nous avons montré que les répétitions directes et inversées présentent des caractéristiques très différentes : les duplications directes sont plus longues et plus similaires que les duplications inversées. De plus, contrairement aux répétitions inversées, une grande majorité des répétitions directes possèdent leurs deux copies proches l'une de l'autre. La distance (*delta*) entre les deux copies est généralement inférieure à 1 kb. Une analyse fine de ces «répétitions directes proches» montre une corrélation négative entre *delta* et le pourcentage d'identité et une corrélation positive entre *delta* et la longueur de la répétition. De plus, contrairement aux autres catégories de répétition, les répétitions directes proches sont principalement localisées dans les régions codantes (CDS). Nous proposons deux hypothèses pour interpréter ces observations : soit le taux de délétion/conversion est négativement corrélé avec delta, soit il existe un mécanisme de duplication qui crée en permanence de nouvelles répétitions directes proches, les autres répétitions intrachromosomiques étant produites par des réarrangements chromosomiques à partir de ces répétitions «primaires».

Analysis of Intrachromosomal Duplications in Yeast *Saccharomyces cerevisiae*: A Possible Model for Their Origin

Guillaume Achaz, Eric Coissac,* Alain Viari,† and Pierre Netter**

*Structure et dynamique des génomes, Institut Jacques Monod, Paris, France; and †Atelier de Bioinformatique, Université Paris VI, Paris, France

The complete genome of the yeast *Saccharomyces cerevisiae* was investigated for intrachromosomal duplications at the level of nucleotide sequences. The analysis was performed by looking for long approximate repeats (from 30 to 3,885 bp) present on each of the chromosomes. We show that direct and inverted repeats exhibit very different characteristics: the two copies of direct repeats are more similar and longer than those of inverted repeats. Furthermore, contrary to the inverted repeats, a large majority of direct repeats appear to be closely spaced. The distance (delta) between the two copies is generally smaller than 1 kb. Further analysis of these ''close direct repeats'' shows a negative correlation between delta and the percentage of identity between the two copies, and a positive correlation between delta and repeat length. Moreover, contrary to the other categories of repeats, close direct repeats are mostly located within coding sequences (CDSs). We propose two hypotheses in order to interpret these observations: first, the deletion/conversion rate is negatively correlated with delta; second, there exists an active duplication mechanism which continuously creates close direct repeats, the other intrachromosomal repeats being the result, by chromosomal rearrangements of these ''primary repeats.''

Introduction

Since the first complete bacterial genome (Fleischmann et al. 1995), 22 new eubacteria, 6 archebacteria, and 3 eukaryote sequences have been completed, and several new genomics fields, such as ''functional genomics'' (deciphering the function of genes) and ''comparative genomics'' (comparison of entire genomes) (Chervitz et al. 1998), have emerged. Here, we focus on ''dynamical genomics,'' which can be seen as the study of chromosome history and dynamics through the analysis of the structure of current genomes. One way of studying these phenomena is through the analysis of chromosomal rearrangement remnants, such as duplications. Among eukaryotes, budding yeast *Saccharomyces cerevisiae,* which has been completely sequenced (Goffeau et al. 1996), is a good model because of its small size (12.1 Mb) and its comprehensive annotation.

The first evidence of sequence duplication in *S. cerevisiae* came from Lalo et al. (1993), who found a large duplication event between chromosomes II and XIV. More exhaustive studies, based on translated coding sequence (CDS) alignments, have brought prominence into large interchromosomal duplications (Coissac, Maillier, and Netter 1997; Wolfe and Shields 1997). Further analysis revealed that, for the two copies of a duplicated CDS, the distances to the closest telomere are similar (Coissac, Maillier, and Netter 1997). The importance of telomeres underlines the relation between nuclear organization and genome dynamics. Other studies were undertaken at the DNA level leading to the development of a ''duplication databank'' (Mewes et al. 1997) and to the definition of the X2 element in the subtelomeric region (Britten 1998).

However, to our knowledge, apart from the description of gene tandem duplication (CUP1, PMR2, rDNA, ASP3), no systematic study has yet been undertaken on intrachromosomal duplications. In the present work, we searched for intrachromosomal repeats at the level of nucleotide sequences. Through this analysis, we show that direct and inverted repeats exhibit very different characteristics. Moreover, we identify a special class of direct repeats (named close direct repeats) exhibiting several particular features. Finally, we propose a model based on the active flow of creation of these close direct repeats and their dispersion by chromosomal rearrangements.

Materials and Methods

Data

The *S. cerevisiae* complete sequences and annotations were extracted from the Saccharomyces Genome Database (SGD; http://genome-www.stanford.edu/Saccharomyces/). The total size of the 16 chromosomes is 12.1 Mb. We used the entire nuclear sequences as given in the database, including the three tandem clusters (CUP1, rDNA, and PMR2), which were reduced to a single repeat. We additionally built 10 ''random genomes'' by shuffling each chromosome independently with respect to its dinucleotide composition.

Construction of the Repeat Database

Our primary goal was to look for approximate repeats, i.e., repeats whose copies may not be strictly identical but may contain errors (mismatches and indels). The usual procedure for this purpose derives from dynamic programming (Smith and Waterman 1981) but is unfortunately not amenable to the study of very long sequences because of its quadratic time complexity. Although several heuristics have already been proposed to work around this problem (Leung et al. 1991; Vincens

Abbreviation: CDS, coding sequence.

Key words: genome dynamics, evolution, duplication, direct repeats, *Saccharomyces cerevisiae.*

Address for correspondence and reprints: Guillaume Achaz, Structure et dynamique des génomes, IJM, Tour 43–44, 1° étage, 4, place Jussieu, 75251 Paris CEDEX 05, France. E-mail: achaz@ijm.jussieu.fr.

Mol. Biol. Evol. 17(8):1268–1275. 2000

et al. 1998), we chose here to develop our own procedure in order to fit the biological problem more closely. Like most of the already-proposed heuristics, this procedure first looks for ''seeds'' of exact repeats and then extends the seeds by using dynamic programming techniques. This is done for each chromosome independently in four consecutive steps which are described as follows.

First Step: Searching for Seeds

Exact repeats were detected by using the Karp-Miller-Rosenberg (KMR) algorithm (Karp, Miller, and Rosenberg 1972), which finds the largest subword present at least r_{min} times (here $r_{min} = 2$) in a text (here, each chromosome). Since we were interested in ''unusually'' large repeats (i.e., repeats which did not appear by chance), we set a threshold (L_{min}) on the minimal length of repeats of interest. L_{min} was calculated using the statistics developed by Karlin and Ost (1985). For each chromosome, we chose L_{min} such that the probability of finding a three-copy repeat in a random sequence with the same length and base composition on the chromosome was less than 0.001. L_{min} typically ranges from 15 to 17 bp depending on the chromosome length.

In order to avoid the problem of any two subwords of a repeated word being themselves repeated, we devised the following heuristics (Rocha, Danchin, and Viari 1999): first, the longest repeat on the chromosome is sought, and its length is compared with the minimum preset value L_{min}. When a test is successful, both copies of the repeat are masked and excluded from further analysis. The process is iterated up to the point where the length of the largest repeats becomes smaller than L_{min}. It should be pointed out that this process is a heuristic. In particular, if there is a three-copy repeat where the third copy appears a little bit shorter, then this copy will be missed by the method. This explains the rationale behind the procedure to set up the threshold L_{min} (vide supra). We devised two versions of the program: one to detect direct repeats and the other to detect inverted repeats (repeats for which the second copy has the reverse orientation). The two orientation classes (direct and reverse) were further handled separately.

Second Step: Removing Low-Complexity, Overlapping, and Telomeric Seeds

In order to remove low-complexity repeats (like microsatellites), we used an entropy filter. The entropy is taken here in the sense of Shannon (Schneider et al. 1986) for dinucleotide distribution:

$$H = -\sum_{i=AA}^{TT} p_i \log_{16} p_i,$$

where p_i is the frequency of the ith dinucleotide. The entropy (H) is computed on the sequence of a repeat (H_{repeat}) and on the whole chromosome ($H_{chromosome}$). The values are then compared by computing the ratio $H_{repeat}/H_{chromosome}$. This ratio was calibrated by using artificial stretches of mono-, di-, tri-, and tetranucleotides to define a threshold: only repeats whose ratio was greater than 0.6 were kept.

Next, we discarded all repeats for which the two copies overlap. At this stage, these repeats generally correspond to multicopies of small words.

Finally, we removed all subtelomeric duplications. Several well-known elements are located in the subtelomeric regions (Y′ sequence [Louis and Haber 1990], X2 [Britten 1998], seripauperine [Viswanathan et al. 1994]). These elements have already been widely studied and are known to exhibit a highly special plasticity (for review, see Pryde, Gorham, and Louis 1997). We arbitrarily set a subtelomeric barrier at 30 kb and removed all repeats with at least one copy in a subtelomeric region.

Third Step: Extending the Seeds

Exact repeats (seeds) were extended into larger nonstrict repeats by using a local alignment program (Smith and Waterman 1981) developed by P. Hardy and M. Waterman (http://www.hto.usc.edu/software/seqaln/). The sequence of a seed was substituted with X's, and 100 bp were picked on both sides. For example, a seed of 30 bp will become $(A/C/G/T)_{100}$-$(X)_{30}$-$(A/C/G/T)_{100}$. The scoring matrix retained for the alignment was as follows: match(A/T/C/G) = +4; match(X) = +99; mismatch(A/T/G/C) = −4; mismatch(X) = −99; Gap_{open} = −16; $Gap_{extension}$ = −4. The value +99 will force the program to always align the two copies of the seed. When the best local alignment found by the program ended less than 10 bp from one of the sequences termini, the sequences were further extended 200 bp and a new run was performed. This operation was iterated until the alignment eventually ended more than 10 bp from both sides. It should be pointed out that after this step, several different initial seeds may give rise to the same (or a similar) extended repeat. Therefore, when two or more extended repeats occurred at the same location (with a tolerance of 20% of their length), we just kept the longest one.

Fourth Step: Removing Short or Biological Trivial Repeats

In order to remove repeats that were too short or too different, we decided to keep repeats with (1) a minimum percentage of identity and (2) a minimum number of matches between their two copies. These minima were arbitrarily set at 50% identity and 30 matches. Finally, we applied a last filter in order to remove all ''biologically trivial'' duplications, which have their own dynamics. Actually, many of the repeats were due to the 275 tRNAs, 2 rRNAs, 50 Ty's, or 385 solos widespread in the yeast genome and were therefore removed. The positions of these known repeated elements were extracted from the SGD annotations (http://genome-www.stanford.edu/Saccharomyces).

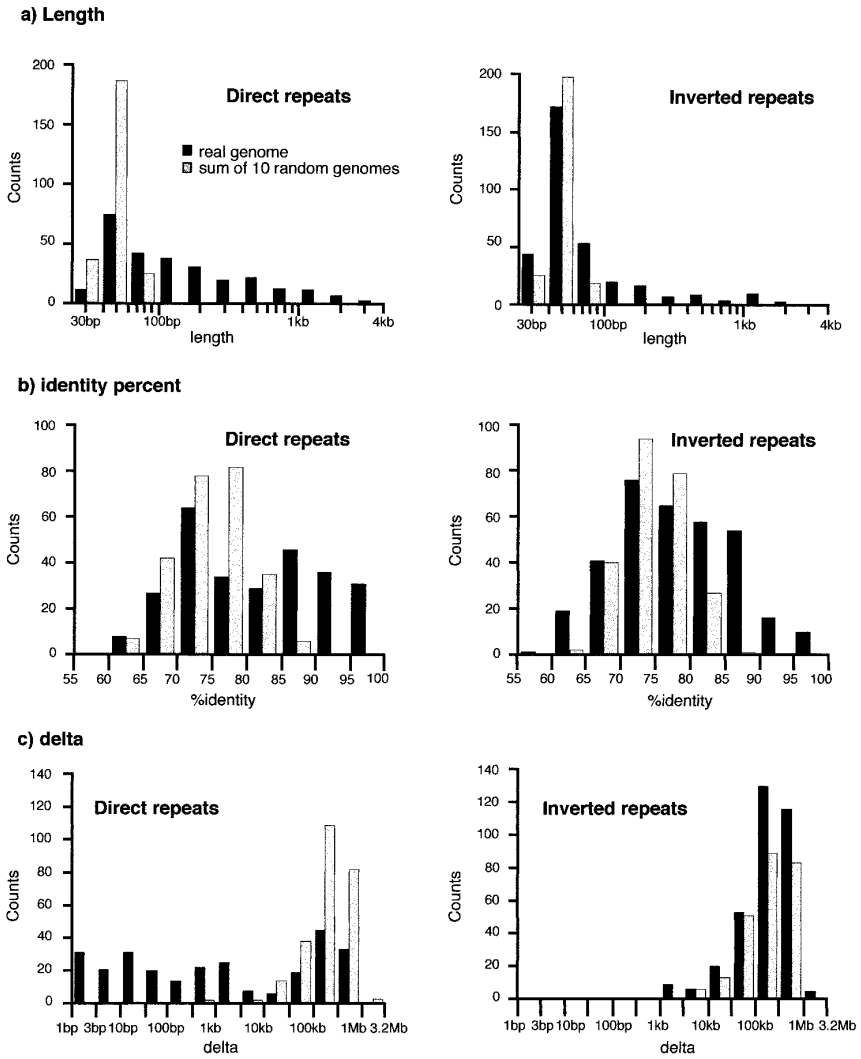


FIG. 1.—Distribution of the three parameters (length, identity, and delta) used in this study for each orientation (direct or inverted) of the repeats. Black boxes represent data observed for the real yeast genome, and gray boxes correspond to shuffled data. Since much fewer repeats are observed on random data (see text), repeats from 10 random genomes (each chromosome is shuffled with respect to the dinucleotide composition) have been pooled. *a*, Histogram of the length of the repeats. *b*, Histogram of the percentage of identity between the two copies of a repeat. *c*, Histogram of the distance (delta) between the two copies of a repeat.

Results

The application of the previously described method yields a total of 275 direct repeats and 340 inverted repeats on the yeast genome. In comparison, the random genomes (see *Materials and Methods*) produce an average of 25 direct repeats and 24 inverted repeats. The number and distribution of repeats differ from one chromosome to the other (data not shown). However, in the rest of this analysis, we pooled together all of the repeats in order to get sufficient statistics to study their global properties. In order to examine more closely the characteristics of the repeats, we focused on three parameters: ''length'' simply denotes the mean length of the two copies; ''identity'' is defined as the ratio of the number of matches between the two copies over the length of the largest copy; and ''delta,'' also called ''spacer'' in the literature (Klein 1995), is defined as the distance between the two copies. For both orientations, delta begins after the 3′ end of the first copy. It stops at the 5′ end of the second copy for direct repeats and at its 3′ end for inverted ones.

Differences Between Direct and Inverted Repeats

Figure 1 shows the distributions of the three parameters described above for the two orientation classes and for real and random genomes. The comparison of real direct repeats with random ones in figure 1*a* reveals important differences: random repeats are all shorter than

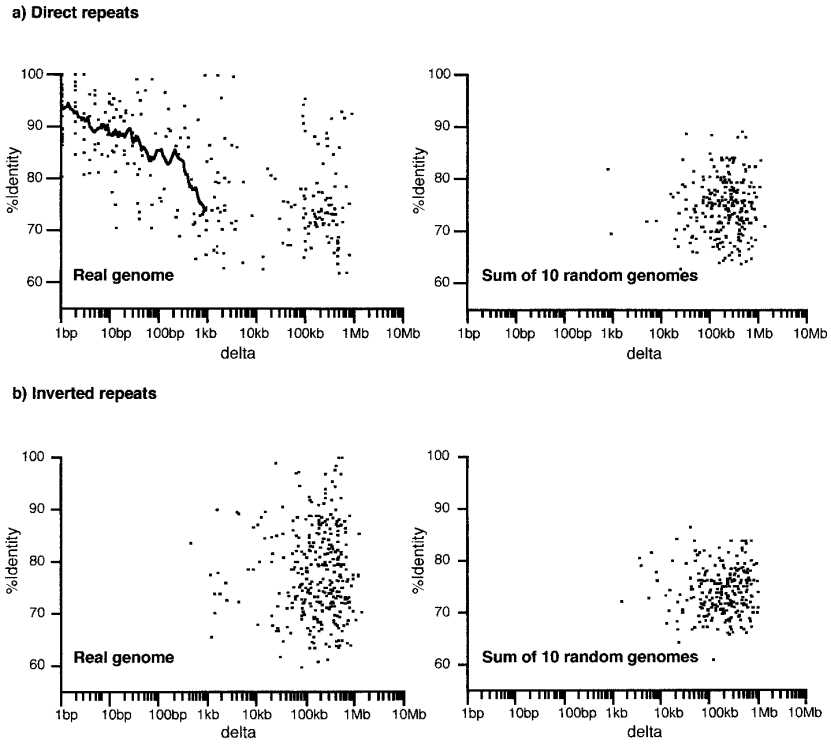


FIG. 2.—Negative correlation between the percentage of identity and the spacing (delta) between the two copies of a repeat. The percentage of identity (y-axis) is plotted as a function of delta (x-axis on a logarithmic scale) for both real (left side) and shuffled (right side) yeast genomes. Direct repeats are given in *a*, and inverted repeats are given in *b*. Since much fewer repeats are observed on random data (see text), and in order to give rise to a comparable total number of points, the plots on the right actually correspond to the sum of 10 random genomes. The black curve (for real data) represents the mean of the y values (identity) computed on a sliding window spanning 20 data points. This visual negative correlation is further confirmed by Kendall tau rank tests (see text).

100 bp, whereas a significant number (146/275) of real ones are much longer than 100 bp (up to 3,885 bp coming from the ENA family on chromosome IV). On the contrary, real inverted repeats behave much like random ones: only a few (71/340) real inverted repeats are significantly longer than 100 bp. Thus, on the sole basis of their length, it seems clear that real direct repeats are different from real inverted ones.

As shown in figure 1*b*, both real direct and inverted repeats show a higher percentage of identity than random ones. Moreover, by comparing the two orientation classes for real data, a major difference appears: direct repeats exhibit a higher degree of similarity than inverted ones (for instance, 103 direct repeats, against only 29 inverted repeats, are found above 90% identity).

Finally, the histograms of delta (fig. 1*c*) highlight another important structural difference between real direct and inverted repeats. Most (139/275) real direct repeats have deltas shorter than 1 kb, while random repeats exhibit almost exclusively deltas longer than 1 kb. In contrast, real inverted repeats display about the same distribution as random inverted ones.

In summary, these results show that both orientation classes are different from random distribution and that real direct and inverted repeats constitute two different populations with distinct properties. The main difference concerns the delta parameter, with the majority of direct repeats being closely spaced (delta smaller than 1 kb). Hereafter, we refer to them as ''close'' (as opposed to ''distant'') direct repeats.

Identity Is Negatively Correlated with Delta for Close Direct Repeats

In order to reveal possible correlations between the parameters, we plotted, for both orientation classes and for real and random genomes, the identity as a function of delta. Figure 2*a* suggests that close direct repeats are negatively correlated to delta. This visual observation is further confirmed by Kendall tau correlation measurement ($\tau = -0.36$; $P \approx 10^{-10}$). In contrast, no such correlation is found for larger deltas nor for random repeats.

Length Is Positively Correlated with Delta for Close Direct Repeats

Similarly, we searched for a correlation between length and delta for both orientation classes and for real and random genomes. Figure 3*a* reveals a peculiar variation of the length as a function of delta. More precisely, close direct repeats exhibit a positive correlation ($\tau = +0.26$; $P \approx 3 \times 10^{-6}$) between length and delta. It should be noted that no significant rank correlation

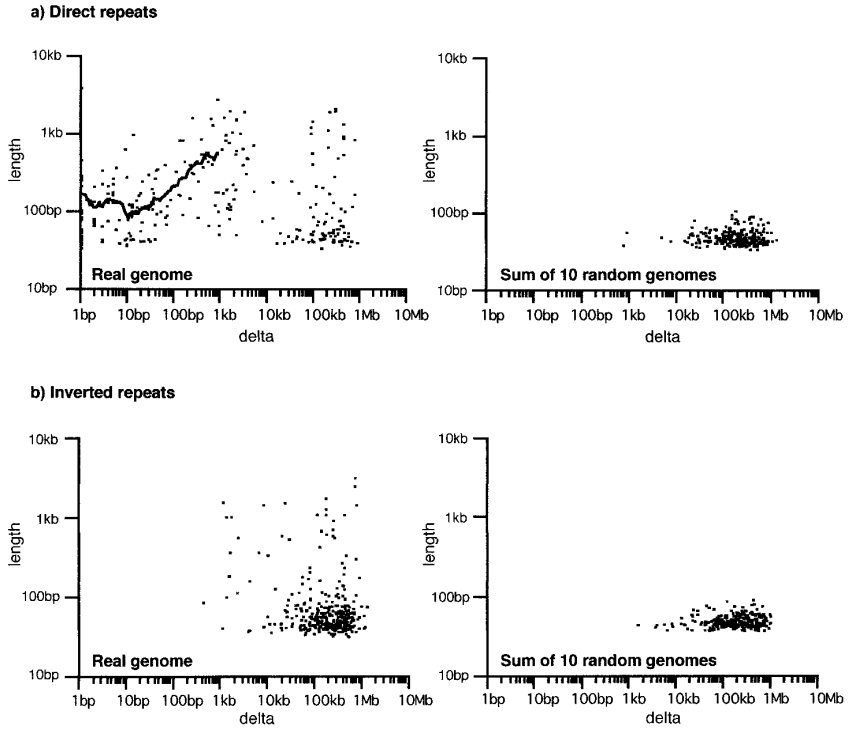


FIG. 3.—Positive correlation between the length and the spacing (delta) between the two copies of a repeat. The mean length of the two copies of a repeat (y-axis) is plotted as a function of delta (x-axis on a logarithmic scale) for both real (left side) and shuffled (right side) yeast genomes. Direct repeats are given in *a*, and inverted repeats are given in *b*. Since much fewer repeats are observed on random data (see text), and in order to give rise to a comparable total number of points, the plots on the right actually correspond to the sum of 10 random genomes. The black curve (for real data) represents the mean of the y values (identity) computed on a sliding window spanning 20 data points. This visual positive correlation is further confirmed by Kendall tau rank tests (see text).

between length and identity was observed for close direct repeats.

Close Direct Repeats Are Mostly ''Coding'' Sequences

In order to find out whether repeats are located inside CDSs, we examined positions of repeats in relation to the CDSs. This brought out two main results. Close direct repeats are mainly located within coding sequences: 85.6% (119/139) of them have their two copies completely included within CDSs and, with two exceptions, these repeats are always located within the same CDS. Moreover, it turns out that for 115 of these 117 repeats, the two copies are in the same coding frame, therefore giving rise to repeats at the protein level too.

In contrast, a much lower percentage of distant repeats (58%; 79/136) and inverted repeats (40.6%; 138/340) are completely included within CDSs. Moreover, only 50.6% (40/79) of distant direct DNA repeats and 34.1% (47/138) of inverted DNA repeats correspond to repeats at the protein level.

Discussion

This investigation on intrachromosomal duplications allows us to bring out several biological results and hypotheses about the dynamics of repeats. The first set of arguments comes from the analysis of the data presented in figures 2 and 3: for direct repeats, the main differences being observed between close repeats (delta $<$ 1 kb) and distant repeats (delta $>$ 1 kb) were as follows:

1. In figure 2, for close direct repeats, one can observe a negative correlation between the percentage of identity and delta: the shorter the delta, the higher the identity. Similar results have already been suggested for *Caenorhabditis elegans* (Semple and Wolfe 1999). This result could be understood if a high percentage of identity is (*i*) the mark of a recent adjacent duplication event and/or (*ii*) the result of an active conversion process (homogenization of the two copies). This latter process may depend upon the relative distance of the two copies (vide infra).
2. Figure 3 shows that, for close direct repeats, there is a positive correlation between length and delta: the shorter the delta, the shorter the length of the repeat. This correlation could be interpreted as the result of (*i*) a specific mechanism preferentially deleting large repeats (the loss of one copy leading to a single sequence) and/or (*ii*) genetic erosion due to the mutational events accumulated

from the initial duplication, therefore leading to a lower identity percent.

Some of the interpretations invoked above could be considered contradictory. In particular, the high percentage of identity of the close direct repeats is considered the mark of a recent duplication event (point 1, *i*), whereas their short length could be a consequence of the long time elapsed from the initial duplication event (point 2, *ii*). Therefore, as explained below, we shall consider the second explanation less probable.

Conversion Versus Deletion: A Plausible Explanation

For close direct repeats, it seems reasonable to think that the extent of the exchange process is negatively correlated with delta. The exchange process can be either deletion (loss of one copy by reciprocal recombination or replication slippage) or conversion (homogenization of the two copies by nonreciprocal recombination). In fact, experimental studies undertaken on *Bacillus subtilis* (Chedin et al. 1994) and *Escherichia coli* (Lovett et al. 1994) have highlighted similar results. Thus, the decrease in identity as a function of delta (fig. 2) could be explained by a decrease in the conversion rate.

To understand the correlation between length and delta (fig. 3), we must put the genetic exchange back in its dynamic context: actually, each close repeat can be submitted to a deletion or to a conversion event. If a deletion occurs, there is no way back. On the contrary, if a conversion occurs, the two copies are still present and a new round of exchange (i.e., conversion or deletion) is possible. So, during a long period, a bias in favor of deletion of one copy should be observed. Furthermore, several experiments have demonstrated a positive correlation between recombination rate and repeat length in yeast (Jinks, Michelitch, and Ramcharan 1993), bacteria (Peeters et al. 1988), and phages (Pierce, Kong, and Masker 1991). Briefly, for a short delta, a long repeat should be too unstable to persist, but by increasing delta, longer repeats could be maintained. This ''length tolerance'' effect could explain the observed positive correlation between length and delta.

Functional Pressures: A Protection from Deletion

Another important difference between close and distant repeats is related to their presence within CDS: close direct repeats are located mainly within CDSs and in the same frame, therefore giving rise to repeats at the protein level as well. On the contrary, distant direct repeats give rise to fewer protein repeats. These observations lead to two nonexclusive hypotheses:

1. Close direct DNA repeats are most probably submitted to an active recombination pressure leading to the deletion of one of the copies. However, the repeat can be fixed if it is submitted to functional pressures at the protein level. The consequence is that one very rarely observes close direct repeats which have not been protected from deletion by this selective advantage (i.e., located outside CDSs), because they have been massively lost by recombination. Finally, close direct repeats which have been fixed by functional pressures at the protein level are still submitted to an active conversion process (vide infra), preventing any further evolution.
2. On the other hand, distant repeats are submitted to less active recombination and conversion pressures. This allows the creation of different proteins from the same repeated DNA sequence located in different CDSs, and even sometimes translated in different reading frames.

It should be pointed out that Marcotte et al. (1998) recently reported that there is a high frequency of internal repeats within proteins sequences of eukaryotes (as compared with prokaryotes). This observation can be in line with the result presented in this study. To summarize, we probably observe a combination of DNA mechanisms which tend to (1) delete the close repeats and (2) keep them identical and which are constrained by functional pressures at the protein level.

Direct Versus Inverted Repeats

The last group of results we take into account concerns the important differences observed between direct and inverted repeats. These differences can be summarized as follows:

1. The direct repeats exhibit a higher similarity than the inverted ones (fig. 1*b*). If we assume that there is a higher conversion rate for the numerous close direct repeats (vide supra), this observation is not surprising.
2. The direct repeats are clearly longer that the inverted ones (fig. 1*a*). This result is surprising, since, as already mentioned, long direct repeats are experimentally known to be more easily deleted (Jinks, Michelitch, and Ramcharan 1993). Therefore, our interpretation is that the observed long direct repeats were produced more recently than inverted ones and have not yet been eliminated.
3. Finally, the repartition of deltas is the main difference between the two groups of repeats (fig. 1*c*). Inverted repeats do not seem to be constrained by delta (with the corresponding distributions being almost identical for real and random genomes). On the contrary, as shown above, close direct repeats are overrepresented and distant direct repeats are underrepresented as compared with inverted ones. We now discuss a possible model to account for these differences.

A Dynamic Model for Intrachromosomal Repeats

We propose a simple model, illustrated in figure 4, to explain all these observations and solve the apparent contradictions. In this model, the initial event is the continuous production of close direct repeats. Whatever the mechanism giving rise to it (unequal crossing over or replication slippage), when a close direct repeat is created, it can be either modified by mutation (although the

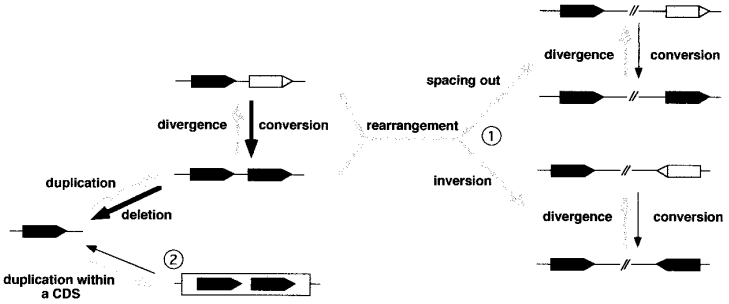


FIG. 4.—A model for the origin and dynamics of intrachromosomal repeats. The initial event is the close duplication of a sequence (oriented black boxes). The two copies can then diverge (oriented gray boxes) or be maintained identical through a conversion process. Alternatively, the repeat can also be deleted, leading back to a single copy. On a long timescale, this second situation prevails. Therefore the only two ways to maintain both copies are (case 1) to move them away through chromosomal rearrangements, since the relative conversion rate then decreases (thin arrows), and (case 2) to protect them from deletion by functional pressures; the two copies are located within CDSs.

high conversion rate will tend to maintain the two copies identical) or deleted (since the deletion rate is high). As long as the repeat remains, a new exchange (conversion/deletion) is possible. Therefore, the fate of a close direct repeat is to disappear sooner or later (depending on the conversion rate vs. the deletion rate). As a consequence, only two kinds of direct repeats can be conserved on a large time scale:

1. Coding repeats, which can be conserved by functional pressures. In this case, they must be short due to the length tolerance effect. One should note, however, that strong functional pressures and/or multiple-copy repeats can lead to maintenance of large tandem clusters (rDNAs, ENA family, ASP3 cluster, CUP1 cluster, etc).
2. Repeats in which one of the two copies is moved away by an interchromosomal (not represented in fig. 4) or intrachromosomal rearrangement: an inversion will lead to inverted repeats, and an insertion will lead to distant direct repeats. Under this model, we could explain the underrepresentation of distant direct repeats by a lower level of insertion as compared to the inversion one.

This model implies that most intrachromosomal repeats originate from close direct duplications but does not preclude any mechanism. Furthermore, it gives rise to several predictions that can be experimentally tested, like a negative correlation of the deletion/conversion rate with delta.

Acknowledgments

We thank E. Rocha, J. Pothier, E. Maillier, and D. Higuet for their scientific help and their friendly support. This work was supported by grants from Association pour la Recherche sur le Cancer. E.C. and P.N. are members of the Université Pierre et Marie Curie, Paris.

LITERATURE CITED

BRITTEN, R. J. 1998. Precise sequence complementarity between yeast chromosome ends and two classes of just-subtelomeric sequences. Proc. Natl. Acad. Sci. USA **95**:5906–5912.

CHEDIN, F., E. DERVYN, R. DERVYN, S. D. EHRLICH, and P. NOIROT. 1994. Frequency of deletion formation decreases exponentially with distance between short direct repeats. Mol. Microbiol. **12**:561–569.

CHERVITZ, S. A., L. ARAVIND, G. SHERLOCK et al. (13 co-authors). 1998. Comparison of the complete protein sets of worm and yeast: orthology and divergence. Science **282**: 2022–2028.

COISSAC, E., E. MAILLIER, and P. NETTER. 1997. A comparative study of duplications in bacteria and eukaryotes: the importance of telomeres. Mol. Biol. Evol. **14**:1062–1074.

FLEISCHMANN, R. D., M. D. ADAMS, O. WHITE et al. (11 co-authors). 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd.* Science **269**:496–512.

GOFFEAU, A., B. G. BARELL, H. BUSSEY et al. (16 co-authors). 1996. Life with 6000 genes. Science **274**:546–567.

JINKS, R. S., M. MICHELITCH, and S. RAMCHARAN. 1993. Substrate length requirements for efficient mitotic recombination in *Saccharomyces cerevisiae.* Mol. Cell. Biol. **13**:3937–3950.

KARLIN, S., and F. OST. 1985. Maximal segmental match length among random sequences from a finite alphabet. Pp. 225–243 *in* L. M. L. CAM and R. A. OLSHEN, eds. Proceedings of the Berkeley Conference in honour of Jerzy Neyman and Jack Kiefer. Vol. 1. Association for Computing Machinery, New York.

KARP, R. M., R. E. MILLER, and A. L. ROSENBERG. 1972. Rapid identification of repeated patterns in strings, trees and arrays. Pp. 125–126 *in* Proceedings 4th Annual ACM Symposium Theory of Computing, New York.

KLEIN, H. L. 1995. Genetic control of intrachromosomal recombination. Bioessays **17**:147–159.

LALO, D., S. STETTLER, S. MARIOTTE, P. P. SLONIMSKI, and P. THURIAUX. 1993. Two yeast chromosomes are related by a fossil duplication of their centromeric regions. C. R. Acad. Sci. **316**:367–373.

LEUNG, M. Y., B. E. BLAISDELL, C. BURGE, and S. KARLIN. 1991. An efficient algorithm for identifying matches with errors in multiple long molecular sequences. J. Mol. Biol. **221**:1367–1378.

LOUIS, E. J., and J. E. HABER. 1990. The subtelomeric Y′ repeat family in *Saccharomyces cerevisiae*: an experimental system for repeated sequence evolution. Genetics **124**:533–545.

LOVETT, S. T., T. J. GLUCKMAN, P. J. SIMON, V. J. SUTERA, and P. T. DRAPKIN. 1994. Recombination between repeats in

Escherichia coli by a recA-independent, proximity-sensitive mechanism. Mol. Gen. Genet. **245**:294–300.

MARCOTTE, E. M., M. PELLEGRINI, T. O. YEATES, and D. EISENBERG. 1998. Census of protein repeats. J. Mol. Biol. **293**:151–160.

MEWES, H. W., K. ALBERMANN, M. BAHR et al. (12 co-authors). 1997. Overview of the yeast genome. Nature **387**: 7–65.

PEETERS, B. P., B. J. DE, S. BRON, and G. VENEMA. 1988. Structural plasmid instability in *Bacillus subtilis*: effect of direct and inverted repeats. Mol. Gen. Genet. **212**:450–458.

PIERCE, J. C., D. KONG, and W. MASKER. 1991. The effect of the length of direct repeats and the presence of palindromes on deletion between directly repeated DNA sequences in bacteriophage T7. Nucleic Acids Res. **19**:3901–3905.

PRYDE, F. E., H. C. GORHAM, and E. J. LOUIS. 1997. Chromosome ends: all the same under their caps. Curr. Opin. Genet. Dev. **7**:822–828.

ROCHA, E. P. C., A. DANCHIN, and A. VIARI. 1999. Analysis of long repeats in bacterial genomes reveals alternative evolutionary mechanisms in *Bacillus subtilis* and other competent prokaryotes. Mol. Biol. Evol. **16**:1219–1230.

SCHNEIDER, T. D., G. D. STORMO, L. GOLD, and A. EHRENFEUCHT. 1986. Information content of binding sites on nucleotide sequences. J. Mol. Biol. **188**:415–431.

SEMPLE, C., and K. H. WOLFE. 1999. Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. J. Mol. Evol. **48**:555–564.

SMITH, T. F., and M. S. WATERMAN. 1981. Identification of common molecular subsequences. J. Mol. Biol. **147**:195–197.

VINCENS, P., L. BUFFAT, C. ANDRE, J. P. CHEVROLAT, J. F. BOISVIEUX, and S. HAZOUT. 1998. A strategy for finding regions of similarity in complete genome sequences. Bioinformatics **14**:715–725.

VISWANATHAN, M., G. MUTHUKUMAR, Y. S. CONG, and J. LENARD. 1994. Seripauperins of *Saccharomyces cerevisiae*: a new multigene family encoding serine-poor relatives of serine-rich proteins. Gene **148**:149–153.

WOLFE, K. H., and D. C. SHIELDS. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387**:708–713.

MANOLO GOUY, reviewing editor

Accepted May 2, 2000

Study of intrachromosomal duplications among the eukaryote genomes

Achaz, G and Netter, P and Coissac, E

Les répétitions intrachromosomiques ont été recherchées au niveau nucléique sur les séquences complètes de chromosomes eucaryotes. Cette analyse a été menée en recherchant les répétitions non exactes sur deux génomes complets, *Saccharomyces cerevisiae* et *Caenorhabditis elegans*, et quatre génomes partiellement séquencés, *Drosophila melanogaster*, *Plasmodium falciparum*, *Arabidopsis thaliana* et *Homo sapiens*. Ces analyses montrent que tous les eucaryotes possèdent des répétitions intrachromosomiques aux des caractéristiques similaires, suggérant une même dynamique. La majorité des répétitions directes possèdent leurs deux copies proches l'une de l'autre et ces deux copies sont plus similaires et plus courtes que les autres répétitions. À l'inverse, il n'y a pratiquement pas de répétitions inversées. Ces résultats sont compatibles avec un modèle dynamique des duplications. Ce modèle est basé sur une genèse continue de répétitions en tandem et implique que la majorité des répétitions possédant leurs deux copies éloignées dérive de ces répétitions en tandem par des remaniements chromosomiques (insertions, inversions et délétions). Des traces de ces réarrangements ont pu être mises en évidence par une analyse fine des séquences chromosomiques. Malgré cette dynamique partagée par tous les eucaryotes, chaque génome possède son propre style de duplication intrachromosomique. La densité des éléments répétés est identique pour tous les chromosomes d'un même organisme, mais différente d'une espèce à l'autre. Cette densité est à mettre en relation avec les taux relatifs de duplication, délétion et de mutation propre à chaque espèce. Il est à noter que la densité de répétition pour le chromosome X de *C. elegans* est plus faible que pour les autosomes de cet organisme, suggérant que l'échange entre chromosomes homologues est important dans le processus de duplication.

Study of Intrachromosomal Duplications Among the Eukaryote Genomes

Guillaume Achaz, Pierre Netter, and Eric Coissac

Structure et Dynamique des Génomes, Institut Jacques Monod, Paris, France

Complete eukaryote chromosomes were investigated for intrachromosomal duplications of nucleotide sequences. The analysis was performed by looking for nonexact repeats on two complete genomes, *Saccharomyces cerevisiae* and *Caenorhabditis elegans*, and four partial ones, *Drosphila melanogaster*, *Plasmodium falciparum*, *Arabidopsis thaliana*, and *Homo sapiens*. Through this analysis, we show that all eukaryote chromosomes exhibit similar characteristics for their intrachromosomal repeats, suggesting similar dynamics: many direct repeats have their two copies physically close together, and these close direct repeats are more similar and shorter than the other repeats. On the contrary, there are almost no close inverted repeats. These results support a model for the dynamics of duplication. This model is based on a continuous genesis of tandem repeats and implies that most of the distant and inverted repeats originate from these tandem repeats by further chromosomal rearrangements (insertions, inversions, and deletions). Remnants of these predicted rearrangements have been brought out through fine analysis of the chromosome sequence. Despite these dynamics, shared by all eukaryotes, each genome exhibits its own style of intrachromosomal duplication: the density of repeated elements is similar in all chromosomes issued from the same genome, but is different between species. This density was further related to the relative rates of duplication, deletion, and mutation proper to each species. One should notice that the density of repeats in the X chromosome of *C. elegans* is much lower than in the autosomes of that organism, suggesting that the exchange between homologous chromosomes is important in the duplication process.

Introduction

All eukaryote genomes exhibit similar physical structures and constraints (i.e., linear chromosomes, scaffold attachment, nucleosome organization). However, many characteristics highlight important differences between them: (1) coding sequences represent 72% of the *Saccharomyces cerevisiae* genome (Goffeau et al. 1996) and only 2%–5% of the *Homo sapiens* genome (Dunham et al. 1999), (2) a centimorgan corresponds to kilobases in *S. cerevisiae* (Baudat and Nicolas 1997) and to megabases in humans (Dunham et al. 1999), (3) the number of introns per gene and the density of transposons increase with the genome size, and (4) the isochores organization has been ascribed mostly to vertebrate genomes (Bernardi 2000). Despite these differences, one would expect to find remnants of a similar nuclear organization in genome sequences. Events of DNA duplication were described in many eukaryote genomes, but are the duplication dynamics similar in all eukaryotes?

Within duplication events, four main subprocesses have been documented: abnormal segregation during cell division (leading to entire-chromosome[s] duplication, viz., hyperploidization, and sometimes to whole-genome doubling, viz. polyploidization), transposition (duplication of transposable elements), expansion of low-complexity sequences (microsatellites and minisatellites), and finally generic duplications of unspecific DNA regions within the same chromosome or between two chromosomes. We shall henceforth refer to this last subprocess as the iteration process. Polyploidization events were proposed to explain the large-scale duplications at the origin of vertebrates (Ohno 1970), in many angiosperms (Masterson 1994)—even in *Arabidopsis thaliana* (Blanc et al. 2000), in the fish lineage (Amores et al. 1998), and in the yeast *S. cerevisiae* (Wolfe and Shields 1997). However, it is not clear if these large-scale duplications are always the result of polyploidization, successive hyperploidizations, or bursts of large iterations (Holland 1999; Llorente et al. 2000; Vision, Brown, and Tanksley 2000; Hughes, Da Silva, and Friedman 2001; Robinson-Rechavi et al. 2001).

In order to investigate the iteration process, we focused our attention on intrachromosomal repeats in the chromosome sequences. Two complete genomes, *S. cerevisiae* and *Caernorhabditis elegans*, and four partial ones, *H. sapiens*, *Drosophila melanogaster*, *A. thaliana*, and *Plasmodium falciparum*, were analyzed. It should be noted that the genome of *S. cerevisiae* was already investigated for its repeats in a previous study (Achaz et al. 2000) in which we proposed a model for the dynamics of the iteration process based on a continuous genesis of close direct repeats (CDR). A CDR is defined here as a repeat with its copies in the same orientation and with a physical distance between them (the spacer) smaller than 1 kb. The model supposes that most of the intrachromosomal repeats originate from these CDRs, the others being the result of further chromosomal rearrangements. In the present study, the model established in yeast was tested for new eukaryote chromosomes. We focused on the differences between genomes and tried to connect them to the genome context. In our model, supposing that most of the intrachromosomal repeats originate from tandem repeats, the chromosome sequences had been investigated to find the remnants of the chromosomal rearrangements. Hence, we view repeats as the markers of genome dynamics.

Materials and Methods

Data

We analyzed the complete eukaryote genomes of *S. cerevisiae*—16 chromosomes—(Goffeau et al. 1996)

Key words: genome dynamics, evolution, duplication, eukaryotes.

Address for correspondence and reprints: Guillaume Achaz, Structure et Dynamique des Génomes, Institut Jacques Monod, Tour 43–44, 1° Étage, 4, Place Jussieu, 75251 Paris Cedex 05, France.
E-mail: achaz@ijm.jussieu.fr.

Mol. Biol. Evol. 18(12):2280–2288. 2001

Table 1
Estimation of the Intrachromosomal Redundancy of Each Genome

Species	Chromosome Mean Size (Mb)	Two-copy Seeds[a] (% bp)	All Seeds[b] (% bp)
S. cerevisiae	0.75	1.77	3.83
P. falciparum	1.01	4.67	9.04
A. thaliana	18.60	4.83	10.13
C. elegans	15.87	13.60	22.19
D. melanogaster	19.07	1.10	3.22
H. sapiens	33.65	3.90	18.68

[a] The proportion (as a percentage of the total base pairs) of the analyzed chromosomes included in the two-copy seeds (exact repeats). It should be noted that only two-copy seeds are kept for further analysis.
[b] The proportion of the analyzed chromosomes included in all seeds.

and *C. elegans*—six chromosomes—(Consortium 1998), and four partial genomes: *H. sapiens*—chromosomes 21 (Hattori et al. 2000) and 22 (Dunham et al. 1999), *P. falciparum*—chromosomes 2 (Gardner et al. 1998) and 3 (Bowman et al. 1999), *A. thaliana*—chromosomes 2 (Lin et al. 1999) and 4 (Mayer et al. 1999), and six chromosomal arms (X, 2L, 2R, 3L, 3R, 4) of *D. melanogaster* (Adams et al. 2000).

Sequences of *H. sapiens*, *C. elegans*, *P. falciparum*, and *A. thaliana* were extracted from GenBank (ftp://ncbi.nlm.nih.gov/genbank/genomes). The *S. cerevisiae* chromosomes were extracted from Saccharomyces Genome Database (http://genome-www.stanford.edu/Saccharomyces). Sequences of *D. melanogaster* were downloaded from Celera database (http://www.celera.com).

It should be pointed out that most sequences contain many gaps (stretches of N). For example, in chromosome 1 of *C. elegans*, 8.8% of its base pairs are N, and 29 gaps are longer than 10 kb. These stretches were not taken into account during the construction of the repeats' database.

Construction of the Repeats Database

General trends of repeats detection, like most of the heuristics already proposed (Leung et al. 1991; Vincens et al. 1998), are based on looking first for seeds (exact repeats) and then extending them with a local alignment program. The detailed methodology is described below through three main steps: searching, filtering, and extending.

First Step: Searching for Seeds

In this step, exact repeats (seeds) were detected by using the REPuter software (Kurtz and Schleiermacher 1999). This software detects all seeds (direct and inverted) in a given sequence that are any distance apart from the chromosome. As we are interested in unusually large seeds, the minimum length of seeds (L_{min}) was calculated using the statistics developed by Karlin and Ost (1985). For each chromosome, we chose L_{min} such that the probability of finding a two-copy word with at least this length in a same-size, same-nucleotide composition random sequence is 0.001. Typically, L_{min} ranges from 21 for the smallest chromosome (chromosome 1 of *S. cerevisiae*) to 28 for the largest ones (chromosomes 21 and 22 of *H. sapiens*).

Second Step: Filtering the Seeds

First, to remove all low-complexity seeds (i.e., microsatellites or poly-A stretches), we used an entropy filter based on dinucleotide composition (Achaz et al. 2000). Second, all multicopy seeds were removed. A chromosome map in which each position is linked to its n-plication degree (duplication, triplication, etc.) was established. To build this map, we counted for each chromosome position the number of times this position is found in seeds (direct and inverted seeds were pooled together). This map is used to estimate the degree of redundancy of chromosomes (i.e., the number of duplications, triplications, etc.). Table 1 presents, for each species, the mean size of the chromosomes, the percentage of chromosomes included in two-copy seeds, and the percentage of chromosomes represented by all the seeds. As we are only interested here in two-copy seeds, we used the map to remove all seeds in which one of the positions is included in a multicopy repeat.

Third Step: Extending the Seeds

Seeds were extended into larger nonstrict repeats by using a local alignment program (Smith and Waterman 1981) available at http://www-hto.usc.edu/software/seqaln. It should be pointed out that many seeds might give rise to the same extended repeat. Therefore, when two or more repeats occurred in the same location, we just kept the first one. Before the alignment is performed, 100 bp were picked up on both sides of the seeds. Thus, for a given seed of size N, the first alignment is computed with two sequences of $2 \times 100 + N$ bp. The following matrix, which was built empirically, was retained for local alignments: $match_{(A/T/C/G)} = 4$, mismatch $= -4$, $Gap_{open} = -16$, and $Gap_{extension} = -4$. When the best local alignment ends at less than 10 bp of a terminus, 200 bp were added at the termini, and a new run was done. As the alignment of a large sequence requires too many computer resources, we devised the following heuristic to compute the alignment of large sequences. If the alignment size was more than 1 kb, the partial alignment was memorized, and only the rest of the alignment was computed in a new

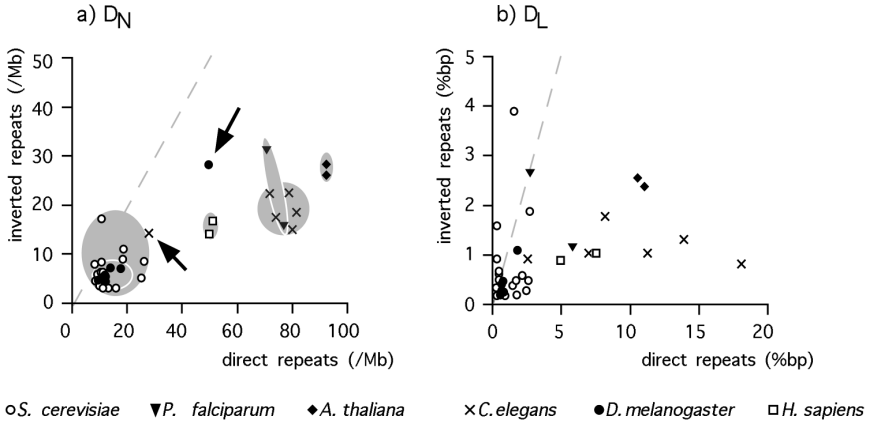


FIG. 1.—Occurrence and density of direct and inverted repeats in each chromosome. For each chromosome of each species, inverted repeats were compared to direct repeats. (*a*) Plot of D_N (density in number: number divided by chromosome length) of inverted repeats as a function of D_N of direct repeats. Chromosomes of the same species are grouped in gray areas, with the exception of the X chromosome of *C. elegans* and the fourth chromosomal arm of *D. melanogaster*, both indicated by black arrows. (*b*) Plot of D_L (density in length: sum of repeats length divided by chromosome length) of inverted repeats as a function of D_L of direct repeats. Each species is represented by a different symbol given just below the plot.

run. The process goes on until both sides of the complete alignment end at more than 10 bp of the termini. Thus, it provides a nonoptimal alignment but allows us to extend very large repeats. Then we removed all repeats in which the copies overlap because they generally correspond, at this stage, to three-copy repeats.

It should be mentioned that the methodology was similar to the one previously used in the *S. cerevisiae* analysis (Achaz et al. 2000), but was modified in order to analyze in the same way the chromosomes of yeast (<1.5 Mb) and man (35 Mb). The major modifications were applied to reduce the number of seeds and to keep only *sensu stricto* duplicated seeds (present only in two copies) for the alignment process.

Results and Discussion

The application of the methodology described above yields for direct and inverted repeats, respectively: 110 and 75 for *S. cerevisiae*, 136 and 48 for *P. falciparum*, 2,407 and 1,068 for *A. thaliana*, 6,885 and 1,845 for *C. elegans*, 1,479 and 691 for *D. melanogaster*, and 3,457 and 2,406 for *H. sapiens*.

Genome Style and History of Chromosomes

In order to analyze the relationship between chromosome size and redundancy level, we measured two parameters D_N and D_L, defined as follows:

$$D_N = \frac{\text{Number of repeats}}{\text{Size of chromosome}}$$

$$D_L = \frac{\sum \text{Length of repeats}}{\text{Size of chromosome}}$$

As predicted by the estimated redundancy of the genome in Table 1, if we exclude the Drosophila chromosomal arms (which are clearly underrepeated for their size), D_N and D_L are positively correlated with the chromosome size, using a Kendall tau-rank test ($\tau = 0.30$, $P < 0.05$ for D_N and $\tau = 0.40$, $P < 0.01$ for D_L). These observations are in agreement with an analysis of gene redundancy undertaken on partial genome sequences (Coissac, Maillier, and Netter 1997).

Two hypotheses can be proposed to explain the low densities of the chromosomal arms of *D. melanogaster*. The first one is a data bias: it should be noted that the analyzed sequences are constituted exclusively of euchromatine (only around two-thirds of the complete genome), and it is known that repeats are concentrated inside heterochromatine (Henikoff 2000). Moreover, assembly errors could lead to artificially deleted tandem repeats. The second hypothesis rests on biological grounds. One can imagine that Drosophila's genome has a special status in the duplication process (because there is no meiotic crossing-over in the male, the duplication process can be less active). The achievement of the complete sequence of *D. melanogaster* should solve this problem.

In order to investigate more precisely each chromosome, we analyzed D_N and D_L for direct and inverted repeats (fig. 1). It appears that D_N is similar for all chromosomes within the same species, whereas D_L is not. Thus, D_N could define the style of redundancy of the genome. We assume that D_N results from the iteration events combined with the loss of duplicated sequences, and then propose D_N to be connected to the biological machinery of each species. Because the machinery is clearly different for each species, but similar for all chromosomes within the same genome, D_N should be the consequence of each genome's dynamics. Furthermore, the differences between species come essentially from direct repeats, and less from inverted repeats. This suggests that the biological machinery is more connected to the creation and the loss of direct repeats than to the dynamics of inverted repeats.

The only two exceptions are the fourth chromosomal arm of *D. melanogaster* and the X chromosome of *C.*

elegans. The high density of the small fourth chromosomal arm of *D. melanogaster* could be the result of its particular structure (if there is no data bias): it is mostly constituted of heterochromatin, but, contrary to centromeric chromatin (or Y chromosome), it is partially visible in polytene chromosomes. On the contrary, the X chromosome of *C. elegans* exhibits a lower D_N than that of the other worm's chromosomes. This observation is in good agreement with the unequal distribution of repetitive elements, such as CeRep23 (Barnes et al. 1995), Cele1, Cele2, and Cele42 (Surzycki and Belknap 2000), between the autosomes and the X chromosome in *C. elegans*. It should be pointed out that exchanges between the homologous X chromosomes are only possible in hermaphrodite XX (males are X0), which could explain this lower D_N. If this is true for *C. elegans*, one may expect this to be true for all heterochromosomes. The X chromosomal arms of *D. melanogaster* seem similar to the other chromosomal arms; however, none of the Drosphila male chromosomes is submitted to meitoic crossing-over.

Contrary to D_N, D_L could reflect better the chromosome history than the effects of the cellular machinery: a unique event of iteration can lead to a high D_L for direct or inverted repeats. For example, direct repeats of the chromosome 1 of *C. elegans* exhibit a high D_L and a normal D_N (when compared with the other *C. elegans* chromosome values). This particularity is mainly caused by two large duplicated sequences, one 250-kb long (with an identity of 98.7%) and the other 600-kb long (fractionated into several segments of high identity, often more than 99%). Furthermore, the inverted repeats of the chromosome 1 of *S. cerevisiae* show a high D_L and a normal D_N, as a consequence of two internal regions inversely repeated in subtelomeres (Britten 1998).

A Model of Dynamics of Iteration

Our model of intrachromosomal iteration (Achaz et al. 2000) is based on a permanent genesis of CDR. The CDRs are then submitted to a high level of exchange (conversion and deletion). This high exchange rate tends to maintain the two copies identically (conversion) and also to eliminate them (deletion). At each round of exchange, both events are possible, but whereas conversion may still be followed by deletion, a deletion event cannot be followed by conversion.

Therefore, on a long timescale, a bias in favor of deletion should be observed. A CDR has to disappear sooner or later (depending on the relative rates of conversion and deletion). However, there are two situations where a repeat would be maintained: when it is protected from deletions by functional pressures (i.e., located inside a gene) or when the copies are spaced by further chromosomal rearrangements. This model was mainly based on three observations for CDR: (1) they are overrepresented, (2) they are mostly located inside the same gene, and (3) their length is positively correlated with the spacer (the physical distance between copies), and their identity is negatively correlated with it.

Through the present analysis, the model was tested with other eukaryotes. It should be mentioned that a model of tandem creation and further dispersion was already invoked for the families of two genes (*Hox* and *NBG*) in *C. elegans* (Ruvkun and Hobert 1998). The annotations of eukaryote chromosomes being partial, they were not taken into account. Thus, we did not analyze the relation of repeats position with genes location.

CDRs Are Overrepresented

The repartitions of spacer size for direct and inverted repeats (fig. 2) reveal that CDRs are overrepresented as compared with close inverted repeats. Moreover, in the previous study (Achaz et al. 2000), the repeats of *S. cerevisiae* were compared with the repeats that issued from random chromosomes. From this comparison, we showed that such close repeats (inverted or direct) are absent from random chromosomes. This strongly suggests that these CDRs are not the result of chance. The presence of many CDRs in all chromosomes is in good agreement with the model.

However, the repartition spacer's length indicates the existence of many direct repeats with a spacer between 1 and 10 kb in *A. thaliana* chromosomes (they represent more than one-third of all direct repeats). We looked for a plausible explanation for this overrepresentation in *A. thaliana* (as compared with other species), with particular attention to the sequence located between the two copies (the spacer). Several hypotheses can be envisaged and rejected: (1) repeats are not the edges of transposons because the spacers are not paralogous, (2) the hypothesis of campbell-like insertions of exogenous DNA, as it was proposed for *B. subtilis* (Rocha, Danchin, and Viari 1999*a*), can be eliminated because there is no difference in nucleotide composition between spacers and chromosomes, (3) there is no clear difference between these repeats and others—no high identity level, no special length, no special physical location. Similar observations can be established for the genome of *C. elegans* and *D. melanogaster*, where direct repeats with a spacer between 1 and 10 kb are also overrepresented.

In conclusion, we did not yet find any plausible hypothesis to understand why these repeats are overrepresented.

CDRs Are Identical and Short

We started by characterizing CDRs in terms of the distribution of their identity (fig. 3). Except for *S. cerevisiae*, CDRs have their two copies more identical than distant direct repeats ($P < 10^{-4}$, Mann-Whitney rank test).

In order to explain the *S. cerevisiae* exception, one should take into consideration that the distinction between close and distant repeats has been arbitrarily fixed at the same spacer size (1 kb) for each organism. The biological difference between close and distant repeats is connected to the recombination machinery. As this machinery varies from yeast to human, the limit between close and direct repeats should not be identical for all species. In that way, it can be shown that for *S. cerevisiae*, direct repeats with a spacer smaller than 500

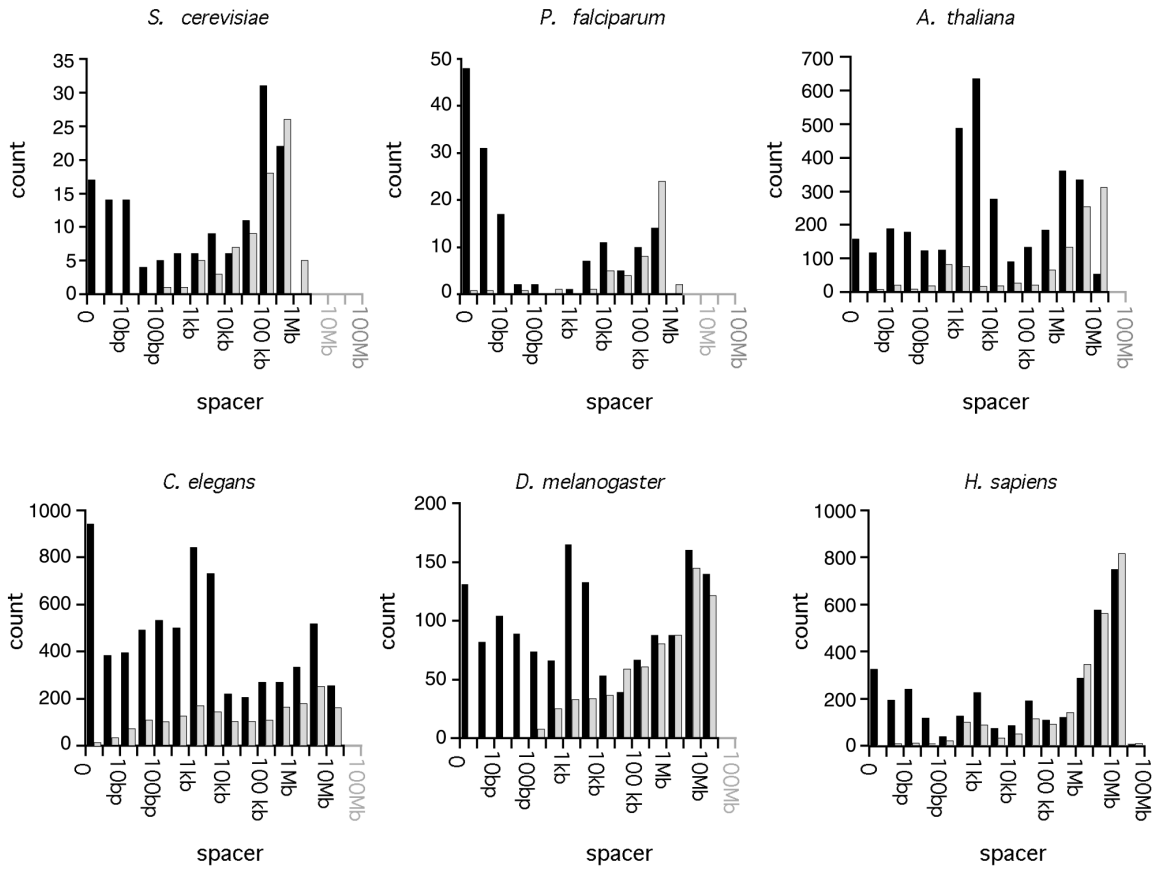


FIG. 2.—Distribution of spacers for each orientation (direct and inverted). Each histogram shows the distribution of spacers (the distance between the two copies) for direct and inverted repeats. Black boxes represent the direct repeats, and gray boxes represent the inverted ones. The distribution is established by the log 10 of the spacer size (in steps of 0.5). These histograms show clearly that CDRs (direct repeats with a spacer smaller than 1 kb) are overrepresented in all species.

bp are more identical than other direct repeats ($P < 0.05$, Mann-Whitney rank test).

This greater similarity could be explained, on the one hand, by the recent origin of these repeats and, on the other, by a high conversion rate between the two copies when they are close together. As previously discussed, CDR could also be submitted to a high deletion rate. It has been reported that recombination rate is positively correlated with repeat length in yeast (Jinks, Michelitch, and Ramcharan 1993) and in mammalian cells (Rubnitz and Subramani 1984). Thus, CDRs with long copies are too unstable to persist, and only small CDRs are conserved. In order to test this hypothesis, the length distributions of close and distant direct repeats were compared: it appeared that CDRs are smaller than the distant ones ($P < 10^{-4}$, Mann-Whitney rank test).

CDRs Exhibit an Exchange Rate Negatively Correlated with the Spacer Size

We previously observed a positive rank correlation between length and spacer and a negative rank correlation between identity and spacer for CDR in yeast: the closer the repeats, the more identical and shorter they are. Except for the *P. falciparum* chromosomes, correlations between identity, length and spacer were found in all eukaryotes (Table 2). This is in good agreement with an observation reported in *C. elegans* that the similarity between paralogous genes is negatively correlated with the physical distance between them (Semple and Wolfe 1999).

In order to understand such a result, we proposed that, as in bacteria (Peeters et al. 1988), the exchange rate between the two copies is negatively correlated with the spacer size. A higher conversion rate will increase the identity percentage, and a higher deletion rate will tend to remove large repeats.

In conclusion, the properties which supported the model of iteration dynamics established in *S. cerevisiae* are shared by other eukaryotes. This suggests that the model could be extended to all eukaryotes.

The Case of *P. falciparum*: How Parasitism Influences the Genome Style

P. falciparum chromosomes exhibit a high level of redundancy as compared with similar-sized chromo-

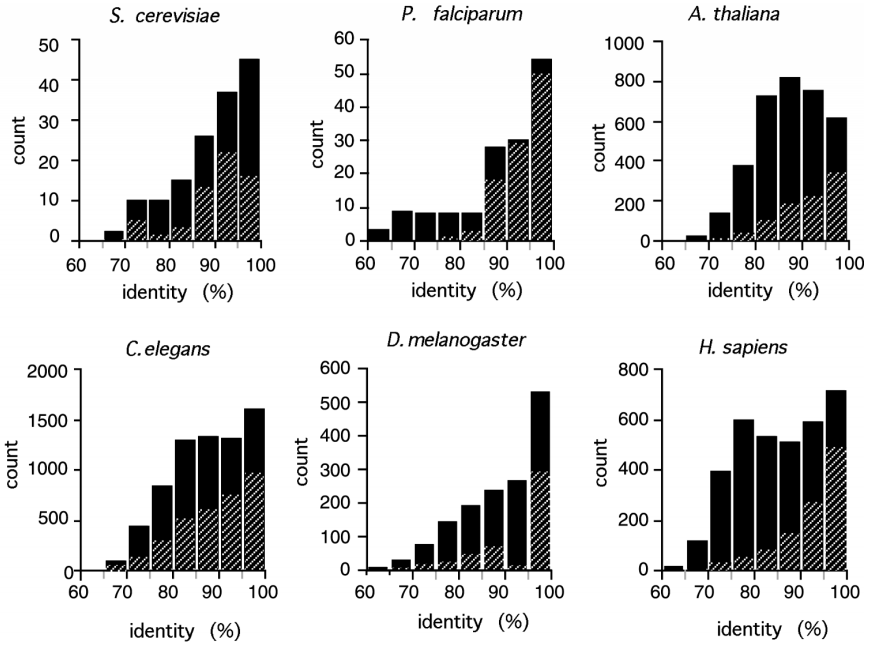


FIG. 3.—Distribution of the identity percentage for direct repeats. The histograms show the distribution of the identity percentage of a given species for distant direct repeats and CDRs. The hatched black boxes represent only the CDRs, and the plain black boxes represent the distant direct repeats. It can be shown using a Mann-Whitney rank test that, except for yeast, CDRs are more identical than distant direct repeats.

somes of *S. cerevisiae* (fig. 1), and their CDRs are extremely overrepresented: 74% have a spacer smaller than 1 kb (fig. 2). They are very identical (fig. 3) and very small (data not shown). However, no correlation between spacer, identity, and length can be highlighted (Table 2).

Two-thirds of the inverted repeats are located near the telomeres (one copy in each subtelomere), suggesting a peculiar history and a high exchange rate for these repeats. It was suggested that all subtelomeres exhibit a very plastic dynamics in *S. cerevisiae* (Pryde, Gorham, and Louis 1997) and in *H. sapiens* (Coleman, Baird, and Royle 1999). Their importance in the interchromosomal iteration process was demonstrated in *S. cerevisiae* (Coissac, Maillier, and Netter 1997).

All these observations are consistent with what was described previously: the highly repeated gene families and the special status of subtelomeres in *P. falciparum* (Gardner et al. 1998; Bowman et al. 1999).

Do These Observations Mean that This Ciliate Does Not Follow the Same Dynamics as the Other Eukaryotes?

P. falciparum is a human pathogenic parasite, the main agent of malaria. It has been reported that many bacterial pathogens exhibit a high redundancy level (Rocha, Danchin, and Viari 1999*b*) which has been related to high selective pressures for sequence variation. A significant number of repeats allows many recombination events, leading to a high plasticity of the genome, and then to a high evolution rate. As for these bacteria, the high redundancy level of *P. falciparum* could be a consequence of its parasitism.

Table 2
Computed Kendall Rank Correlations for CDRs of Each Species

	Close Direct Repeats[a]		Identity[b]		Length[b]	
Species	N	D_N	τ	P	τ	P
S. cerevisiae	60	5.0	−0.32	$<10^{-3}$	0.45	$<10^{-4}$
P. falciparum	100	49.8	−0.08	>0.05	0.06	>0.05
A. thaliana	889	23.9	−0.35	$<10^{-4}$	0.39	$<10^{-4}$
C. elegans	3,242	34.0	−0.31	$<10^{-4}$	0.24	$<10^{-4}$
D. melanogaster . . .	546	4.7	−0.36	$<10^{-4}$	0.41	$<10^{-4}$
H. sapiens	1,042	15.5	−0.30	$<10^{-4}$	0.33	$<10^{-4}$

[a] The number of CDRs (N) and the density in number (D_N) of these CDRs.

[b] The results of the tested correlations: the correlation coefficient (τ) and the probability value associated with (P) are given for the correlations between identity and spacer, and between length and spacer.

The quasi-absence of distant repeats and the absence of correlation indicate that there are almost only young repeats. The absence of correlation is, in this way, not caused by the absence of the mechanism leading to them but by too short a time of evolution. Population studies suggest that *P. falciparum* spread worldwide from a limited area (Rich and Ayala 2000). The absence of old repeats could be a consequence of the recent change in the ecological conditions of *P. falciparum*, associated with a burst of evolution. In conclusion, *P. falciparum* follows the same iteration dynamics as the other eukaryotes. However, because it is a recent parasite, its chromosomes are more repeated than those of the other eukaryotes (as a result of parasitism), and there are almost no ancient repeats (because of its recent emergence).

How Tandem Repeats Can Be Turned into Spaced Repeats

Intrachromosomal repeats, in our model, are mostly created in tandem (by recombination between sister chromatides or by replication slippage), and are turned into distant repeats by chromosomal rearrangements. Analyzing all the ending states after several rearrangements is difficult. However, it is interesting to examine all the theoretical resulting states obtained after only one rearrangement event. Three kinds of rearrangement have been taken into account (fig. 4): deletion of a part of the tandem, insertion of a sequence inside the tandem repeat, and inversion taking away a piece of the tandem. The insertion process can be the result of either the insertion of a transposable element or the reparation of a double-strand break by sequence conversion (Voelkel and Roeder 1990). Small inversions have been suggested to explain the evolution of the genes' order between *C. albicans* and *S. cerevisiae* (Seoighe et al. 2000), highlighting their role in genome dynamics.

If the model is valid, one should find the vestiges of tandem rearrangement in the chromosome sequences. Thus, we used the wublastn software (http://blast.wustl.edu) to look for paralogs of the spacers in the complete chromosomes. Only spacers with size between 50 bp and 10 kb and flanked by direct repeats were taken into account. It should be stressed that the queried databases were constructed for each species with complete chromosomes only (the same that we used for the detection of the repeats). A sequence was arbitrarily considered as a paralog of the spacer if the sequence length was at least 80% of the spacer length, and if the two sequences were identical by more than 80%. Using this approach, large insertions (fig. 4.2*a*) or some inversions (fig. 4.3*b*) can be undoubtedly identified, but small internal deletions and small internal insertions (fig. 4.1*b* and 4.2*b*) cannot be clearly differentiated. One should notice that deletion of an edge of a copy (fig. 4.1*a*) or a complete inversion of a copy (fig. 4.3*a*) cannot be detected by this method.

Results were sorted as a function of the number of paralogs detected in the chromosomes. For most spacers, no paralog was found. This has several possible reasons: (1) our criteria were very stringent, (2) the research was performed against the whole genome only for *S. cerevisiae* and *C. elegans*, and (3) we only detected paralogs for spacers issued from a recent unique event of rearrangement. Multiparalog families (when a spacer presented at least two paralogs) were separated because they give an idea of the relative transposition rate. All cases where the spacer had only one paralog have been analyzed more precisely as they appeared in figure 4.

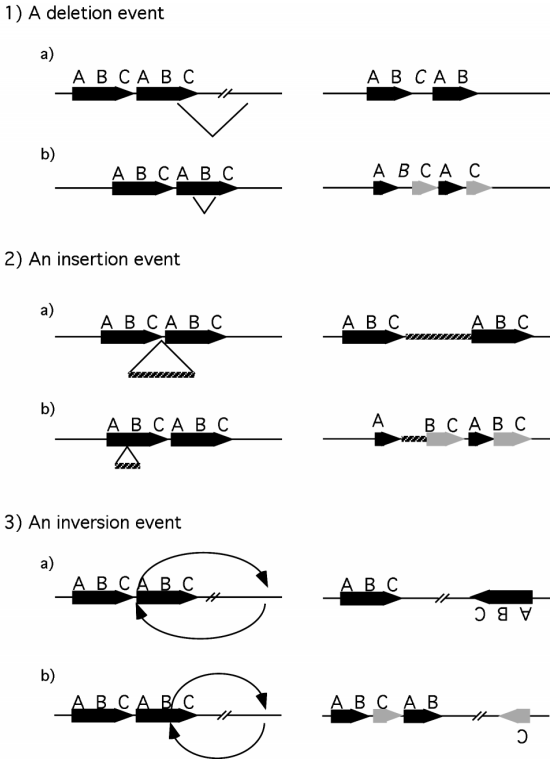


FIG. 4.—How tandem repeats can be turned into spaced repeats. This figure presents the three main ways to create a spacer between two tandem repeats: (1) a deletion event could occur inside the tandem repeats, leading to the creation of a spaced direct repeat (*a*) and two spaced direct repeats (*b*), (2) an insertion event could arise inside the tandem repeat, leading to one direct repeat with a spacer similar to another region in the genome (*a*) and two direct repeats (*b*), and (3) an inversion event could lead to one inverted repeat (*a*) or one direct repeat in which the spacer is an inverted repeat (*b*).

As shown in Table 3, all possible remnants of the tandem rearrangement were detected in the sequence of chromosomes. These observations indicate that the theoretical rearrangements arise in the genome history, reinforcing the model of the iteration dynamics.

A striking result was the overrepresentation of intrachromosomal direct paralogs in *C. elegans*. A detailed analysis of these paralogs revealed that they are mostly part of larger old tandem repeats. This observation has to be connected to the presence of large tandem repeats in the chromosomes of this species (i.e., a 600-kb repeat in the first chromosome), also recently described by Friedman and Hughes (2000). It seems probable that the

Table 3
Detected Paralogs for Spacers of Direct Repeats

Species	No Paralog	Only One Paralog[a]				At Least Two Paralogs	Total[c]
		Another Chromosome	Same Chromosome[b]				
			Close Direct	Direct	Inverted		
S. cerevisiae	20	1	1	—	—	6	28
P. falciparum	10	—	1	—	—	—	11
A. thaliana	1,318	11	46	17	6	77	1,475
C. elegans.	1,960	61	70	308	5	505	2,909
D. melanogaster[d] . . .	473	—	3	4	1	14	495
H. sapiens.	419	1	12	14	4	62	512

[a] The spacer and its paralog were either in two different chromosomes or in the same one.

[b] The spacer and its paralog were either in the same orientation and physically closer than 1 kb, in the same orientation but not close, or not in the same orientation.

[c] Blastn results obtained in paralogs research of the spacer sequences of direct repeats. Only spacers with a size between 50 bp and 10 kb were queried against the complete chromosomes of a given species. A sequence is considered as a paralog if its length is at least 80% of the spacer length and if the identity percentage is at least 80%.

[d] For *D. melanogaster*, the arms 2L and 2R (as well as 3L and 3R) were considered as the same chromosome and in the same orientation.

worm genome has exhibited an active process of intrachromosomal iteration.

All generic duplications of nonspecific DNA regions within the same chromosome or between two chromosomes were referred to in this study as iteration. However, this iteration process should be divided into at least two distinct mechanisms. The first is the creation of tandem repeats (by sister chromatide exchange or replication slippage), which creates (under our model) most of the intrachromosomal repeats. The second is the genesis of repeats (inter- or intrachromosomal) by a double-strand break repair. Actually, this repair can lead to duplication when the repair is associated with a conversion mechanism. This implies that the duplication process can at least be divided into four mechanisms: abnormal chromosome segregation (hyperploidization); transposition (transposable elements); sister chromatide exchange, replication slippage (tandem repeats and satellites), or both; and double-strand break repair (iteration by conversion).

Conclusions

Through this study of eukaryotes' intrachromosomal repeats, several biological results were highlighted. We extended our model, proposed for *S. cerevisiae*, to other eukaryote chromosomes (*S. cerevisiae*, *C. elegans*, *P. falciparum*, *A. thaliana*, *D. melanogaster*, and *H. sapiens*). This suggests that despite the differences in chromosomal properties, the iteration process follows globally the same dynamics in the eukaryote kingdom and thus has to be connected to structures and mechanisms shared by all eukaryote chromosomes.

The density of repeats number defines a genome style where the evolution rate results from iteration, deletion, rearrangement, and mutation. This rate is similar for all chromosomes within the same genome and is specific to each species. The main exception being the X chromosome of *C. elegans*, it suggests that exchanges between homologous chromosomes are important in the genesis of repeats. Thus, we propose that the genesis of tandem repeats is at least a consequence of exchange between homologous chromosomes.

Finally, we brought out the remnants of rearrangements of tandem repeats into spaced repeats. This suggests that tandem repeats, which can be easily created, are submitted to rounds of chromosomal rearrangements leading to the pattern of repeats observed today. Hence, repeats can be used to follow chromosome rearrangements and are markers of genome dynamics.

Acknowledgments

We thank I. Gonçalves, E. Rocha, D. Higuet, E. Maillier, J. Pothier, and A. Viari for their scientific help and their friendly support. This work was supported by grants from Association pour la Recherche sur le Cancer. E.C. and P.N. are members of Université Pierre et Marie Curie, Paris.

Literature Cited

Achaz, G., E. Coissac, A. Viari, and P. Netter. 2000. Analysis of intrachromosomal duplications in yeast *Saccharomyces cerevisiae*: a possible model for their origin. Mol. Biol. Evol. **17**:1268–1275.

Adams, M. D., S. E. Celniker, R. A. Holt et al. (195 co-authors). 2000. The genome of *Drosophila melanogaster*. Science **287**:2185–2195.

Amores, A., A. Force, Y. L. Yan et al. (13 co-authors). 1998. Zebrafish *hox* clusters and vertebrate genome evolution. Science. **282**:1711–1714.

Barnes, T. M., Y. Kohara, A. Coulson, and S. Hekimi. 1995. Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*. Genetics **141**:159–179.

Baudat, F., and A. Nicolas. 1997. Clustering of meiotic double-strand breaks on yeast chromosome III. Proc. Natl. Acad. Sci. USA **94**:5213–5218.

Bernardi, G. 2000. Isochores and the evolutionary genomics of vertebrates. Gene **241**:3–17.

Blanc, G., A. Barakat, R. Guyot, R. Cooke, and M. Delseny. 2000. Extensive duplication and reshuffling in the Arabidopsis genome. Plant Cell **12**:1093–1101.

Bowman, S., D. Lawson, D. Basham et al. (36 co-authors). 1999. The complete nucleotide sequence of chromosome 3 of Plasmodium falciparum. Nature **400**:532–538.

Britten, R. J. 1998. Precise sequence complementarity between yeast chromosome ends and two classes of just-sub-

telomeric sequences. Proc. Natl. Acad. Sci. USA **95**:5906–5912.

COISSAC, E., E. MAILLIER, and P. NETTER. 1997. A comparative study of duplications in bacteria and eukaryotes: the importance of telomeres. Mol. Biol. Evol. **14**:1062–1074.

COLEMAN, J., D. M. BAIRD, and N. J. ROYLE. 1999. The plasticity of human telomeres demonstrated by hypervariable telomeres repeat array that is located on some copies of 16p and 16q. Hum. Mol. Genet. **8**:1637–1646.

CONSORTIUM. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science **282**: 2012–2018.

DUNHAM, I., N. SHIMIZU, B. A. ROE et al. (239 co-authors). 1999. The DNA sequence of human chromosome 22. Nature **402**:489–495.

FRIEDMAN, R., and A. L. HUGHES. 2000. Gene duplication and the structure of eukaryotic genomes. Genome Res. **11**:373–381.

GARDNER, M. J., H. TETTELIN, D. J. CARUCCI et al. (27 co-authors). 1998. Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. Science **282**:1126–1132.

GOFFEAU, A., B. G. BARRELL, H. BUSSEY et al. (16 co-authors). 1996. Life with 6000 genes. Science **274**:546.

HATTORI, M., A. FUJIYAMA, T. D. TAYLOR et al. (62 co-authors). 2000. The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. Nature **405**:311–319.

HENIKOFF, S. 2000. Heterochromatin function in complex genomes. Biochem. Biophys. Acta **1470**:O1–O8.

HOLLAND, P. W. 1999. Gene duplication: past, present and future. Semin. Cell Dev. Biol. **10**:541–547.

HUGHES, A. L., J. DA SILVA, and R. FRIEDMAN. 2001. Ancient duplication did not structure the human *Hox*-bearing chromosomes. Genome Res. **11**:771–780.

JINKS, R. S., M. MICHELITCH, and S. RAMCHARAN. 1993. Substrate length requirements for efficient mitotic recombination in *Saccharomyces cerevisiae*. Mol. Cell. Biol. **13**:3937–3950.

KARLIN, S., and F. OST. 1985. Maximal segmental match length among random sequences from a finite alphabet. Pp. 225–243 *in* L. M. L. CAM and R. A. OLSHEN, eds. Proceedings of the Berkeley Conference in honor of Jerzy Neyman and Jack Kiefer, Vol. 1. Association for Computing Machinery, New York.

KURTZ, S., and C. SCHLEIERMACHER. 1999. REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics **15**:426–427.

LEUNG, M. Y., B. E. BLAISDELL, C. BURGE, and S. KARLIN. 1991. An efficient algorithm for identifying matches with errors in multiple long molecular sequences. J. Mol. Biol. **221**:1367–1378.

LIN, X., S. KAUL, S. ROUNSLEY et al. (39 co-authors). 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature **402**:761–768.

LLORENTE, B., A. MALPERTUY, C. NEUVEGLISE et al. (24 co-authors). 2000. Genomic exploration of the hemiascomycetous yeasts: 18. Comparative analysis of chromosome maps and synteny with *Saccharomyces cerevisiae*. FEBS Lett. **487**:101–112.

MASTERSON, J. 1994. Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. Science **264**: 421–424.

MAYER, K., C. SCHULLER, R. WAMBUTT et al. (234 co-authors). 1999. Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. Nature **402**:769–777.

OHNO, S. 1970. Evolution by gene duplication. Springer-Verlag, Heidelberg, Germany.

PEETERS, B. P. H., J. H. DE BOER, S. BRON, and G. VENEMA. 1988. Structural plasmid instability in *Bacillus subtilis*: effect of direct and inverted repeats. Mol. Gen. Genet. **212**: 450–458.

PRYDE, F. E., H. C. GORHAM, and E. J. LOUIS. 1997. Chromosome ends: all the same under their caps. Curr. Opin. Genet. Dev. **7**:822–828.

RICH, S. M., and F. J. AYALA. 2000. Population structure and recent evolution of *Plasmodium falciparum*. Proc. Natl. Acad. Sci. USA **97**:6994–7001.

ROBINSON-RECHAVI, M., O. MARCHAND, H. ESCRIVA, P. L. BARDET, D. ZELUS, S. HUGHES, and V. LAUDET. 2001. Euteleost fish genomes are characterized by expansion of gene families. Genome Res. **11**:781–788.

ROCHA, E. P., A. DANCHIN, and A. VIARI. 1999a. Analysis of long repeats in bacterial genomes reveals alternative evolutionary mechanisms in *Bacillus subtilis* and other competent prokaryotes. Mol. Biol. Evol. **16**:1219–1230.

ROCHA, E. P., A. DANCHIN, and A. VIARI. 1999b. Functional and evolutionary roles of long repeats in prokaryotes. Res. Microbiol. **150**:725–733.

RUBNITZ, J., and S. SUBRAMANI. 1984. The minimum amount of homology required for homologous recombination in mammalian cells. Mol. Cell Biol. **4**:2253–2258.

RUVKUN, G., and O. HOBERT. 1998. The taxonomy of developmental control in *Caenorhabditis elegans*. Science **282**: 2033–2041.

SEMPLE, C., and K. H. WOLFE. 1999. Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. J. Mol. Evol. **48**:555–564.

SEOIGHE, C., N. FEDERSPIEL, and T. JONES et al. (20 co-authors). 2000. Prevalence of small inversions in yeast gene order evolution. Proc. Natl. Acad. Sci. USA **97**:14433–14437.

SMITH, T. F., and M. S. WATERMAN. 1981. Identification of common molecular subsequences. J. Mol. Biol. **147**:195–197.

SURZYCKI, S. A., and W. R. BELKNAP. 2000. Repetitive-DNA elements are similarly distributed on *Caenorhabditis elegans* autosomes. Proc. Natl. Sci. USA **97**:245–249.

VINCENS, P., L. BUFFAT, C. ANDRE, J. P. CHEVROLAT, J. F. BOISVIEUX, and S. HAZOUT. 1998. A strategy for finding regions of similarity in complete genome sequences. Bioinformatics **14**:715–725.

VISION, T. J., D. G. BROWN, and S. D. TANKSLEY. 2000. The origins of genomic duplications in *Arabidopsis*. Science **290**:2114–2117.

VOELKEL, K., and G. S. ROEDER. 1990. Gene conversion tracts stimulated by HOT1-promoted transcription are long and continuous. Genetics **126**:851–867.

WOLFE, K. H., and D. C. SHIELDS. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387**:708–713.

MANOLO GOUY, reviewing editor

Accepted August 27, 2001

Origin and fate of repeats in bacteria

Achaz, G and Rocha, E P C and Netter, P and Coissac, E

Nous avons recherché les répétions intrachromosomiques dans 53 génomes complets de bactéries. Lors d'études précédentes, réalisées sur des chromosomes eucaryotes, nous avions proposé un modèle expliquant la dynamique des répétitions s'appuyant sur une genèse permanente de répétitions en tandem, suivi par un processus actif de délétion contrecarré par des réarrangements chromosomiques protégeant les duplications de la délétion. Cette étude réalisée à partir de génomes bactériens et archaebactériens montre que notre modèle s'applique aussi à ces deux règnes. Aussi le mécanisme de duplication doit être la conséquence de mécanismes très anciens partagés par les trois règnes. Nous avons de plus mis en évidence une forte corrélation négative entre le biais de composition en nucléotides et la densité de répétitions des génomes. Nous suggérons que dans les génomes fortement biaisés, des petites répétitions non issues de phénomènes de duplication sont fréquemment crées par hasard et que ces petites répétitions servent d'amorces aux mécanismes de duplication, expliquant ainsi les forts taux de répétions des grandes répétitions.

Nucleic Acids Research, 2002, Vol. 30, No. 13 **2987–2994**

Origin and fate of repeats in bacteria

G. Achaz*, E. P. C. Rocha[1,2], P. Netter and E. Coissac

Structure et Dynamique des Génomes, Institut Jacques Monod, Tour 43-44, 1° Étage, 4 Place Jussieu, F-75251 Paris Cedex 05, France, [1]Atelier de Bioinformatique, Université Pierre et Marie Curie, Paris, France and [2]URA2171, Unité GGB, Institut Pasteur, Paris, France

Received December 12, 2001; Revised April 12, 2002; Accepted May 8, 2002

ABSTRACT

We investigated 53 complete bacterial chromosomes for intrachromosomal repeats. In previous studies on eukaryote chromosomes, we proposed a model for the dynamics of repeats based on the continuous genesis of tandem repeats, followed by an active process of high deletion rate, counteracted by rearrangement events that may prevent the repeats from being deleted. The present study of long repeats in the genomes of Bacteria and Archaea suggests that our model of interspersed repeats dynamics may apply to them. Thus the duplication process might be a consequence of very ancient mechanisms shared by all three domains. Moreover, we show that there is a strong negative correlation between nucleotide composition bias and the repeat density of genomes. We hypothesise that in highly biased genomes, non-duplicated small repeats arise more frequently by random effects and are used as primers for duplication mechanisms, leading to a higher density of large repeats.

INTRODUCTION

DNA repeats can be defined as sequences sharing extensive similarity with other sequences of the same genome. It is usually supposed that repeats arise by successive duplications and several causal mechanisms, including hyperploidisation (even polyploidisation), tandem duplication, double-strand break repair by insertion or transposition, have been proposed to be involved. The underlying mechanisms are thought to act at different levels depending on the kingdom, or even on organism [i.e. polyploidisation has been proposed to explain the presence of large repeats in eukaryotes (1,2), but is probably absent in Archaea and Bacteria]. Once a repeat is created, it can be targeted by the recombination apparatus and be subject to deletion. Thus, genome size results from a balance between duplication and deletion events. The importance of deletion processes seems crucial in compact genomes, especially in those of intracellular endosymbionts or pathogens (3).

Usually, repeats in Bacteria are divided into two subclasses: low complexity repeats (sometimes mislabeled 'tandem repeats') and longer repeats (the centre of our interest). The first category is constituted of small oligonucleotides (typically ranging from mononucleotide to pentanucleotide in size) repeated many times in a head-to-tail configuration. These low complexity repeats, e.g. microsatellites, are very abundant in the genomes of eukaryotes, in which they have been widely studied (4). Although less abundant in bacterial and archaeal genomes (5), the mechanisms of their origin (6), their function (7), the consequences for genome dynamics (8) and the structural constraints imposed on the chromosome (9) have all been studied

Longer repeats include transposable elements, minisatellites (mostly in Eukarya), large tandem repeats and spaced repeats. DNA transposable elements (like IS) are widely distributed among the Archaea and Bacteria. As specific mechanisms for the duplication of mobile elements have been identified (10), such self-replicating elements have to be considered separately when the origin of repeats is analysed. However, they must be taken into account when the influence of repeats on genome stability is considered.

Several mechanisms have been proposed for the genesis of tandem repeats: slipped strand mispairing, unequal crossover (by homologous recombination), rolling circle and circle excision with reinsertion (11). Some of these mechanisms could also result in a tandem repeat deletion. These mechanisms render tandem repeats unstable, easy to create but also easy to delete. In contrast, distant repeats can almost only be deleted by homologous recombination and at the cost of large deletions of genetic material. As a consequence, they may persist more easily during genome evolution. Two mechanisms have been envisaged to create spaced repeats *ex nihilo*. The first, known as Campbell-like insertion, creates repeats by inserted exogenous sequences and has been proposed to explain the peculiar distribution of many repeats in *Bacillus subtilis* (12). The second, referred to as 'conversion' or 'insertion', repairs a double-strand break by copying a sequence sharing similarity with the edges of the broken sequence: this mechanism works either by break-induced replication or by gap repair (for reviews in yeast see 13,14).

The first question we tackled in this work concerns the origin of interspersed repeats (excluding transposable elements). Our previous studies (15,16) had led us to propose a model (Fig. 1) for the origin of eukaryote intrachromosomal repeats based on the permanent genesis of close direct repeats (CDR, repeats with copies separated by <1 kb). Since our model is compatible with all mechanisms, we do not assume any particular one for the creation of CDR. Newly created CDR are then subject to a strong rate of exchange (conversion and deletion). Experimental studies undertaken on *B.subtilis* (17) and *Escherichia coli* (18–20) have shown that the rate of illegitimate recombination is negatively correlated with the distance between the copies (spacer size) and positively correlated with repeat length. Recombination between close repeats tends to maintain

*To whom correspondence should be addressed. Tel: +33 1 44 27 76 94; Fax: +33 1 44 27 82 05; Email: achaz@ijm.jussieu.fr

2988 Nucleic Acids Research, 2002, Vol. 30, No. 13

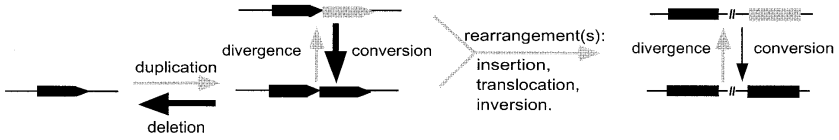


Figure 1. A model of interspersed repeats dynamics. In this model, interspersed repeats originate mainly from tandem repeats, which can be separated by further chromosomal rearrangements. In newly created repeats with a small spacer (i) the conversion rate is high, keeping the two copies identical and (ii) the deletion rate is also high, so that over a longer time scale only small repeats are retained. However, if one or more rearrangements (e.g. insertion, translocation and/or inversion) occur separating the two copies, both deletion and conversion rates decrease markedly. Both copies are then free to evolve.

neighbouring repeats identical (by conversion) but also to eliminate them (by deletion). At each round of exchange, both events are possible (although we ignore whether they are equally likely). If conversion can be followed by deletion, the opposite is not true: a deletion event cannot be followed by conversion. Over a long time, this will result in a bias in favour of deletions, with CDR disappearing sooner or later (depending on the relative rates of conversion and deletion). Thus, in the absence of strong selective pressure, long CDR are too unstable to persist, except if the copies are moved further apart by chromosomal rearrangements (i.e. insertion, translocation and inversion). In this case, the rate of illegitimate recombination will drop severely and the repeats may be maintained.

In this context, one expects CDR to be more similar than distant repeats, since either they are more recent or they are more subject to conversion. On the other hand, one expects that larger repeats will only survive fast deletion by frequent illegitimate recombination if they are placed distantly. Thus, under our model, CDR tend to have smaller and more identical repeats whereas distant repeats tend to be longer and less similar. This matches the observations we have made in eukaryote genomes, where repeats are both more identical and smaller when they are closer (15). The main goal of this work was to test if this model, first established in Eukarya, could be applied to Bacteria and Archaea.

The second focus of our attention concerns the factors influencing the dynamics of our model, i.e. rates of duplication, deletion and rearrangement. Here we analyse precisely the relation between the origin of tandem repeats and the genome composition biases. Duplication mechanisms typically require the pre-existence of a region of similarity. Levinson and Gutman (8) proposed that small non-duplicated repeats (afterwards referred to as repeats appearing by chance) are primers for mechanisms such as slipped strand mispairing, thus creating larger repeats. We have tried to analyse this proposition by deciphering the relations between repeat density and the relative frequencies of nucleotides in the chromosome.

MATERIALS AND METHODS

Data

We analysed the complete genomes of 40 Bacteria and 11 Archaea (Table 1). All sequences were extracted from GenBank (ftp://ftp.ncbi.nih.gov/genbank/Bacteria), except for those of *Pyrococcus furiosus*, downloaded from http://www.genome.utah.edu.

Construction of the repeats database

We followed the methodology previously developed to detect repeats in eukaryote genomes (15,16), but made an extra effort to detect smaller, but significant, repeats, since bacterial chromosomes are smaller. The methodology is described below and follows four main steps.

First step: detection of seeds. In this step, exact direct and inverse repeats (seeds) of 15 bp were detected using the REPuter software (21). Many seeds with lengths that are not statistically significant according to Karlin and Ost statistics were retained (22). The second step is intended to further extend these seeds into larger, non-strict repeats.

Second step: from seeds to repeats. Local alignment (23) is used to extend the edges of the seeds into larger repeats. Except for the construction of the score matrix, the extension process is the same we used to analyse eukaryote chromosomes (16). This method produces non-exact repeats by extending a seed on both sides when similarity is high. To do so, we used an algorithm based on a local alignment procedure (23).

Nucleotide frequencies differ widely between species genomes, from 25 to 75% (24). Therefore, if an identity matrix is used for the local alignment, seeds of the same size in chromosomes with a very unbalanced distribution of nucleotides (e.g. *Ureaplasma urealiticum* where A ≈ T ≈ 0.37 and C ≈ G ≈ 0.13) tend to produce larger repeats than in genomes with equal frequencies (e.g. *E.coli*). In order to avoid this effect, we used an empirical scoring matrix for each chromosome, which takes into account its specific composition. These matrices provide a better score for matches between rare nucleotides:

$\text{match}_{i/i} = 100 \times (1 - p_i^2)$; $\text{match}_{N/i} = 25$

$\text{mismatch}_{i/j} = -100 \times (1 - p_i \times p_j)$; $\text{gap}_{open} = -400$; $\text{gap}_{ext} = -100$

where p_i is the frequency of nucleotide i. By building these matrices for all species, we observed scores for matches ranging from 86 to 98 and scores for mismatches ranging from –98 to –86. Thus the score of gap_{open} is always less than 4× mismatch and the score of gap_{ext} always less than 1× mismatch. We also tried other matrices that gave similar results.

Third step: removing repeats that are not statistically significant. Since seeds are rather small, many repeats may not have statistically significant lengths. To remove these non-significant repeats, we built, for each chromosome, 10 additional random chromosomes by shuffling it with respect to its trinucleotide composition (Markov chains of order 2). In these random sequences, repeats were detected as in real sequences (steps 1 and 2). Afterwards, we built a distribution of observed alignment scores from the set of repeats detected in the 10 random chromosomes. We then defined a threshold of significance, corresponding to 0.001 of this distribution. Below this minimal score (S_{min}), repeats were regarded as non-significant and removed from further analysis. S_{min} depends essentially on the size and

Table 1. Organisms analysed

40 Bacteria	16 Proteobacteria	3 alpha subdivision	*Caulobacter crescentus*	*Cacr*
			Mesorhizobium loti	*Melo*
			Rickettsia prowazekii	*Ripr*
		2 beta subdivision	*Neisseria meningitidis MD58*	*NemeM*
			Neisseria meningitidis Z2491	*NemeZ*
		8 gamma subdivision	*Buchnera species*	*Busp*
			Escherichia coli K12	*EscoK*
			Escherichia coli O157 H7	*EscoO*
			Haemophilus influenzae	*Hain*
			Pasteurella multocida	*Pamu*
			Pseudomonas aeruginosa	*Psae*
			Vibrio cholerae	*Vich_1/Vich_2*
			Xylella fastidiosa	*Xyfa*
		3 delta and epsilon subdivision	*Campylobacter jejuni*	*Caje*
			Helicobacter pylori 26695	*Hepy*
			Helicobacter pylori J99	*HepyJ*
	10 Low G+C Gram-Positive Bacteria	2 Streptococcaceae	*Streptococcus pyogenes*	*Stpy*
			Lactococcus lactis	*Lala*
		2 Staphylococcus	*Staphylococcus aureus Mu50*	*StauM*
			Staphylococcus aureus N315	*StauN*
		2 Bacillus	*Bacillus halodurans*	*Baha*
			Bacillus subtilis	*Basu*
		4 Mycoplasmataceae	*Mycoplasma genitalium*	*Myge*
			Mycoplasma pneumoniae	*Mypn*
			Mycoplasma pulmonis	*Mypu*
			Ureaplasma urealyticum	*Urur*
	3 High G+C Gram-Positive Bacteria	3 Mycobacteriaceae	*Mycobacterium leprae*	*Myle*
			Mycobacterium tuberculosis CDC1551	*MytuC*
			Mycobacterium tuberculosis HR7Rv	*MytuH*
	5 Chlamydiales	5 Chlamydiaceae	*Chlamydia pneumoniae AR39*	*ChpnA*
			Chlamydia pneumoniae CWL029	*ChpnC*
			Chlamydia pneumoniae J138	*ChpnJ*
			Chlamydia muridarum	*Chmu*
			Chlamydia trachomatis	*Chtr*
	2 Spirochaetales	2 Spirochaetaceae	*Borrelia burgdorferi*	*Bobu*
			Treponema pallidum	*Trpa*
	1 Cyanobacteria	1 Chroococcales	*Synechocystis sp.*	*Sysp*
	1 Thermus/Deinococcus	1 Deinococcus	*Deinococcus radiodurans*	*Dera_1/Dera_2*
	1 Thermotogales	1 Thermotoga	*Thermotoga maritima*	*Thma*
	1Aquificales	1 Aquificaceae	*Aquifex aeolicus*	*Aqae*
11 Archea	2 Crenarchaeota	1 Aeropyrum	*Aeropyrum pernix*	*Aepe*
		1 Sulfolobaceae	*Sulfolobus solfataricus*	*Suso*
	9 Euryarchaeota	1 Archaeoglobaceae	*Archaeoglobus fulgidus*	*Arfu*
		1 Halobacteriaceae	*Halobacterium species*	*Hasp*
		1 Methanobacteriaceae	*Methanobacterium thermoautotrophicum*	*Meth*
		1 Methanococcaceae	*Methanococcus jannaschii*	*Meja*
		3 Thermococcaceae	*Pyrococcus abyssi*	*Pyab*
			Pyrococcus furiosus	*Pyfu*
			Pyrococcus horikoshii	*Pyho*
		2 Thermoplasmaceae	*Thermoplasma acidophilum*	*Thac*
			Thermoplasma volcanium	*Thvo*

composition of the genome (and naturally on our choice of scoring system) and ranges from 2052 (*Chlamydia pneumoniae*) to 2258 (*Mycoplasma pulmonis*). Using score (*S*), length (*L*) and identity (*Id*), characteristics of some pertinent repeats from these two organisms are given with more details: (i) for *C.pneumoniae*, the smallest score corresponds to $S = 2052$, $L = 36$ and $Id = 80.6\%$; the medians of the distributions being $S = 4505$, $L = 220$ and $Id = 63.1\%$; (ii) for *M.pulmonis*, the smallest score corresponds to $S = 2258$, $L = 82$, $Id = 71.7\%$; the medians being $S = 3005$, $L = 90$ and $Id = 68.9\%$.

Fourth step: determining family sizes. At this stage, all significant repeats are given as a series of pairs. However, many repeats are organised in multicopy families (i.e. IS and rRNA operons). Hence, we developed a procedure to detect such multicopy families in our data set.

To do so, we built, for each chromosome, a map in which each position is linked to its '*n*-plication' degree: unique, duplicated, triplicated, etc. These maps were built by counting, for each chromosome position, the number of times this position is found in repeats (direct and inverted ones were pooled

2990 *Nucleic Acids Research, 2002, Vol. 30, No. 13*

together). Each pair was then associated with the map and the family size of each repeat was determined.

Density of repeats

In order to characterise the repeats, we used two measures of density, the density in number and the density in length. They are defined as:

D_N = no. of copies/size of chromosome (Mb)

$D_L = 100 \times$ [size of repeat sequence (bp)/size of chromosome (bp)]

Nucleotide complexity

Complexity is frequently used as a compact measure of the difference of the nucleotide distribution to equal repartition. In this context, information entropy has been proposed to describe biases of mononucleotide distributions (25):

$$H = -\sum_{i=A}^{T} p_i \log_4 p_i$$

where p_i is the frequency of nucleotide i. If a sequence exhibits an equal repartition of its four nucleotides (maximum complexity), its entropy is 1. In bacterial chromosomes it ranges from 0.91 to 1.

Proportion of CDR

CDR were originally defined as repeats with a distance between their two copies of <1 kb. We estimated the proportion of CDR expected if repeats are spread randomly along a chromosome. The proportion of CDR is calculated as the ratio between the number of CDR and the total number of repeats. Two cases were taken into account. (i) If the chromosome is circular, the largest spacer size is $L/2$, where L is the chromosome length. The distribution of spacer size is constant from 0 to $L/2$. So, the proportion of CDR in a circular chromosome is $1000 \times 2/L$. (ii) If the chromosome is linear, the largest spacer size is L and the spacer distribution decreases linearly from 0 to L. Using the intercept theorem of Thales (or any analytical demonstration), it could easily be demonstrated that the proportion of CDR is $1000/L \times (2 - 1000/L)$.

RESULTS AND DISCUSSION

What repeats have we detected?

We have found a large number of repeats in most (but not all) bacterial genomes (Table 2). In order to characterise these repeats, we used two measures of repeat density, D_N and D_L (see Materials and Methods). As expected, both densities were positively correlated ($\tau = 0.63$, $P < 10^{-4}$, Kendall τ rank test): a chromosome with many repeats also exhibits a high proportion of duplications in its chromosome. However, the biological interpretation of these measures may be quite different: D_N can be assimilated to the rate of amplification (a balance between duplication and deletion processes) and D_L to the history of the chromosomes, a measure of the redundancy tolerated by a chromosome. Thus, D_N and D_L should be analysed in parallel as they give complementary information on chromosomal redundancy. The data in Table 2 brings to the fore two issues. (i) Chromosomes of related organisms often exhibit similar densities of repeats: both *Chlamydia trachomatis* strains, the three *C.pneumoniae* strains, the three *Pyrococcus* strains, both *Mycobacterium tuberculosis* strains, both *Staphylococcus aureus* strains, both *Neisseria meningitidis* strains and both *Helicobacter pylori* strains. However, exceptions do exist. *Escherichia coli* O157:H7 is more repeated than K12, in agreement with previous observations (26). Also, when we broaden the phylogenetic range, we observe that the four *Mycoplasma* spp. show very different densities (D_N and D_L), indicating fast divergence, possibly due to their rudimentary

Table 2. Densities of repeats

Species[b]	Size[c] (Mb)	D_N[d]			D_L[e]		
		D_N	D_{N2}	D_{N2}/D_N	D_L	D_{L2}	D_{L2}/D_L
Cacr[a]	4.0	1085.7	667.2	0.61	19.8	11.6	0.58
Melo[a]	7.0	987.1	543.8	0.55	20.8	12.4	0.60
Ripr	1.1	206.0	180	0.87	2.4	2.2	0.90
NemeM[a]	2.3	472.2	211.2	0.45	20.4	6.8	0.33
NemeZ[a]	2.2	461.0	218.8	0.47	19.1	6.0	0.32
Busp	0.6	561.9	178	0.32	11.3	5.3	0.47
EscoK[a]	4.6	261.3	151.8	0.58	11.4	5.9	0.52
EscoO[a]	5.5	322.3	153.6	0.48	17.9	6.9	0.38
Hain[a]	1.8	517.5	215.2	0.42	8.9	4.9	0.55
Pamu[a]	2.3	275.1	133	0.48	6.2	3.6	0.58
Psae[a]	6.3	1663.1	819	0.49	28.2	15.6	0.55
Vich_1[a]	3.0	150.3	96.6	0.64	5.5	2.7	0.48
Vich_2[a]	1.1	71.8	54	0.75	10.1	3.4	0.33
Xyfa[a]	2.7	229.9	119.4	0.52	13.7	8.5	0.62
Caje	1.6	503.8	359.4	0.71	9.1	5.8	0.64
Hepy[a]	1.7	463.5	259	0.56	11.5	7.1	0.62
HepyJ[a]	1.6	481.8	288.4	0.60	10.7	7.2	0.67
Stpy[a]	1.9	203.5	119.8	0.59	7.2	3.7	0.51
Lala[a]	2.4	528.8	257	0.49	14.1	7.4	0.52
StauM[a]	2.9	368.0	212	0.58	13.1	7.2	0.55
StauN[a]	2.8	354.9	199.6	0.56	13.4	6.3	0.47
Baha[a]	4.2	225.8	130.8	0.58	9.4	3.6	0.38
Basu	4.2	264.3	185	0.70	9.8	6.7	0.69
Myge	0.6	287.9	131	0.46	6.9	2.6	0.37
Mypn	0.8	400.5	95.6	0.24	24.0	5.5	0.23
MypuC[a]	1.0	632.9	292.6	0.46	18.3	10.3	0.56
Urur	0.8	530.8	367.2	0.69	14.0	11.1	0.8
Myle	3.3	232.9	186.6	0.80	6.1	3.4	0.56
MytuC[a]	4.4	496.2	273	0.55	21.0	8.7	0.41
MytuH[a]	4.4	501.2	265.6	0.53	21.3	8.7	0.41
ChpnA	1.2	132.5	84.6	0.64	4.9	3.1	0.63
ChpnC	1.2	144.7	96	0.66	4.9	3.0	0.62
ChpnJ	1.2	132.7	86.4	0.65	4.6	2.8	0.61
Chmu	1.1	86.0	56.2	0.65	3.4	2.0	0.58
Chtr	1.0	68.1	53.8	0.79	2.3	2.2	0.96
Bobu	0.9	295.4	226.2	0.77	4.6	3.6	0.77
Trpa	1.1	126.5	77.4	0.61	3.9	2.4	0.6
Sysp[a]	3.6	402.4	254.6	0.63	9.1	5.1	0.55
Dera_1[a]	2.7	871.4	618.4	0.71	14.0	8.2	0.59
Dera_2[a]	0.4	303.1	179.4	0.59	7.9	5.7	0.73
Thma	1.9	187.6	111.8	0.60	6.6	3.8	0.58
Aqae	1.6	226.9	176.6	0.78	5.3	4.3	0.81
Aepe	1.7	226.4	136.6	0.60	4.1	2.5	0.61
Suso[a]	3.0	577.8	157	0.27	25.0	9.1	0.36
Arfu	2.2	349.8	202	0.58	10.3	6.8	0.66
Hasp[a]	2.0	1366.8	712	0.52	19.8	12.7	0.64
Meth[a]	1.8	302.6	229.6	0.76	9.2	6.8	0.74
Meja[a]	1.7	773.0	408.4	0.53	13.5	7.9	0.58
Pyab[a]	1.8	186.4	132.6	0.71	4.7	3.4	0.72
Pyfu	1.9	181.3	141.6	0.78	7.5	4.4	0.59
Pyho	1.7	198.5	123.2	0.62	5.3	3.5	0.66
Thac[a]	1.6	118.2	74.2	0.63	2.9	2.1	0.71
Thvo	1.6	162.1	83.2	0.51	5.1	3.1	0.61

[a]Chromosomes containing transposable elements.
[b]Abbreviations and order are those used in Table 1.
[c]Size of the chromosome (in Mb)
[d]D_N, number of copies per Mb. D_{N2} is the density for two-copy CDR only.
[e]D_L is the proportion of the chromosome included in repeats. D_{L2} is the proportion of the chromosome included in two-copy CDR only.

Nucleic Acids Research, 2002, Vol. 30, No. 13 **2991**

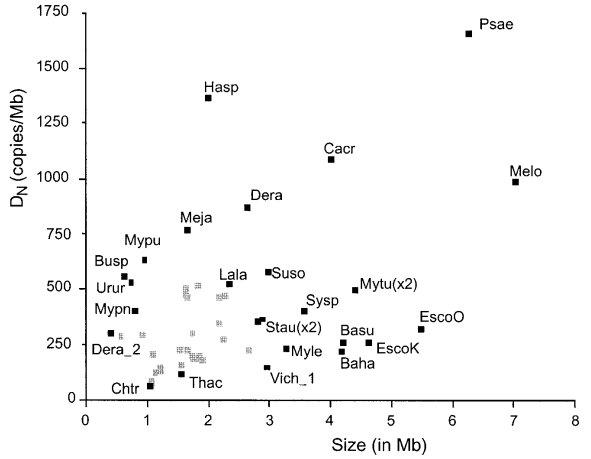


Figure 2. Repeat density as a function of chromosome size. Plot of D_N as a function of chromosome size. This figure illustrates the positive correlation between D_N and chromosome size.

repair mechanisms and to the selective pressure for variation in these pathogens (27). (ii) Both D_N and D_L exhibit a positive correlation with chromosome size ($\tau = 0.24$, $P < 10^{-3}$ for D_N and $\tau = 0.37$, $P < 10^{-4}$ for D_L),. These observations are in good agreement with previous observations on parts of both bacterial genomes and eukaryote genomes (16,28) (Fig. 2).

Since we were interested in the repeats' origins and in the supposition that it proceeds by duplication, we determined the proportions of two-copy repeats (and respective densities D_{N2} and D_{L2}) among all repeats (Table 2). As expected, D_{N2} is positively correlated with D_N ($\tau = 0.77$, $P < 10^{-4}$) and D_{L2} with D_L ($\tau = 0.73$, $P < 10^{-4}$). It could be noticed that, in contrast to eukaryote genomes in which D_{N2} is similar for chromosomes of the same species (16), densities varied between the two chromosomes of *Deinococcus radiodurans* and also between the two chromosomes of *Vibrio cholerae*.

Chromosomes containing transposable elements exhibit lower D_{N2}/D_N and D_{L2}/D_L ratios ($P < 0.01$, Mann–Whitney rank tests). Since transposable elements are mostly multicopy families, this can be easily understood. We observed few exceptions (low ratios in the absence of transposable elements), involving small genomes and, in particular, *Mycoplasma genitalium* and *Mycoplasma pneumoniae*. These repeats are associated with the immunodominant proteins of these genomes and are related to antigenic and tissue tropism variation (27).

Did interspersed repeats originate from tandems?

In order to test whether our model holds for Bacteria and Archaea we have tested its four major predictions. If interspersed repeats originate massively from tandem repeats, one might expect that (i) direct repeats are more numerous than inverted ones and that (ii) CDR are in large excess. Since the exchange rate between CDR is expected to be negatively correlated with spacer size and positively correlated with repeat length there should be (iii) a negative correlation between repeat similarity and spacer size and (iv) a positive correlation between repeat length and spacer size. Since we are interested in the origin of repeats, we decided to analyse only two-copy repeats further. This removed all low complexity repeats from our data set. Based on the annotations, we show

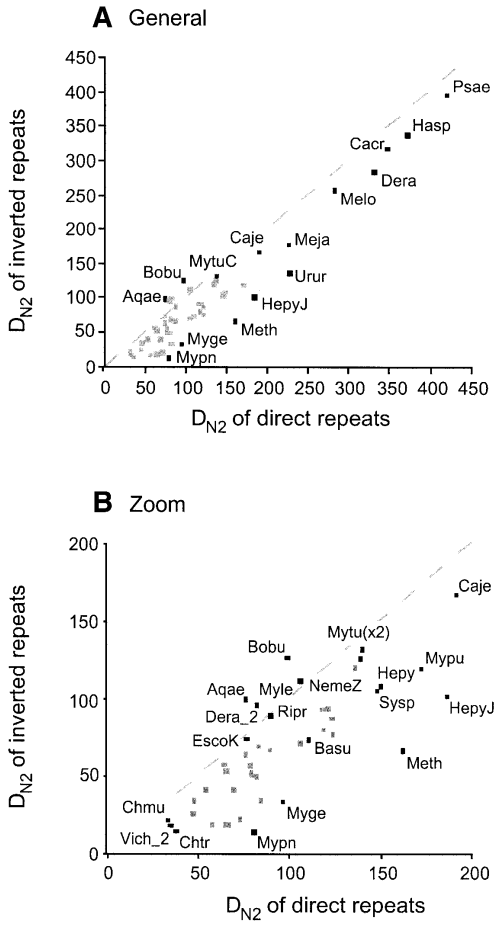


Figure 3. Densities of inverted repeats versus direct repeats. For each of the 53 chromosomes, we plotted the densities in number (D_{N2} = two-copy number/size in Mb) of inverted repeats as a function of D_{N2} of direct repeats. Because of the large difference in densities between genomes, two scales have been used. Abbreviations of species used in this figure correspond to those described in Table 1. The density of direct repeats is generally greater than the density of inverted repeats, but both are of the same order of magnitude.

that repeats located at least half in rRNA, tRNA or functional transposase represent ≤5% of our two-copy repeats, except for *C.trachomatis* (14%, four of 28) and the second chromosome of *V.cholerae* (7%, two of 29) (data not shown).

Direct repeats are more numerous than inverted ones. The large majority of the chromosomes (47 of 53) exhibit a higher density of two-copy direct repeats as compared with inverted ones ($P < 0.001$, binomial test), although sometimes the relative difference is not very high (Fig. 3). It is worth noticing that the two chromosomes that exhibit the largest excess of direct repeats are *M.genitalium* and *M.pneumoniae*. This is due to the previously described repeats located inside the adhesin genes.

CDR are over-represented. We estimated the numbers and densities of two-copy CDR, N_{2CDR} and D_{N2CDR}, respectively, and the theoretical number of CDR as a function of the number of direct repeats in linear and circular chromosomes (see

2992 *Nucleic Acids Research, 2002, Vol. 30, No. 13*

Table 3. Close direct repeats

Species[a]	N_{2CDR}[b]	p[c]	D_{N2cdr}[d]
Cacr	37	$<10^{-4}$	18.4
Melo	71	$<10^{-4}$	20.2
Ripr	4	$<10^{-4}$	7.2
NemeM	33	$<10^{-4}$	29.0
NemeZ	26	$<10^{-4}$	23.8
Busp	1	0.101	3.2
EscoK	23	$<10^{-4}$	10.0
EscoO	34	$<10^{-4}$	12.4
Hain	22	$<10^{-4}$	24.0
Pamu	23	$<10^{-4}$	20.4
Psae	51	$<10^{-4}$	18.2
Vich_1	12	$<10^{-4}$	8.2
Vich_2	4	$<10^{-4}$	7.4
Xyfa	46	$<10^{-4}$	34.4
Caje	16	$<10^{-4}$	19.6
Hepy	28	$<10^{-4}$	33.6
HepyJ	35	$<10^{-4}$	42.6
Stpy	14	$<10^{-4}$	15.2
Lala	26	$<10^{-4}$	22.0
StauM	30	$<10^{-4}$	20.8
StauN	29	$<10^{-4}$	20.6
Baha	26	$<10^{-4}$	12.4
Basu	31	$<10^{-4}$	14.8
Myge	7	$<10^{-4}$	14.2
Mypn	12	$<10^{-4}$	29.4
Mypu	17	$<10^{-4}$	35.2
Urur	12	$<10^{-4}$	32.0
Myle	19	$<10^{-4}$	11.6
MytuC	34	$<10^{-4}$	15.4
MytuH	33	$<10^{-4}$	15.0
ChpnA	12	$<10^{-4}$	19.6
ChpnC	12	$<10^{-4}$	19.6
ChpnJ	10	$<10^{-4}$	16.2
Chmu	6	$<10^{-4}$	11.2
Chtr	5	$<10^{-4}$	9.6
Bobu	9	$<10^{-4}$	19.8
Trpa	11	$<10^{-4}$	19.4
Sysp	29	$<10^{-4}$	16.2
Dera_1	37	$<10^{-4}$	28.0
Dera_2	5	$<10^{-4}$	24.2
Thma	5	$<10^{-4}$	5.4
Aqae	4	$<10^{-4}$	5.2
Aepe	10	$<10^{-4}$	12.0
Suso	18	$<10^{-4}$	12.0
Arfu	15	$<10^{-4}$	13.8
Hasp	22	$<10^{-4}$	21.8
Meth	71	$<10^{-4}$	91.0
Meja	24	$<10^{-4}$	28.8
Pyab	10	$<10^{-4}$	11.4
Pyfu	8	$<10^{-4}$	8.4
Pyho	11	$<10^{-4}$	12.6
Thac	4	$<10^{-4}$	5.2
Thvo	4	$<10^{-4}$	5.0

[a]Abbreviations and order are those used in Table 1.
[b]Observed number of two-copy CDR.
[c]Probability of finding N_{2CDR} or more under a random model. In the random model, one can estimate the probability of finding at least N_{2CDR} in N_2 two-copy direct repeats. This probability is $1 - B(0) + \ldots + B(N_{CDR} - 1)$, where $B(n)$ is the probability of finding n CDR in N direct repeats using a binomial law where the frequency of CDR is $2000/L$ for circular chromosomes and $2000/L - (1000/L)^2$ for linear ones (L = chromosome length). At an α risk of 10^{-4}, we assumed that 52/53 chromosomes are over-represented in CDR (with a risk of 0.005 of getting one or more false positives).
[d]Density in number (copies/Mb) for two-copy CDR.

Materials and Methods). As predicted by the model, CDR are over-represented in all chromosomes, taking into account the number of repeats (Table 3). The only exception is *Buchnera* sp., for which there are few CDR repeats, but it is unclear if this is a statistical artifact or has biological meaning. The *Buchnera* sp. genome is thought to be undergoing reductive evolution (3) and lacks an evident RecA homologue (29). Further, there is evidence that intracellular bacteria are subject to weaker selection (30). Thus, the absence of CDR could be the result of the reductive evolution process. Even if CDR are created, selection will not prevent them from being deleted. This deletion could arise easily since CDR deletion is mainly RecA independent.

Identity and length are constrained by spacer size. We looked for correlations between identity and spacer size within two-copy CDR for species in which there were at least 20 CDR (24 chromosomes). In 18 chromosomes identity was significantly negatively correlated with spacer size ($P < 0.01$, Table 4). In order to extend our analysis, we also took into account multi-copy repeats for chromosomes with less than 20 two-copy CDR or for those exhibiting a non-significant correlation for two-copy CDR (17 + 6 chromosomes). However, because the number of couples increases when families become very large [$c = n \times (n - 1)/2$, where c is the number of couples and n the number of copies], we retained only repeats with between two and five copies. This test identified significant positive correlations for 15 additional chromosomes ($P < 0.01$). Thus, out of the 41 chromosomes tested, 33 exhibited a significant negative correlation between identity and spacer size. Table 4 suggests that many others are weakly correlated.

Correlations between length and spacer size were tested under the same conditions as for identity (Table 5) and were also in agreement with the model. A negative correlation was found in 24 of the 41 chromosomes at $P < 0.01$ and in nine further chromosomes at a less significant α level ($P < 0.05$). Although very significant, these results are weaker than for the correlation between identity and spacer size and this deserves some comment. In the model, interspersed repeats are mostly created as identical tandem repeats, but their size can vary. Successive rounds of recombinational exchange constrain these repeats to be both highly identical and small due to the deletion bias mentioned above. Therefore, while the conversion process only maintains the pre-existing characteristics of the repeats (a high identity), the deletion process establishes an additional new constraint (small length). It is then conceivable that more rounds of exchange are required to establish the correlation between length and spacer size, thereby justifying weaker correlations.

Is tandem repeat creation modulated by chromosomal characteristics?

Since the previous results suggest the adequateness of our model, we proceeded to test the influence of chromosomal features on the duplication process, and in particular of nucleotide composition biases. Bacterial chromosomes exhibit large differences in their nucleotide composition, especially in terms of G + C composition, which can vary from 25 to 75% (24). We used the information entropy to measure the composition bias and found a significant negative correlation between entropy (and then composition bias) and the density of two-copy repeats, D_{N2} ($\tau = -0.34$, $P < 10^{-3}$, Fig. 4), as well as with total repeat densities, D_N ($\tau = -0.34$, $P < 10^{-3}$, Fig. 4). One would expect more biased random chromosomes to be more repetitive, since they use a subset of the possible symbols more frequently. However, our methodology to search for repeats already tackles this effect: we determined threshold scores

Nucleic Acids Research, 2002, Vol. 30, No. 13 **2993**

Table 4. Correlations between identity and spacer size for CDR

2-copy CDR[a]				2-, 3-, 4-, and 5-copy CDR[e]			
Species[b]	CDR	τ[c]	p[d]	Species[b]	CDR	τ[c]	p[f]
Baha	26	-0.31	0.013	*Aepe*	20	-0.06	0.371
Basu	31	-0.42	$<10^{-3}$	*Arfu*	38	-0.33	0.002
Cacr	37	-0.10	0.190	*Baha*	57	-0.38	$<10^{-3}$
Dera_1	37	-0.24	0.021	*Dera_1*	64	-0.29	$<10^{-3}$
EscoK	23	-0.54	$<10^{-3}$	*Bobu*	21	-0.57	$<10^{-3}$
EscoO	34	-0.50	$<10^{-3}$	*Cacr*	63	-0.27	$<10^{-3}$
Hain	22	-0.73	$<10^{-3}$	*Caje*	31	-0.39	0.001
Hasp	22	-0.17	0.141	*Chpn*	24	-0.34	0.011
Hepy	28	-0.21	0.066	*ChpnJ*	21	-0.31	0.024
HepyJ	35	-0.37	$<10^{-3}$	*Hasp*	44	-0.15	0.071
Lala	26	-0.51	$<10^{-3}$	*Hepy*	88	-0.15	0.022
Meja	24	-0.07	0.318	*Meja*	63	-0.19	0.015
Melo	71	-0.33	$<10^{-3}$	*Myge*	41	-0.15	0.090
Meth	71	-0.41	$<10^{-3}$	*Myle*	26	-0.72	$<10^{-3}$
Mytu	34	-0.52	$<10^{-3}$	*Mypn*	38	-0.26	0.010
MytuC	33	-0.48	$<10^{-3}$	*MypuC*	63	-0.40	$<10^{-3}$
NemeM	33	-0.43	$<10^{-3}$	*Pyab*	21	-0.39	0.008
NemeZ	26	-0.41	$<10^{-3}$	*Pyho*	21	-0.47	0.002
Pamu	23	-0.50	0.002	*Stpy*	42	-0.35	$<10^{-3}$
Psae	51	-0.31	$<10^{-3}$	*Suso*	40	-0.42	$<10^{-3}$
StauM	30	-0.34	0.004	*Sysp*	77	-0.32	$<10^{-3}$
StauN	29	-0.32	0.007	*Trpa*	22	-0.49	$<10^{-3}$
Sysp	29	-0.27	0.021	*Vich_1*	61	-0.10	0.130
Xyfa	46	-0.53	$<10^{-3}$				

[a]24 chromosomes with more than 20 two-copy CDR were used to test correlations.
[b]Abbreviations are those used in Table 1.
[c]Coefficients of Kendall τ rank tests between spacer size and identity for CDR.
[d]Probability associated with the Kendall τ rank tests. We assumed that, at an α risk of 0.01, 18/24 correlations are significant (with a risk of 0.21 of getting at least one false positive and of 0.02 of getting at least two false positives).
[e]We used two-copy, three-copy, four-copy and five-copy CDR to test 17 new chromosomes and re-test the six non-significant ones, where the two-copy CDR were less than 20 or were the tested correlation was $P < 0.01$.
[f]Probability associated with the Kendall τ rank tests. We assumed that, at an α risk of 0.01, 15/23 correlations are significant (with a risk of 0.21 of getting at least one false positive and of 0.02 of getting at least two false positives).

Table 5. Correlations between length and spacer size for CDR

2-copy CDR[a]				2-, 3-, 4-, and 5-copy CDR[e]			
Species[b]	CDR	τ[c]	p[d]	Species[b]	CDR	τ[c]	p[f]
Baha	26	0,46	$<10^{-3}$	*Aepe*	20	0,19	0.134
Basu	31	0,36	0,002	*Arfu*	38	0,39	$<10^{-3}$
Cacr	37	0.18	0.061	*Bobu*	21	0,50	$<10^{-3}$
Dera_1	37	0.16	0.089	*Cacr*	63	0,27	0.001
EscoK	23	0,50	$<10^{-3}$	*Caje*	31	0,30	0.009
EscoO	34	0,39	$<10^{-3}$	*Chpn*	24	-0,08	0.310
Hain	22	0,38	0,007	*ChpnJ*	21	0,01	0.488
Hasp	22	0.10	0.276	*Dera_1*	64	0,10	0.134
Hepy	22	0.28	0.018	*Hasp*	44	0,07	0.254
HepyJ	35	0,47	$<10^{-3}$	*Hepy*	88	0,25	$<10^{-3}$
Lala	26	0,26	0.033	*Lala*	66	0.18	0.019
Meja	24	0.12	0.212	*Meja*	63	0,19	0.015
Melo	71	0,30	$<10^{-3}$	*Myge*	41	-0,01	0.451
Meth	71	0,23	0.002	*Myle*	26	0,56	$<10^{-3}$
Mytu	34	0,41	$<10^{-3}$	*Mypn*	38	0,18	0.059
MytuC	33	0,38	$<10^{-3}$	*MypuC*	63	0,41	$<10^{-3}$
NemeM	33	0,14	0.127	*NemeM*	81	0.19	0.006
NemeZ	26	0,28	0.026	*NemeZ*	67	0.21	0.006
Pamu	23	0,32	0.018	*Pamu*	62	0.36	$<10^{-3}$
Psae	51	0,24	0.006	*Pyab*	21	0,20	0.113
StauM	30	0,35	0.003	*Pyho*	21	0,28	0.040
StauN	29	0,30	0.012	*StauN*	128	0.22	$<10^{-3}$
Sysp	29	0,19	0.079	*Stpy*	42	0,17	0.055
Xyfa	46	0,55	$<10^{-3}$	*Suso*	40	0,13	0.121
				Sysp	77	-0.05	0.264
				Trpa	22	0,35	0.011
				Vich_1	61	0,08	0.198

[a]24 chromosomes with more than 20 two-copy CDR were used to test correlations.
[b]Abbreviations are those used in Table 1.
[c]Coefficients of Kendall τ rank tests between spacer size and length for CDR.
[d]Probability associated with the Kendall τ rank tests. We assumed that, at an α risk of 0.01, 13/24 correlations are significant (with a risk of 0.21 of getting at least one false positive and of 0.02 of getting at least two false positives).
[e]We used two-copy, three-copy, four-copy and five-copy CDR to test 17 new chromosomes and re-test the 11 non-significant ones.
[f]Probability associated with the Kendall τ rank tests. We assumed that, at an α risk of 0.01, 11/27 correlations are significant (with a risk of 0.24 of getting at least one false positive and of 0.03 of getting at least two false positives).

based on empirical distributions for each genome and also defined specific scoring matrices, calculated taking into account the nucleotide compositions of the genomes (see Materials and Methods). This is why the minimal significant alignment score is larger for more biased genomes, such as some *Mycoplasma* spp. Since methodological biases were taken into account in the search for repeats, one is inclined to explain these results from a biological point of view.

Whatever the mechanism of tandem repeat genesis, it always requires pre-existing small repeats (11). Levinson and Gutman (8) have proposed that small repeats appear by chance and are at the origin of larger repeats that are created by slipped strand mispairing between these small repeats. It so happens that low complexity genomes, by chance alone, present a larger number of small repeats. If we accept the hypothesis that tandem genesis mechanisms are not down-regulated in low complexity genomes, then we are immediately led to the conclusion that tandem genesis must be more frequent in these genomes, simply due to their higher compositional bias. Thus, we propose that in such genomes a higher number of primers appear by chance and lead to more abundant repeats.

Small, non-duplicated repeats can be used as primers for initiation of tandem duplications. Thus, many types of repeats are related: small repeats are transformed into tandem repeats, which are then turned into interspersed repeats. As a consequence one gains by analysing these repeats together, instead of dividing them into different classes.

In this respect, it is interesting to note that chromosomes 2 and 3 of *Plasmodium falciparum* exhibit a very high density of repeats (as compared with eukaryote chromosomes of the same size) (16) which is associated with a very low G + C content (18%). It is therefore tempting to suggest that in eukaryote chromosomes complexity of the genome also plays an important role in the mechanisms of repeat generation. Naturally, the statistical testing of this generalisation will have to await the availability of a larger sample of complete eukaryote genomes.

CONCLUSION

We have shown that a model for the dynamics of repeats (previously established in Eukarya), based on tandem genesis with further dispersion, holds for most Bacteria and Archaea. As predicted by the model, we show that in most genomes (i) direct repeats are more numerous than inverted repeats, (ii) CDR are in large excess, (iii) there is a negative correlation between repeat identity and spacer size and (iv) there is a positive correlation between repeat length and spacer size. This strongly suggests that despite their diversity, intrachromosomal repeats of

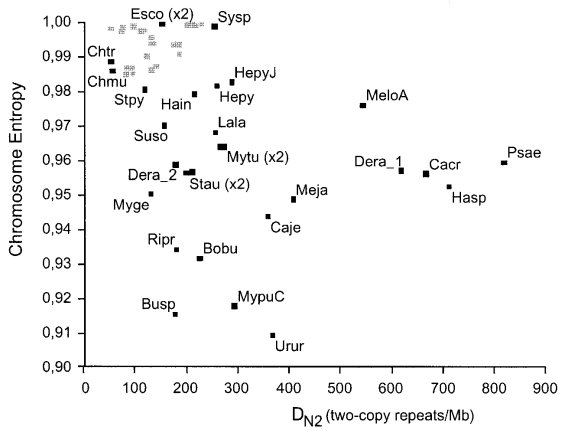


Figure 4. Complexity of chromosomes as a function of repeat density. Entropy (a measure of nucleotide complexity) of each of the 53 chromosomes as a function of their global repeat density. Entropy measures the nucleotide complexity of a sequence: if each nucleotide frequency is 0.25, then entropy is maximum (1), else it is lower. This figure illustrates that entropy is negatively correlated with repeat density.

all genomes share similar dynamics that are probably related to very ancient mechanisms shared by the three domains of life. Naturally, this model is not exclusive of other mechanisms of duplication (transposition, horizontal gene transfer, insertions, hyperploidisation, etc.).

We have also shown that nucleotide composition biases of the chromosome strongly influence the rate of tandem repeat creation and thus the rate of repeat amplification. Other effects are likely to shape the dynamics of bacterial repeats and the large availability of complete genomes will shed light on them. This will certainly provide new clues in deciphering the dynamics of repeats in bacterial genomes and shed additional light on genome evolution.

ACKNOWLEDGEMENTS

We would like to thank I. Gonçalves, D. Higuet, E. Maillier and J. Pothier for their scientific help and their friendly support. We would also like to thank P. Avner and E. Leguern for their helpful remarks on previous versions of this manuscript. This work was supported by grants from the Association pour la Recherche sur le Cancer. G.A. was funded by the Fondation pour la Recherche Médicale. E.C. and P.N. are members of Université Pierre et Marie Curie (Paris, France).

REFERENCES

1. Ohno,S. (1970) *Evolution by Gene Duplication*. Springer-Verlag, Heidelberg, Germany.
2. Wolfe,K.H. (2001) Yesterday's polyploids and the mystery of diploidization. *Nature Rev. Genet.*, **2**, 333–341.
3. Andersson,S.G. and Kurland,C.G. (1998) Reductive evolution of resident genomes. *Trends Microbiol.*, **6**, 263–268.
4. Katti,M.V., Ranjekar,P.K. and Gupta,V.S. (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol. Biol. Evol.*, **18**, 1161–1167.
5. Le Fleche,P., Hauck,Y., Onteniente,L., Prieur,A., Denoeud,F., Ramisse,V., Sylvestre,P., Benson,G., Ramisse,F. and Vergnaud,G. (2001) A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*. *BMC Microbiol.*, **1**, 2.
6. Levinson,G. and Gutman,G.A. (1987) High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. *Nucleic Acids Res.*, **15**, 5323–5338.
7. van Belkum,A., van Leeuwen,W., Scherer,S. and Verbrugh,H. (1999) Occurrence and structure-function relationship of pentameric short sequence repeats in microbial genomes. *Res. Microbiol.*, **150**, 617–626.
8. Levinson,G. and Gutman,G.A. (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.*, **4**, 203–221.
9. Yeramian,E. and Buc,H. (1999) Tandem repeats in complete bacterial genome sequences: sequence and structural analyses for comparative studies. *Res. Microbiol.*, **150**, 745–754.
10. Mahillon,J. and Chandler,M. (1998) Insertion sequences. *Microbiol. Mol. Biol. Rev.*, **62**, 725–774.
11. Romero,D. and Palacios,R. (1997) Gene amplification and genomic plasticity in prokaryotes. *Annu. Rev. Genet.*, **31**, 91–111.
12. Rocha,E.P.C., Danchin,A. and Viari,A. (1999) Analysis of long repeats in bacterial genomes reveals alternative evolutionary mechanisms in *Bacillus subtilis* and other competent prokaryotes. *Mol. Biol. Evol.*, **16**, 1219–1230.
13. Kraus,E., Leung,W.Y. and Haber,J.E. (2001) Break-induced replication: a review and an example in budding yeast. *Proc. Natl Acad. Sci. USA*, **98**, 8255–8262.
14. Paques,F. and Haber,J.E. (1999) Multiple pathways of recombination induced by double-strand breaks in *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.*, **63**, 349–404.
15. Achaz,G., Coissac,E., Viari,A. and Netter,P. (2000) Analysis of intrachromosomal duplications in yeast *Saccharomyces cerevisiae*: a possible model for their origin. *Mol. Biol. Evol.*, **17**, 1268–1275.
16. Achaz,G., Netter,P. and Coissac,E. (2001) Study of intrachromosomal duplications among the eukaryote genomes. *Mol. Biol. Evol.*, **18**, 2280–2288.
17. Chedin,F., Dervyn,E., Dervyn,R., Ehrlich,S.D. and Noirot,P. (1994) Frequency of deletion formation decreases exponentially with distance between short direct repeats. *Mol. Microbiol.*, **12**, 561–569.
18. Lovett,S.T., Gluckman,T.J., Simon,P.J., Sutera,V.J. and Drapkin,P.T. (1994) Recombination between repeats in *Escherichia coli* by a recA-independent, proximity-sensitive mechanism. *Mol. Gen. Genet.*, **245**, 294–300.
19. Bi,X. and Liu,L.F. (1996) recA-independent DNA recombination between repetitive sequences: mechanisms and implications. *Prog. Nucleic Acid Res. Mol. Biol.*, **54**, 253–292.
20. Peeters,B.P., de Boer,J.H., Bron,S. and Venema,G. (1988) Structural plasmid instability in *Bacillus subtilis*: effect of direct and inverted repeats. *Mol. Gen. Genet.*, **212**, 450–458.
21. Kurtz,S. and Schleiermacher,C. (1999) REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics*, **15**, 426–427.
22. Karlin,S. and Ost,F. (1985) Maximal segmental match length among random sequences from a finite alphabet. In Cam,L.M.L. and Olshen,R.A. (eds), *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*. Association for Computing Machinery, New York, NY, Vol. 1, pp. 225–243.
23. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
24. Sueoka,N. (1962) On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl Acad. Sci. USA*, **48**, 582–592.
25. Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, **188**, 415–431.
26. Hayashi,T., Makino,K., Ohnishi,M., Kurokawa,K., Ishii,K., Yokoyama,K., Han,C.G., Ohtsubo,E., Nakayama,K., Murata,T. *et al.* (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.*, **8**, 11–22.
27. Rocha,E.P.C. and Blanchard,A. (2002) Genomic repeats, genome plasticity and the dynamics of *Mycoplasma* evolution. *Nucleic Acids Res.*, **30**, 2031–2042.
28. Coissac,E., Maillier,E. and Netter,P. (1997) A comparative study of duplications in bacteria and eukaryotes: the importance of telomeres. *Mol. Biol. Evol.*, **14**, 1062–1074.
29. Shigenobu,S., Watanabe,H., Hattori,M., Sakaki,Y. and Ishikawa,H. (2000) Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature*, **407**, 81–86.
30. Ochman,H. and Moran,N.A. (2001) Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science*, **292**, 1096–1099.