Diplôme National d'Habilitation à Diriger des Recherches

Evolution des Génomes et du Métabolisme : Annotation, Analyse et Exploration de données

présenté par

Olivier Lespinet

le vendredi 10 décembre 2010

devant le jury constitué de :

Simonetta Gribaldo	Chargée de Recherche, Institut Pasteur	Rapporteur
Claudine Médigue	Directeur de Recherche, CNRS	Rapporteur
Dominique Higuet	Professeur, Université Pierre et Marie Curie	Rapporteur
Christine Froidevaux	Professeur, Université Paris-Sud 11	Examinateur
Herman van Tilbeurgh	Professeur, Université Paris-Sud 11	Examinateur

Remerciements

Je souhaite profiter de ces quelques lignes pour remercier tous ceux, famille, amis et collègues, qui m'ont accompagné et encouragé tout au long de ces années. Merci aux nombreux collègues et étudiants sans lesquels le travail de ces dernières années n'aurait pas été possible, en particulier Bernard, Quentin, Frédéric, Stéphane, Matthieu et Sandrine. Je ne voudrais pas non plus oublier André et Michel qui m'ont guidé lors de mes premières années de thèse.

Merci enfin, aux membres du jury d'avoir accepté d'évaluer ce travail.

A Laure, Camille et Manon.

Table des matières

- Chapitre 1 -

Curriculum vitae

Olivier Lespinet

Maître de Conférences des Universités

Spécialités : Bioinformatique, Génomique et Evolution

Institut de Génétique et Microbiologie
UMR 8621 CNRS/Université Paris-Sud 11
Bâtiment 400, UFR des Sciences d'Orsay
91405 Orsay Cedex, FRANCE

Téléphone +33 (0)1 69 15 35 59
Fax +33 (0)1 69 15 72 96
E-mail olivier.lespinet@igmors.u-psud.fr
Page web http://olivier-lespinet.info

ETAT CIVIL

40 ans, né à Niort (Deux-Sèvres) le 28 octobre 1970
Nationalité française
Marié (2 enfants)

CURSUS, FORMATIONS ET TITRES UNIVERSITAIRES

15 janvier 2001 — Doctorat en Génétique Cellulaire et Moléculaire
mention *Très Honorable*, Université de Paris-Sud 11, Faculté des Sciences d'Orsay
Titre : **La famille des genes *Snail*. Caractérisation de deux nouveaux membres chez le mollusque *Patella vulgata*. Hypothèse sur leur fonction ancestrale chez les *Bilateria*.**
Composition du jury de thèse : Pr Bernadette Limbourg-Bouchon (Présidente), Pr Jo van den Biggelaar (Rapporteur), Pr Michel Volovitch (Rapporteur), Pr Andre Adoutte (Examinateur), Dr Angela Nieto (Examinatrice) et Dr Michel Cassan (Directeur de thèse)

1999 — European Advanced Practical Training Course in Evolutionnary Developmental Biology
Station Marine de Roscoff, France

1997 — "*Molecular and Cellular Approaches to Genetic Analysis of development"*
Erasmus Intensive Course, ENS, Paris, France

1996 — DEA de Génétique Cellulaire et Moléculaire
Mention *Très Bien*, Université de Paris-Sud 11, Faculté des Sciences d'Orsay

1995-1996 Service National

1994 — Maîtrise de Génétique Moléculaire et d'Informatique Appliquée
Mention *Bien*, Université de Paris-Sud XI (Orsay)

1993 — Licence de Biochimie
Mention *Assez Bien*, Université de Paris-Sud XI (Orsay)

1992 — DEUG B, Sciences de la Nature et de la Vie, option Physiologie et Biologie Cellulaire
Mention *Assez Bien*, Université de Poitiers

1989-1990 — Classe préparatoire aux grandes écoles (Math-sup Bio), Lycée Camille Guérin (Poitiers)

1989 — Baccalauréat série C, Lycée Jean Macé (Niort)

PARCOURS PROFESSIONNEL

Depuis 2002	Maître de Conférences des Universités Equipe Evolution Moléculaire et Bioinformatique des Génomes (Directeur : B. Labedan) Institut de Génétique et Microbiologie, UMR CNRS 8621/Université Paris-Sud 11, Orsay, France
2001-2002	NIH Visiting Fellow Evolutionary Genomic Research Group (Principal Investigator : E. Koonin) National Center for Biotechnology Information, NIH, Bethesda MD, USA
1999-2001	ATER (demi-poste) en Bioinformatique, Université Paris-Sud 11 Equipe Evolution et Développement des Métazoaires (Directeur : A. Adoutte) Centre de Génétique Moléculaire, UPR CNRS 2167, Gif-sur-Yvette,France
1996-2001	Doctorant en Génétique Cellulaire et Moléculaire (Allocataire MENRT) Equipe Evolution et Développement des Métazoaires (Directeur : A. Adoutte) Centre de Génétique Moléculaire, UPR CNRS 2167, Gif-sur-Yvette,France

ACTIVITES D'ENSEIGNEMENT

Depuis 2002 — Maître de conférences des Universités, Faculté des Sciences d'Orsay, Université Paris-Sud 11

Je suis actuellement responsable (ou co-responsable) des unités d'enseignement suivantes :

- Génomique Comparée (M2 Biologie-Santé et ED Gènes, Génomes,Cellules)
- Evolution en questions ? (M1 Biologie-Santé)
- Génomes et Systèmes (M1 Bioinformatique et Biostatistiques)
- Analyse *in silico* des génomes (L3 Biologie, Magistère de Biotechnologie)

Je participe également aux unités d'enseignement :

- BioInformatique Appliquée (M2 MAPS, Faculté de Pharmacie de Châtenay-Malabry)
- Génomique Fondamentale (M1 Biologie-Santé)
- Analyse des séquences (M1 Bioinformatique et Biostatistiques)
- Mathématiques et Biologie (L3 Mathématiques)
- Informatique Appliquée à la Biologie (L2 Biologie)

2001	Teaching Assistant (40 heures), Practical on Phylogenetic reconstruction Responsable: Pr André Adoutte Embryology Summer Course, Woods Hole Marine Laboratory, Woods Hole MA, USA
1999-2001	ATER en Bioinformatique, poste à mi-temps, 96 h ETD Responsable: Dr Michel Termier Enseignement intégré de Bioinformatique (Licence de Biologie)
1998-1999	Vacataire en Informatique, Université de Versailles Saint-Quentin-en-Yvelines, 32 h ETD Responsable: Dr Claudine Lalande-Devauchelle Enseignement de découverte de la programmation Visual Basic en DEUG Sciences de la Vie

ACTIVITES ADMINISTRATIVES

2010	*Membre des comités de sélection des postes :* • chaire INSERM/Université Paris-Sud 11, 67/65 MCF 2165 profil : *Innovations méthodologiques et transfert pour l'analyse de données de génomique clinique* • 27 MCF 1471 de l'université Paris-sud 11 profil : *Informatique*
Depuis 2010	• Nommé au conseil Scientifique de l'Institut de Génétique et Microbiologie • Nommé au comité de pilotage de la plateforme e-BIO (RENABI Ile de France-Sud)
2009	Membre du comité de sélection du poste 65 MCF 1600 de l'université Paris-Sud 11 profil : *Trafic membranaire de phagocytes.*
2008	Membre du comité de sélection du poste 65/27 MCF 0396 de l'université d'Evry-Val-d'Essonne profil : *Biologie Systémique et Architecture des Génomes*
Depuis 2007	• Président de la Commission de la Pédagogie de la Faculté des Sciences d'Orsay • Membre Invité du conseil des formations de la faculté des Sciences d'Orsay • Elu au conseil de la faculté des Sciences d'Orsay
2006-2008	• Elu au bureau de la commission de spécialistes CNU 64-65 de l'Université Paris-Sud 11 • Participe au bureau du département de Biologie de la Faculté des sciences d'Orsay
2005-2010	• Président de jury des semestres S5 et S6 de la licence de Biologie de l'Université Paris-Sud 11 • responsable du parcours Bioinformatique et Biostatistiques du L3 de Biologie de Paris-Sud 11
2004-2008	Elu au département de Biologie de la faculté des Sciences d'Orsay
2003-2010	Elu à la commission de spécialistes puis à la CCSU CNU 64-65 de l'université Paris-Sud 11
1999-2001	Elu étudiant au Conseil de Laboratoire du Centre de Génétique Moléculaire, CNRS Gif-sur-Yvette
1996-1998	Elu étudiant au Conseil Scientifique de l'Université de Paris-Sud 11

ACTIVITES LIEES A LA RECHERCHE

J'ai expertisé des articles pour les revues suivantes : Bioinformatics, BMC Bioinformatics, BMC Genomics, Gene, Briefings in Bioinformatics, Genetica, Molecular Biology and Evolution, PLoS Computational Biology, Yeast.

J'ai expertisé des projets dans le cadre des programmes suivants : ACI IMPBio, ANR Systèmes Complexes et Modélisation Mathématique, Génome Québec.

Depuis janvier 2010 j'anime la thématique Diversité, Dynamique et Evolution des Systèmes Biologiques au

sein de l'Institut de Génétique et Microbiologie. Le statut de porteur de thématique me permet de développer ma propre activité de recherche tout en étant hébergé au sein de l'équipe Evolution Moléculaire et Bioinformatique des Génomes dirigée par Bernard Labedan.

Depuis le 1er octobre 2008, je suis bénéficiaire d'une Prime d'Enseignement Doctorale et de Recherche (PEDR).

ENCADREMENT D'ETUDIANTS ET DE STAGIAIRES

- Doctorants

Depuis 2008	Matthieu Barba Université Paris-Sud 11, Ecole Doctorale Gènes, Génomes et Cellules (bourse MESR) Sujet : Evolution de réseaux métaboliques complexes chez les microorganismes. Encadrement à hauteur de 10%, 1 acte de congrès publié
Depuis 2007	Sandrine Grossetête Lalami Université Paris-Sud 11, Ecole Doctorale Gènes, Génomes et Cellules (bourse CNRS) Sujet : Comparaison du métabolisme chez les champignons Encadrement à hauteur de 100%, 1 acte de congrès et 2 articles publiés
2005-2008	Frédéric Lemoine Université Paris-Sud 11, Ecole Doctorale Informatique (bourse MESR) Titre : Intégration, Interrogation et Analyse de données de génomique comparative Thèse soutenue le 08 décembre 2008 Encadrement à hauteur de 10%, 3 articles publiés
2002-2007	Quentin Sculo Université Paris-Sud 11, Ecole Doctorale Gènes, Génomes et Cellules (bourse MESR) Titre : Arbres génomiques du vivant : nouvelles approches expérimentales Thèse soutenue le 09 mars 2007 Encadrement à hauteur de 20% / 1 acte de congrès et 1 article publiés

- M2, DEA, DESS

2008	Matthieu Barba Université Paris-Sud 11, M2 (R) Sciences-Technologie-Santé, mention BIBS Sujet : Etude des gènes de la voie de biosynthèse de l'arginine et des voies liées : mise au point de méthodes d'annotations. Encadrement à hauteur de 30%
2007	Sandrine Grossetête Lalami Université Paris 7, M2 (R) Sciences et Applications, mention Biologie-Informatique Sujet : Génomique comparée des champignons Encadrement à hauteur de 90% Céline Antoine Université Paris-Sud 11, M2 (P) Sciences-Technologie-Santé, mention BIBS Sujet : Treefactory : un service web pour la phylogénie Encadrement à hauteur de 100%

2006 | Jean-Patrick Bordrez
Université Paris-Sud 11, M2 (P) Sciences-Technologie-Santé, mention BIBS
Sujet : Conception d'un outil intégré pour la phylogénie
Encadrement à hauteur de 100%

2005 | Frédérique Lemoine
Université Paris-Sud 11, M2 (R) Sciences-Technologie-Santé, mention BIBS
Sujet : Analyse du contexte génétique chez les procaryotes
Encadrement à hauteur de 30%

Vivianne Ndonko
Université Paris-Sud 11, M2 (R) Sciences-Technologie-Santé, mention ISFMB
Sujet : Comparaison exhaustive de données génomiques microbiennes
Encadrement à hauteur de 100%

Hamid Khalili
Université Evry-Val-d'Essonne, 3ième année IUP Génie Biologique et Informatique
Sujet : Conception d'outils pour l'assemblage du génome du champignon *Podospora anserina*
Encadrement à hauteur de 50%

2004 | Laurent Fant
Université Rennes 1, DEA Génomique et Informatique
Sujet : Estimation de la complexité modulaire des protéines
Encadrement à hauteur de 10%

2003 | Adrien Bazureau
Université Bordeaux 2, DESS Informatique des Systèmes Complexes
Sujet : Structuration de quantités massives de données biologiques
Encadrement à hauteur de 10%

- M1, L3, L2

Depuis 2002, j'ai encadré ou co-encadré 16 stages d'étudiants de niveau L2 à M1.

PUBLICATIONS ET COMMUNICATIONS

Articles publiés dans des revues internationales avec comité de lecture :

1. **A general framework for optimization of probes for gene expression microarray and its application to the fungus *Podospora anserina*.** (2010) *F Bidard, S Imbeaud, N Reymond, O Lespinet, P Silar, C Clavé, H Delacroix, V Berteaux-Lecellier and R Debuchy - BMC Research Notes ; Jun 18;3:171.*

2. **FUNGIpath: a tool to assess fungal metabolic pathways predicted by orthology[1].** (2010) *S Grossetête, B Labedan and O Lespinet* - BMC Genomics; Feb 1;11(1):81.

3. **Assessment of phylogenetic diversity of bacterial microflora in drinking water using serial analysis of ribosomal sequence tags.** (2009) *JB Poitelon, M Joyeux, B Welté, JP Duguet, E Prestel,*

1 Articles n°1 et 6 : Ces deux articles ont obtenu l'étiquette ”Highly Accessed” décernée par les journaux du groupe BMC

O Lespinet and MS DuBow - Water Research; Sep;43(17):4197-206.

4. **SynteBase/SynteView: a tool to visualize gene order conservation in prokaryotic genomes.** (2008) *F Lemoine, B Labedan and O Lespinet* - BMC Bioinformatics; Dec 16;9:536.

5. **Genome-wide analysis of the *Fusarium oxysporum* mimp family of MITEs and mobilization of both native and de novo created mimps.** (2008) *M Bergemann, O Lespinet, SB M'Barek, MJ Daboussi and M Dufresne* - Journal of Molecular Evolution; Dec;67(6):631-42.

6. **The Genome Sequence of the Model Ascomycete Fungus *Podospora anserina*[1].** (2008) *E Espagne[2], O Lespinet[2], F Malagnac[2], C Da Silva, O Jaillon, BM Porcel, A Couloux, JM Aury, B Ségurens, J Poulain, V Anthouard, S Grossetête, H Khalili, E Coppin, M Déquard-Chablat, M Picard, V Contamine, S Arnaise, A Bourdais, V Berteaux-Lecellier, D Gautheret, RP de Vries, E Battaglia, PM Coutinho, EGJ Danchin, B Henrissat, R El Khoury, A Sainsard-Chanet, A Boivin, B Pinan-Lucarré, CH Sellem, R Debuchy, P Wincker, J Weissenbach and P Silar* - Genome Biology; 9(5):R77.

7. **The multiple facets of homology and their use in comparative genomics to study the evolution of genes, genomes, and species.** (2008) *S Descorps-Declère, F Lemoine, Q Sculo, O Lespinet and B Labedan* - Biochimie; Apr;90(4):595-608.

8. **Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data.** (2007) *F Lemoine, O Lespinet and B Labedan* - BMC Evolutionary Biology; 7(1):237.

9. **ORENZA: a web resource for studying ORphan ENZyme Activities.** (2006) *O. Lespinet and B Labedan* - BMC Bioinformatics; 7:436.

10. ***Orphan enzymes could be an unexplored reservoir of new drug targets.*** (2006) *O Lespinet and B Labedan* - Drug Discovery Today; 11(7-8):300-5.

11. **Puzzling over orphan enzymes.** (2006) *O Lespinet and B Labedan* - Cellular and Molecular Life Science; 63(5):517-23.

12. **Orphan enzymes?[3]** (2005) *O Lespinet and B Labedan* - Science; 307(5706):42.

13. **Retrieving the Whole Set of Protein Modules of *Campylobacter jejuni* and *Helicobacter pylori*.** (2003) *Q Sculo, O Lespinet and B Labedan* - Genome Letters; 2, 8–15.

14. **The role of lineage-specific gene family expansion in the evolution of eukaryotes[3].** (2002) *O Lespinet, YL Wolf, EV Koonin and L Aravind* - Genome Research; 12(7):1048-59.

15. **A lophotrochozoan *twist* gene is expressed in the ectomesoderm of the gastropod mollusc *Patella vulgata*.** (2002) *AJ Nederbragt[4], O Lespinet[4], S van Wageningen, AE van Loon, A Adoutte, and WJAG Dictus* - Evolution & Development; 4(5):334-43.

16. **Characterisation of two Snail genes in the gastropod mollusc *Patella vulgata*. Implications for understanding the ancestral function of the Snail-related genes in *Bilateria*.** (2002) *O Lespinet[4], AJ Nederbragt[4], M Cassan, WJAG Dictus, AE van Loon, and A Adoutte* - Development, Gene and Evolution; 212(4):186-95.

2 Article n°6 : Les trois premiers auteurs ont contribué à part égale à ce travail.
3 Articles n°12 et 14: La lecture de ces deux articles est "Recommendée" par Faculty of 1000 Biology
4 Articles n°15 et 16: Les deux premiers auteurs ont contribué à part égale à ce travail.

17. **Expression pattern of Brachyury in the mollusc *Patella vulgata* suggests a conserved role in the establishment of the A-P axis in *Bilateria***. (2002) *N Lartillot, O Lespinet, M Vervoort, JAM van den Biggelaar, and A Adoutte* - Development; 129(6):1411-21.

18. **Genome quality control: RIP (Repeat-Induced Point mutation) comes to *Podospora*.** (2001) *F Graïa, O Lespinet, B Rimbault, M Dequard-Chablat, E Coppin, and M Picard* - Molecular Microbiology; 40(3):586-95.

19. **The new animal phylogeny: Reliability and Implications.** (2000) *A Adoutte, G Balavoine, N Lartillot, O Lespinet, B Prud'homme, and R de Rosa* - PNAS USA; 97(9): 4453-4456.

Chapitres d'ouvrages:

1. **Interspecies and intraspecies comparison of microbial proteins: learning about gene ancestry, protein function, and species life style.** (2006) *B Labedan and O Lespinet* - WHILEY, Microbial Proteomics: Functionnal Biology of Whole Organisms. Editeurs Ian Humphery-Smith, Michael Hecker.

Articles publiés dans des actes de congrès nationaux avec comité de lecture :

1. **Fast and accurate multiple sequence alignment of large and diversified sets of distant homologues.** (2010) *M Barba, O Lespinet and B Labedan* - JOBIM, Montpellier 7-9 Septembre, Editeurs O Gascuel, MF Sagot.

2. **FUNGIpath: a new tool for analysing the evolution of fungal metabolic pathways.** (2009) *S Grossetête, B Labedan and O Lespinet* - JOBIM, Nantes 9-11 Juin, Editeurs E Rivals, I Rusu.

3. **New approaches to improve the soundness of the deep evolutionary relationships in genomic trees of microorganisms.** (2005) *Q Sculo, O Lespinet and B Labedan* - JOBIM, Lyon 6-8 Juillet, Editeurs G Perrière, A Guénoche, C Geourjon.

Articles de vulgarisation :

1. **Darwin et l'arbre du vivant.** (2009) *O Lespinet, B Labedan, N Glansdorff et Y Xu* - Plein Sud, Le magazine d'information de l'université Paris-Sud; 73:14-15

Communications orales:

1. **FUNGIpath: a new tool for analysing the evolution of fungal metabolic pathways.** (2009) JOBIM, 9-11 Juin 2009, Nantes (France).

2. **When data integration leads to a new concept:the orphan enzymes.** (2008) EMBnet Conference, Leading Applications and Technologies in Bioinformatics. 18-20 Septembre 2008, Martina Franca (Italie)

3. **Archéologie des protéines.** (2007) Exobio'07. Des soleils à la vie : où, quand, comment ? Ecole thématique du CNRS. 22-29 septembre 2007 propriano (France)

4. **Vers une base de données des Activités Enzymes Orphelines : ORENZA.** (2006) Séminaire de l'Institut de Biochimie et Biophysique Moléculaire et Cellulaire, 20 juin 2006, Orsay, France.

5. **Des molécules aux génomes : une vision évolutive.** (2006) Journées scientifiques de l'Institut de Génétique et Microbiologie, 2 juin 2005, Orsay, France.

6. **Le paradoxe des enzymes orphelines.** (2006) Colloque du Programme de Pluriformation "Bioinformatique et Génomique" de l'Université Paris XI. Tours, 9-10 mars 2006.

Communications affichées:

1. **A whole genome oligo microarray approach to decipher *Podospora anserina* sexual development** (2008) F Bidard, S Imbeaud, N Reymond, O Lespinet, P Silar, C Clave, H Delacroix, V Berteaux-Lecellier and R Debuchy. 9th European Congress on Fungal Genetics, 5-8 avril 2008, Edinbourgh, Grande-Bretagne.

2. **Selection of Optimal Oligonucleotide Probes for a Whole Genome Microarray Approach to Decipher *Podospora anserina* Sexual Development.** (2007) F Bidard, S Imbeaud, S Arnaise, A Bourdais, K Budin, E Coppin, E Espagne, O Lespinet, F Malagnac, M Paoletti, L Peraza- Reyes, N Reymond, S Saupe, E Sicault-Sabourin, D Zickler, P Silar, C Clave, H Delacroix, V Berteaux-Lecellier and R Debuchy. 24th Fungal Genetics Conference, 20-25 mars 2007, Asilomar, Californie, USA.

3. **From genomics to systems biology: playing with the evolution of synteny blocks in microbial genomes.** (2007) *F Lemoine, O Lespinet and B Labedan - JOBIM, 10-12 Juillet 2007,* Marseille, France.

4. **Analyzing gene context to better understand the evolutionary mechanisms underlying gene order conservation in prokaryotes.** (2006) *F Lemoine, O Lespinet and B Labedan - JOBIM, 5-7 Juillet 2006, Bordeaux, France.*

5. **The *Podospora anserina* Genome Project.** (2005) P Silar, P Wincker, F Malagnac, E Espagne, A Boivin, O Lespinet, A Khalili, R Debuchy, S Arnaise, V Berteaux-Lecellier, C Clave, V Contamine, E Coppin, A Couloux, C Dasilva, F Debets, M Dequard-Chablat, R Hoekstra, M Picard, B Pinan-Lucarre, A Sainsart-Chanet, S Saupe, CH Sellem and J Weissenbach. ESF-EMBO Symposium on Comparative Genomics of Eukaryotic Microorganisms - November 12-17, 2005, Sant Feliu de Guixols, Spain

6. **The *Podospora* Genome Project.** (2005) P Silar, P Wincker, F Malagnac, E Espagne, A Boivin, O Lespinet, A Khalili, R Debuchy, S Arnaise, V Berteaux-Lecellier, C Clave, V Contamine, E Coppin, A Couloux, C Dasilva, F Debets, M Dequard-Chablat, R Hoekstra, M Picard, B Pinan-Lucarre, A Sainsart-Chanet, S Saupe, CH Sellem and J Weissenbach. 23th Fungal Genetics Conference, 15-20 mars 2005, Asilomar, Californie, USA.

7. **The role of Lineage-Specific Gene Family Expansion in the Evolution of Eukaryotes** (2002) O Lespinet, YL Wolf, EV koonin and L Aravind – Gordon Research Conference “Structural , Functionnal and Evolutionnary Genomics”, 29 juillet-02 août, Mount Holyoke College, (MA, USA)

8. ***Twist* and *snail* homologues in the gastropod mollusk *Patella vulgata*: a new twist for an old gene network?** A.J. Nederbragt, O. Lespinet, W.J.A.G. Dictus, A.E. van Loon, A. Adoutte, and J.A.M. van den Biggelaar. Society for Integrative and Comparative Biology, Annual Meeting - January 3-7, 2001, Chicago, USA

9. **Evolutionary aspect of mesoderm specification.** (1998) *Olivier Lespinet, Nicolas Lartillot, Michel Cassan & André Adoutte.* Conférences Jacques MONOD, "Developmental Biology: Early steps in embryogenesis" June 1-5, 1998, Aussois (France)

10. **A comparative approach of mesoderm determination.** *(1998) Nicolas Lartillot & Olivier Lespinet*

- *NASA - CASSLS Workshop on Developmental Gene Regulation and Mechanisms of Evolution* June 10-13, 1998, Woods Hole, USA

11. **Genetic conservation of mesoderm specification in *metazoa.*** (1997) *Olivier Lespinet, Nicolas Lartillot, Michel Cassan & André Adoutte Annual Meeting of the French Developmental Society* May 29-31, 1997, Dourdan (France)

- Chapitre 2 -

Synthèse des travaux

2.0. Homologues, Orthologues et Paralogues.

Les termes d'homologues, d'orthologues et de paralogues seront utilisés à de nombreuses reprises au cours de ce manuscript, il m'a donc semblé utile d'en préciser ici le sens. Deux caractères (par exemple des gènes) sont considérés comme homologues s'ils partagent un ancêtre commun. Des caractères orthologues sont des caractères homologues qui ont divergé uniquement par la suite d'évènements de spéciation. Enfin, des caractères paralogues sont des caractères homologues résultant d'un évènement de duplication qui précède (out-paralogues) ou qui suit (in-paralogues) un événement de spéciation (Descorps-Declère *et al.*, 2008).

2.1. Les prémices *(1994-1996 / Orsay, Gif-sur-Yvette)*

J'ai débuté mon initiation à la recherche au printemps 1994 au sein de l'Institut de Biologie Animale Intégrative et Cellulaire (IBAIC) d'Orsay et plus précisément dans le Laboratoire de Biologie Cellulaire 4 (BC4) qui était dirigé par le professeur André Adoutte. Il s'agissait d'un stage d'informatique appliquée effectué dans le cadre de ma maîtrise de Génétique Cellulaire et Moléculaire. Durant ces trois mois, sous la direction d'Hervé Philippe, j'ai tenté de développer une nouvelle approche stochastique pour la recherche de l'arbre le plus parcimonieux en phylogénie moléculaire.

Au delà du côté anecdotique de cette première expérience c'est dans le cadre de ce travail que j'ai d'une part découvert ce qu'était une activité de recherche et d'autre part que j'ai acquis mon intérêt pour l'évolution des séquences et des génomes. J'hésitais cependant à l'époque entre orienter la suite de ma formation vers une recherche expérimentale en biologie classique ou bien me consacrer à ce que j'avais entrevu durant ce premier stage, à savoir une recherche biologique *in silico*.

Afin de disposer de suffisamment d'arguments pour choisir entre ces deux options j'ai donc décidé l'été suivant d'effectuer un stage plus expérimental. Ce second stage de 3 mois à l'Institut de Génétique et Microbiologie (IGM), sous la direction de Vincent Colot, m'a permis de me familiariser avec la méthode de séquençage de Sanger. Il m'a surtout encouragé à poursuivre dans la voie de la recherche expérimentale et m'a incité à suivre les enseignements du DEA de Génétique Cellulaire et Moléculaire.

A l'issue de 10 mois de service national, j'ai donc poursuivi mes études par un DEA dans le cadre duquel j'ai effectué deux premiers stages de génétique classique. Le premier au Laboratoire de Génétique des Virus de Gif-sur-Yvette sous la direction de Didier Contamine et le second à l'Institut de Génétique et Microbiologie d'Orsay, sous la direction conjointe de Fatima Graïa et de Marguerite Picard. Si le premier stage m'a permis de me familiariser avec les notions de co-évolution hôte/parasite et d'évolution localisée et hétérogène des protéines *via* l'étude des ratio Ka/Ks, le second m'a permis de découvrir un nouvel aspect de l'analyse des séquences, à savoir la recherche et la découverte des mutations provoquées par le mécanisme de RIP (Repeat-Induced Point mutation) chez le champignon filamenteux *Podospora anserina*. Ce dernier travail a fait l'objet d'une publication à laquelle j'ai été associé (*Graïa et al., 2001*).

J'ai effectué mon stage long de DEA, puis ma thèse, sous la direction de Michel Cassan. A cette époque, Michel venait d'entamer une collaboration avec Guillaume Balavoine et André Adoutte sur un nouveau projet alliant biologie du développement et évolution.

2.2. Etude de la conservation des gènes impliqués dans la formation du mésoderme chez le mollusque gastéropode *Patella vulgata*. *(1996-2001 / Orsay, Gif-sur-Yvette, Roscoff, Utrecht)*

L'étude de l'évolution des gènes et des réseaux de gènes gouvernant le développement des organismes multicellulaires constitue une discipline familièrement appelée "Evo-Dévo" pour évolution du développement. L'objectif de cette discipline est d'essayer d'établir une correspondance entre les modifications du contrôle génétique des mécanismes moléculaires et cellulaires impliquées lors de l'établissement du plan d'organisation de ces organismes et l'évolution des plans d'organisations entre différentes espèces ou différents phylums.

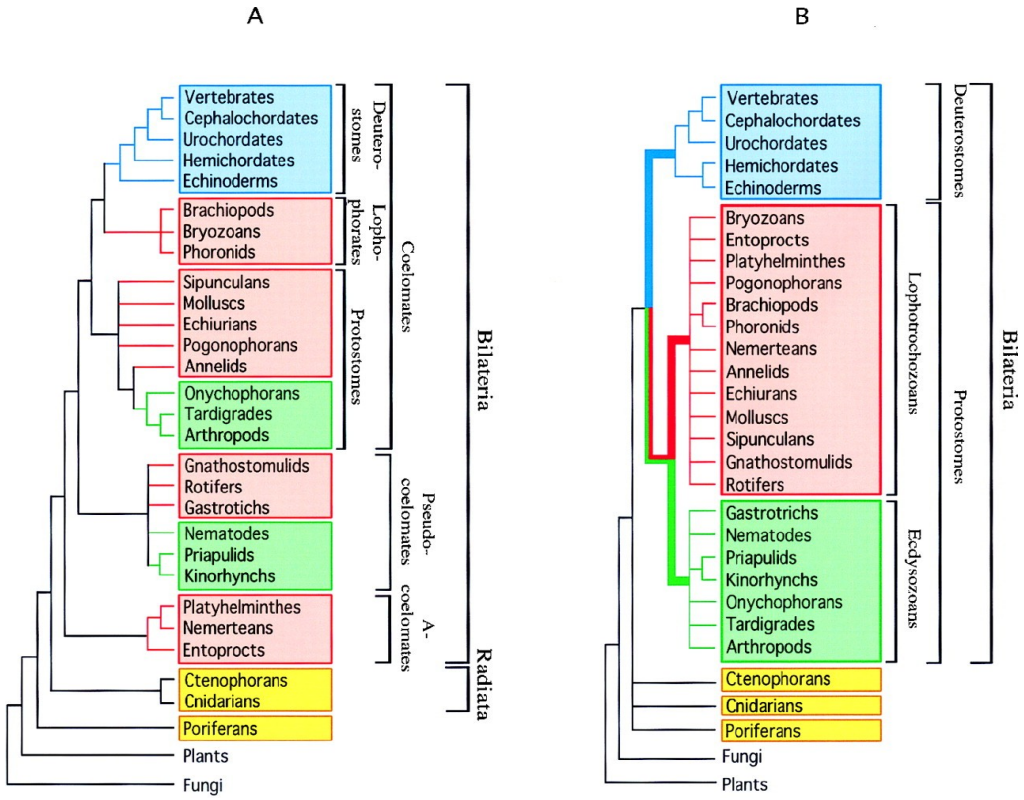


Figure 1 : *Ancienne (A) vs Nouvelle phylogénie (B) des métazoaires. Les deutérostomiens sont en bleu, les Lophotrochozoaires en rouge, les Ecdyzozoaires en vert et les métazoaires les plus basaux en jaune (d'après Adoutte et al., 2000).*

Au milieu des années 1990, les résultats obtenus par l'analyse des caractères moléculaires (Halanych *et al.*, 1995; Aguinaldo *et al.*, 1997; de Rosa *et al.*, 1999) ont permis de proposer une nouvelle phylogénie des métazoaires qui réfute la vision classique d'une évolution lente et très graduelle des animaux, allant des phylums à plan d'organisation simple vers ceux à plan d'organisation plus complexe (Figure 1 A). Cette nouvelle phylogénie réfute en particulier l'existence de phylums intermédiaires (acoelomates et pseudocoelomates) situés entre ceux à plans d'organisation simples (cnidaires et cténophores) et ceux à plans d'organisation plus complexes (coelomates) (Figure 1 A). Elle souligne au contraire le caractère élaboré du plan d'organisation de l'ancêtre des animaux à symétrie bilatérale et propose une subdivision des métazoaires en deux grands groupes, les deutérostomiens et les protostomiens (Figure 1 B). Les protostomiens étant eux même subdivisés en ecdysozoaires et lophotrochozoaires (Figure 1 B) (Adoutte e*t al.*, 2000) .

Ce nouveau scénario a conduit à proposer l'utilisation du groupe sous-étudié des lophotrochozoaires

pour essayer de valider l'hypothèse "haeckelienne" d'une origine unique du mésoderme chez l'ancêtre des métazoaires à symétrie bilatérale. En effet, en dépit d'un mode de formation très différent, un certain nombre de gènes semblent conservés lors de la spécification du mésoderme chez les ecdysozoaires et les deutérostomiens. La recherche de gènes homologues à ceux-ci chez un lophotrochozoaire devait donc nous permettre de confirmer leur fonction mésodermique ancestrale.

Pour ce faire, j'ai utilisé une approche expérimentale qui devait permettre d'apporter un élément de réponse à cette question. Cela a consisté à vérifier dans un premier temps si les gènes impliqués dans la formation du mésoderme chez les *Bilateria* étaient conservés. Il m'a fallu pour cela établir l'inventaire des gènes conservés et exprimés spécifiquement dans le territoire mésodermique des organismes modèles les plus classiques (drosophile et vertébrés). Dans un second temps j'ai essayé de déterminer si ces gènes s'exprimaient également dans le mésoderme du troisième groupe de métazoaires, les lophotrochozoaires.

2.2.1. Recherche des gènes candidats

Chez la Drosphile (ecdysozoaire), l'induction du mésoderme est contrôlée par l'établissement d'un gradient protéique maternel qui résulte directement de la mise en place de l'axe dorso-ventral. La mise en place de cet axe va aboutir à la subdivision de l'embryon en domaines se caractérisant par l'expression différentielle de certains gènes zygotiques (Figure 2).

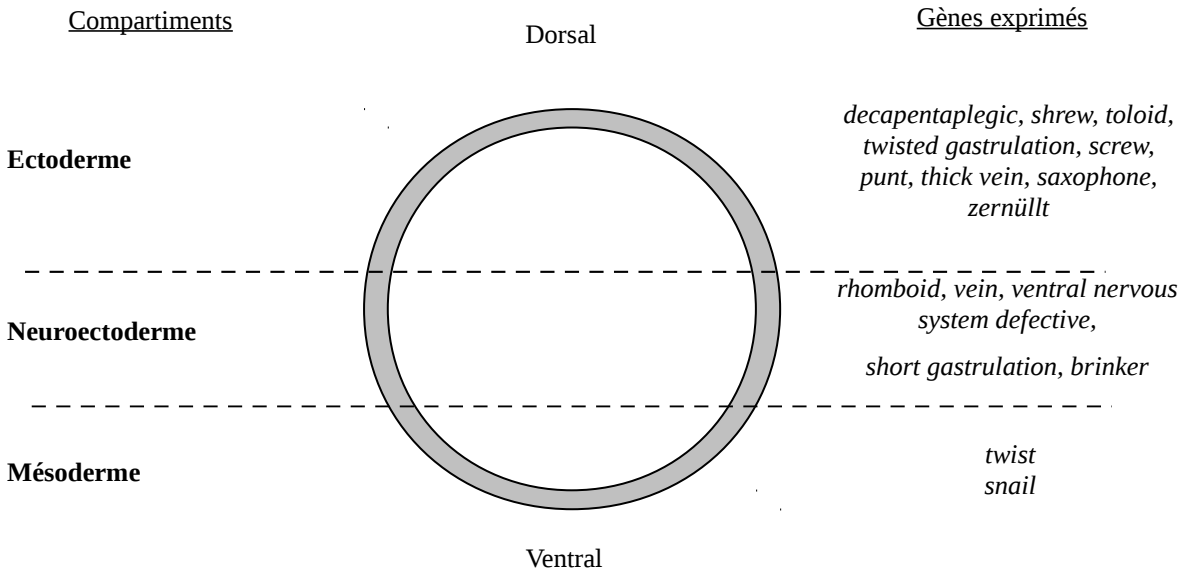


Figure 2 : *Subdivision dorso-ventrale de l'embryon de Drosophile (d'après Lespinet, 1999).*

Ainsi, dans les noyaux des cellules ventrales où la concentration du morphogène Dorsal est la plus élevée, la transcription zygotique des gènes *twist* et *snail* sera activée tandis que celle des gènes *decapentaplegic*, *zernüllt* et *tolloid* sera réprimée (Pan *et al.*, 1991; Thisse *et al.*, 1991; Ip *et al.*, 1992; Akimaru *et al.*, 1997) (Figure 2).
Dans les cellules latérales (neuro-ectoderme) où l'expression nucléaire de Dorsal est moindre, c'est la transcription des gènes *rhomboid*, *vein*, *ventral nervous system defective*, *short gastrulation* et *brinker* qui sera activée (Kosman *et al.*, 1991; Ip *et al.*, 1992; François *et al.*, 1994).
Enfin, l'absence de la protéine Dorsal, dans les noyaux les plus dorsaux, va permettre l'expression

des gènes *decapentaplegic*, *shrew*, *toloïd*, *twisted gastrulation*, *screw*, *punt*, *thick vein*, *saxophone* et *zernüllt* (Ray *et al.*, 1991; Morisato and Anderson, 1995).

Bien que leurs domaines d'expression ne soient pas strictement restreints au mésoderme on considère généralement que les gènes *twist* et *snail* constituent chez la drosophile les marqueurs les plus précoces de la détermination du mésoderme (Leptin, 1991; Leptin *et al.*, 1992). Ces deux gènes s'expriment en effet dans le mésoderme au tout début de la gastrulation. Leur expression s'atténue par la suite avant de s'exprimer de nouveau de façon dynamique dans les trois feuillets embryonnaires (Thisse *et al.*, 1988; Leptin and Grunewald, 1990; Alberga *et al.*, 1991). Mais surtout, on observe que chez les mutants pour ces gènes le mésoderme est absent ou extrêmement réduit (Simpson, 1983; Grau *et al.*, 1984; Anderson *et al.*, 1984).

En résumé, si la protéine Dorsal peut être considérée comme la molécule inductrice du mésoderme chez la Drosophile, Twist et Snail peuvent alors être considérés comme les facteurs de la détermination du mésoderme, du moins telle était la vision dominante au début de mon travail (Hammerschmidt and Nusslein-Volhard, 1993).

Ces deux gènes codent pour des facteurs de transcription qui, à ce titre, vont pouvoir activer ou réprimer l'expression de toute une série de gènes (Leptin, 1991). Ainsi, Twist est un facteur de transcription de type bHLH qui permet la transcription d'une série de gènes exprimés spécifiquement dans le mésoderme: *snail*, *twist*, *tinman*, *bagpipe*, *nautilus*, *mef2* et *zinc-finger homeodomain protein-1* (Lai *et al.*, 1991). Snail est, quant à lui, un facteur de transcription à doigt de zinc qui intervient à la fois pour réprimer dans le mésoderme l'expression de gènes neuroectodermiques, mais aussi pour permettre l'activation de gènes contrôlant les mouvements morphogénétiques qui ont lieu au cours de la gastrulation (Ip *et al.*, 1994; Nibu *et al.*, 1998; Leptin, 1999). Ainsi, associé au co-répresseur dCtBP, Snail réprimerait ventralement les gènes neuroectodermiques *rhomboïd*, *lethal of scute* et *single minded* (Kosman *et al.*, 1991; Nibu *et al.*, 1998), alors qu'il activerait les gènes *serpent*, *zinc-finger homeodomain protein-1*, *hearthless* et *folded gastrulation* (Lai *et al.*, 1991; Ip *et al.*, 1992; Kasai *et al.*,1992; Costa *et al.*, 1994; Casal et Leptin, 1996).

Si l'on connaît certains des gènes qui chez les vertébrés (deutérostomiens) s'expriment dans le mésoderme, une grande partie de l'organisation des réseaux génétiques gouvernant la détermination de ce feuillet secondaire restait encore à déterminer lorsque j'ai commencé ma thèse. Une caractéristique importante du développement des vertébrés et plus généralement des chordés est l'existence de régions dites "organisatrices": organisateur de Spemann chez le Xénope, écusson chez le Danio, nœud de Hensen chez le Poulet et nœud chez la Souris. Déterminés parfois dès la ségrégation cytoplasmique, ces organisateurs sont capables à eux seuls de spécifier l'axe dorso-ventral et jouent de ce fait un rôle prépondérant dans la différenciation dorso-ventrale des feuillets embryonnaires et en particulier du mésoderme.

Bien que l'on ignore précisément quelles sont les voies qui activent *in vivo* leurs expressions, les gènes *Xsnail* et *Xtwist*, chez le Xénope, sont également connus pour être exprimés de façon générique dans le mésoderme ou dans des régions spécifiques du mésoderme (chorde neurale) (Essex *et al.*, 1993; Hopwood *et al.*, 1989; Sargent and Bennett, 1990; Smith *et al.*, 1991).

Des homologues des gènes *snail* et *twist* ont également été isolés chez le Danio, le Poulet et la Souris (Thisse *et al.*, 1993 ; Thisse *et al.*, 1995 ; Sefton *et al.*, 1998). D'une espèce à l'autre, il peut cependant exister des différences en ce qui concerne les domaines d'expression, les fonctions ainsi que les relations d'épistasie et de régulation à l'intérieur du réseau constitué de ces gènes. Toutes ces modifications sont bien évidemment à mettre en corrélation avec des mécanismes de

développement sensiblement différents.

2.2.2. *Choix d'un organisme modèle chez les lophotrochozoaires*

Avant de déterminer le rôle des gènes *snail* et *twist* chez les lophotrochozoaires, il fallait choisir un organisme représentatif de ce groupe sur lequel faire porter notre étude.

Nous avons donc réalisé en parallèle chez deux Annélides (*Enchytraeus albidus* et *Sabellaria alveolata*) et deux Mollusques (*Mytilus galloprovincialis* et *Patella vulgata*), une première étude de faisabilité afin d'estimer la possibilité de se procurer des animaux matures, d'étudier les conditions d'entretien des cultures et de mesurer la facilité d'obtention et de manipulation des embryons. Le mollusque *Patella vulgata* s'est alors avéré être l'espèce qui répondait le mieux à l'ensemble de ces critères. Nous avons donc décidé de concentrer l'essentiel de notre travail sur cet organisme.

Le mollusque *Patella vulgata* possède un mode de développement de type spiral dont l'une des principales caractéristiques embryologiques est que le mésoderme dérive, pour l'essentiel, d'une cellule souche unique: le mésentoblaste (4d) (Figure 3). L'expression de gènes connus pour leur fonction mésodermique dans ce blastomère constituerait donc un argument fort en faveur de l'homologie du mésoderme chez les trois grands groupes de métazoaires.

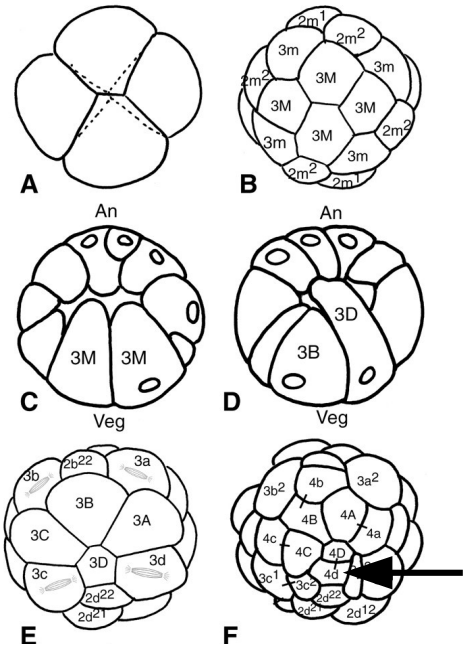


Figure 3 : *Segmentation spirale de Patella vulgata. Les noms des différents blastomères sont indiqués. (A) stade 4 cellules. (B, C et D) stade 32 cellules. (E) stade 60 cellules. (F) stade 64 cellules. La position du mésentoblaste 4d est indiquée par une flèche (d'après Lartillot et al., 2002).*

La recherche et la caractérisation des gènes de la famille s*nail* chez le mollusque *Patella vulgata* m'a cependant conduit à revoir cette hypothèse sans doute trop naïve du rôle de la fonction "spécificateur du mésoderme" pour ce gène et à éclairer ce problème sous un jour nouveau.

2.2.3. Principaux résultats

Un homologue du gène *twist* et deux homologues du gènes *snail* ont été isolés et séquencés chez *Patella vulgata*.

Le gène orthologue de *twist* chez la Patelle (*Pv-twist*) s'exprime chez la larve trochophore dans une partie du mésoderme (ectomésoderme), supportant ainsi l'idée que ce gène est impliqué dans la différenciation du mésoderme et confortant ainsi son rôle ancestral supposé. L'absence de *Pv-twist* dans l'endomésoderme suggère cependant que d'autres gènes qui restent à découvrir sont également impliqués dans la différenciation d'au moins une autre partie du mésoderme (Nederbragt *et al.*, 2002)

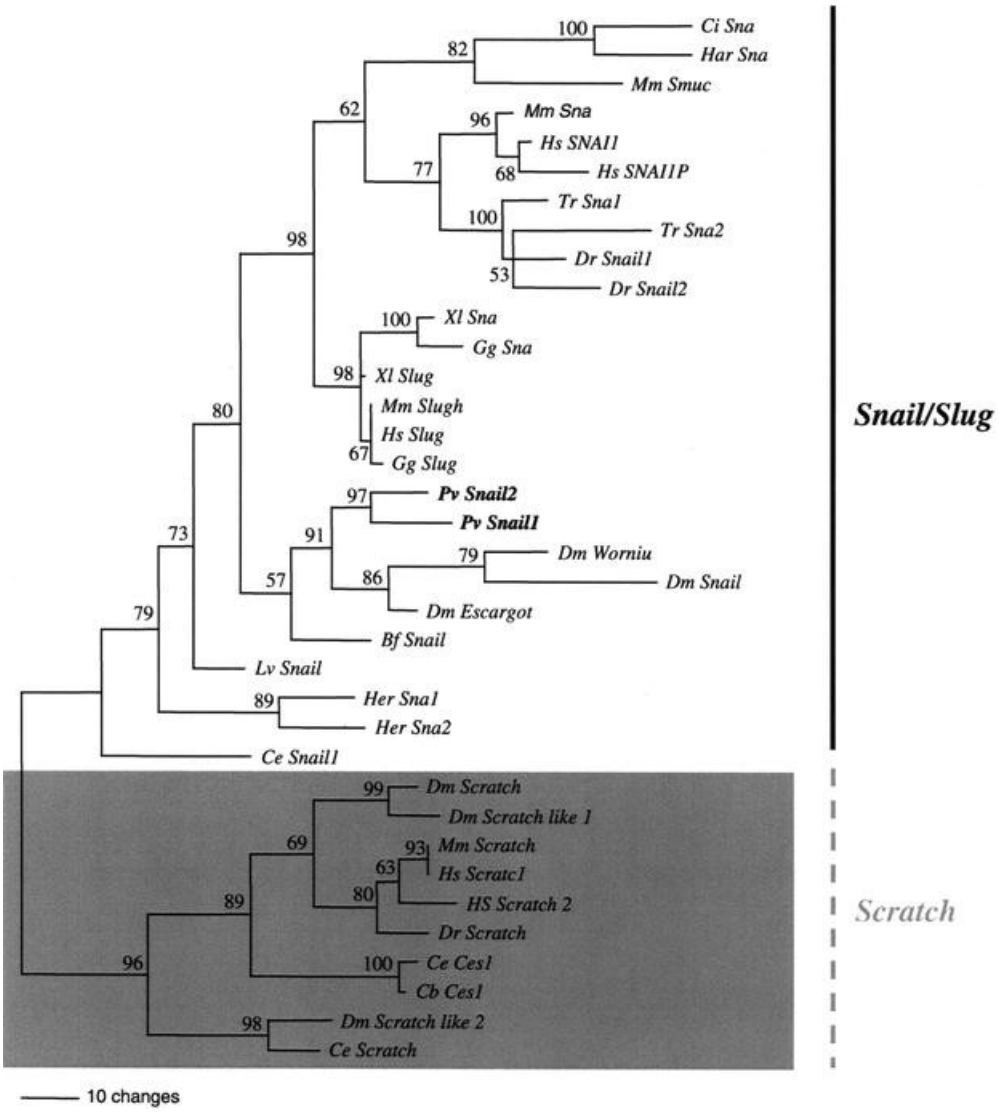


Figure 4 : *Arbre de maximum de vraisemblance des gènes snail. Après édition manuelle de l'alignement des séquences, l'analyse a été réalisée à l'aide du programme ProtML (Adachi and Hasegawa, 1996) sous le modèle JTT-F (Jones et al. 1992). Les probabilités de bootstrap (RELL - BP, 10,000 replications; Hasegawa and Kishino 1994) supérieures à 50% sont indiquées sur les noeuds. Bf Branchiostoma floridae, Cb Caenorhabditis briggsae, Ce Caenorhabditis elegans, Ci Ciona intestinalis, Dm Drosophila melanogaster, Dr Danio rerio, Gg Gallus gallus, Har Halocynthia roretzi, Her Helobdella robusta, Hs Homo sapiens, Pv Patella vulgata, Lv Lytechinus variegatus, Mm Mus musculus, Tr Takifugu rubripes, Xl Xenopus laevis (d'après Lespinet et al., 2002a).*

La caractérisation de deux paralogues du gène *snail* (*Pv-snail1* et *Pv-snail2*) chez la Patelle nous indique que ce gène est également conservé chez les lophotrochozoaires. S'il est difficile de définir sans ambiguïté les relations d'orthologie qui peuvent exister entre les gènes de la famille *snail*, l'approche phylogénétique nous a permis néanmoins d'établir l'existence de plusieurs duplications survenues de manière indépendante au cours de l'évolution de ces gènes *(Figure 4)* (Lespinet *et al.*, 2002a). En ce qui concerne les deux gènes de *Patella vulgata*, on constate que la duplication est relativement récente et en tous cas postérieure à la séparation entre lophotrochozaires et ecdysozaoaires (Figure 4).

Les profils d'expression obtenus chez la Patelle pour ces deux gènes nous conduisent aujourd'hui à formuler de nouvelles hypothèses sur la fonction du gène ancestral. Ils nous mènent en effet à rejeter l'hypothèse d'une fonction ancestrale de type "spécificateur du mésoderme", car seul l'un des deux gènes (*Pv-snail1*) semble s'exprimer dans un petit nombre de cellules qui pourraient être d'origine mésodermique (Figure 5). Cette expression mésodermique est cependant très incertaine et en tous cas très postérieure à la formation du mésentoblaste. Il s'agit plus vraisemblablement d'une spécificité mésenchymateuse plutôt qu'une spécificité mésodermique. On retrouve également ce type de spécificité chez le second gène (*Pv-snail2*) puisqu'il s'exprime dans les cellules des cônes d'invagination du manteau (ectoderme) (Lespinet *et al.*, 2002a).

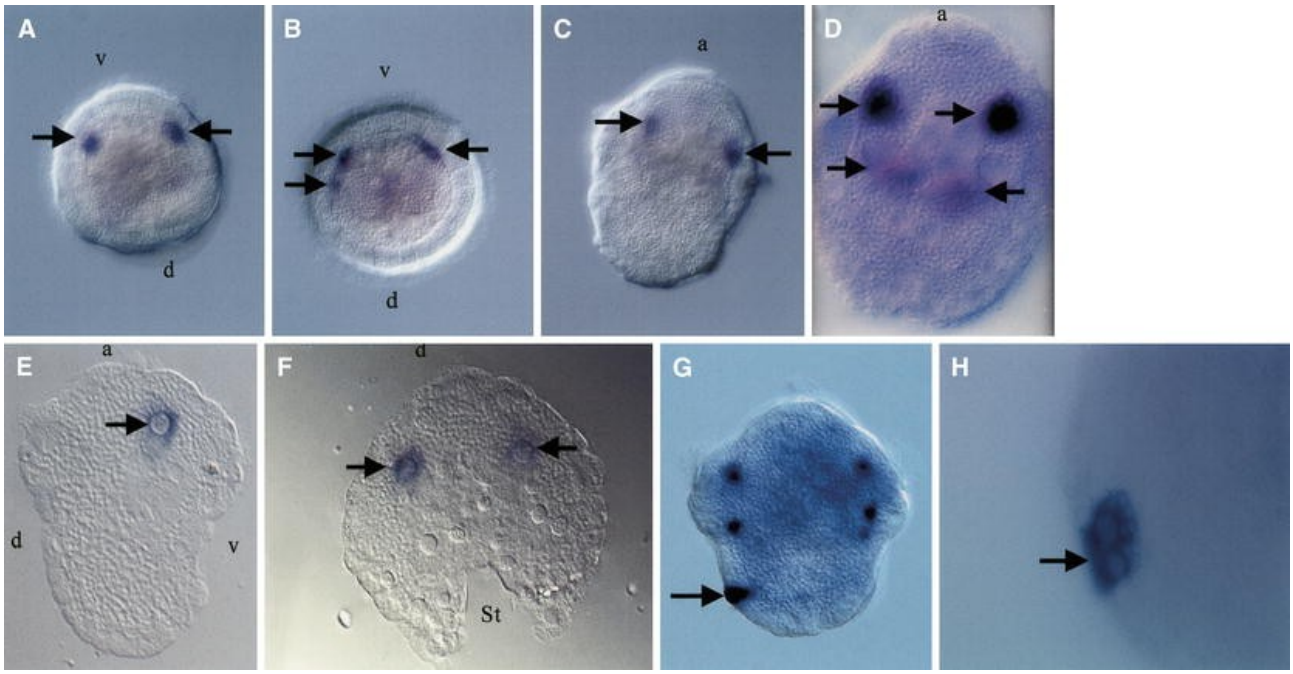


Figure 5 : *L'expression des messagers de Pv-snail1 sont indiqués par des flèches. A, B : vue apicale. C, D : vue ventrale. E : coupe d'une vue latérale. F : coupe d'une vue apicale. A-F : larve trochophore âgée de 16 heures. G-H : Larve trochophore de 24 heures. G : vue ventrale. H : grossissement de la coloration du pied.*

Des données obtenues chez la souris (Cano *et al.*, 2000) et l'homme (*Batlle et al.*, 2000) ont montré que chez ces organismes, les gènes *snail* régulaient négativement l'expression de molécules connues pour leur rôle central dans les processus d'adhésion et de mobilité cellulaire : les E-cadhérines. L'expression des gènes *Pv-snail1* et *Pv-snail2* dans les zones de croissance des bandes mésodermiques d'une part, dans les cônes d'invagination du manteau d'autre part, ainsi que dans les cellules de la plaque neurale peut en effet s'accorder avec l'idée d'une action de ces gènes dans la régulation de l'expression des cadhérines et le contrôle des transitions épithéliale-

mésenchymateuses. De la même façon, le rôle des gènes de la famille *snail* chez la Drosophile, au cours de la gastrulation, lors de la délamination des neuroblastes, ou bien encore lors de la formation du mésoderme, peut tout à fait être ré-interprété en terme de régulateur des transitions épithéliale-mésenchymateuses. Le phénotype des mutants pouvant alors s'expliquer en partie par des défauts de migrations cellulaires.

Nos observations chez la Patelle ont conduit à émettre l'hypothèse d'une fonction ancestrale du gène *snail* dans le contrôle des transitions épithéliale-mésenchymateuses.

L'existence de nombreuses duplications indépendantes au sein de la famille des gènes *snail*, conjuguée à une très faible conservation des séquences régulatrices de ces protéines nous ont également conduit à proposer, pour cette famille de gènes, un scénario d'évolution par duplication/dégénérescence/complémentation (DDC) basé sur le modèle de Force (Force *et al.*, 1999). La variabilité des séquences régulatrices jouerait alors un rôle essentiel dans la sélection des partenaires et la sous-fonctionnalisation de la fonction initiale de ce facteur de transcription. Ainsi, les fonctions spécifiques des différents gènes *snail* auraient été acquises par répartition des fonctions du gène ancestral entre les différents paralogues. On assisterait donc, après duplication, à une spécialisation de chacune des copies de ce gène (Lespinet *et al.* 2002a).

L'ensemble de ses résultats ont été publiés sous la forme de trois articles dans les journaux du domaine, à savoir : Development Gene and Evolution (Lespinet *et al.*, 2002a), Evolution & Development (Nederbragt *et al.*, 2002), et Development (Lartillot *et al.*, 2002).

2.3. Duplication de gènes et évolution des génomes *(Bethesda, MD, USA : 2001-2002)*

L'étude de la famille *snail* m'a conduit à m'interroger sur la fonction des gènes dupliqués au sein des génomes et à m'intéresser d'un peu plus près aux aspects dynamiques des génomes et plus particulièrement aux duplications de gènes. Cet intérêt m'a orienté dans le choix de mon séjour post-doctoral et début 2001 j'ai donc décidé de rejoindre le groupe d'Eugene Koonin (NCBI, Bethesda USA) afin de compléter ma formation initiale de biologiste expérimental, spécialiste de l'évolution et du développement des métazoaires, par une réelle compétence en bioinformatique.

Afin de mieux comprendre l'importance des duplications de gènes au cours de l'évolution des eucaryotes, j'ai réalisé une étude détaillée des groupes de paralogues spécifiques de chacun des 5 génomes suivants: *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Caenorhabditis elegans*, *Drosophila melanogaster* et *Arabidopsis thaliana*.

Plus qu'aux mécanismes conduisant à l'existence de gènes présents en de nombreuses copies, mon objectif était alors d'étudier la nature des gènes dupliqués, l'ampleur des duplications au sein d'une classe de gènes ainsi que les conséquences biologiques de ces amplifications en terme d'avantages sélectifs et d'adaptations pour les organismes chez lesquels elles se produisent.

2.3.1. Méthodologie.

Le protocole utilisé est tout à fait similaire de celui utilisé par King Jordan pour étudier les duplications spécifiques chez les procaryotes (Jordan *et al.*, 2001).

A l'exception du génome du nématode qui provient de la base WormPep20, toutes les séquences proviennent de la base de données protéiques non redondante (nr) du NCBI.

L'identification des LCSs (Lineage-Specific Clusters) a été réalisée en suivant le protocole suivant. A l'aide du programme BLAST (Altschul *et al.*, 1990), chaque séquence a été comparée contre une banque de données constituée de l'ensemble des séquences des 5 génomes étudiés ici. Les protéines correspondant aux meilleurs hits appartenant au génome de la protéine requête et pour lesquels le score est supérieur au meilleur hit provenant d'un autre génome sont conservées et constitueront les graines à partir desquelles seront constituées les LCSs. Les graines seront alors agglomérées en familles si elles possèdent un élément en commun (algorithme de lien simple). Nous obtenons donc des groupes de protéines ressemblantes, dont on pense qu'elles sont homologues, et appartenant toutes au même génome. Cette méthode ayant tendance à produire de très grandes familles elle sera suivie d'une phase de filtre des résultats.

La protéine la plus ressemblante (meilleur score d'identité) en provenance d'un autre génome est incorporée au groupe candidat et une procédure de clustering par UPGMA (Sokal et Michener, 1958) est appliquée. Soit la protéine en provenance d'un autre génome se positionne en groupe externe à la base de l'arbre obtenu confirmant ainsi la validité du groupe trouvé précédemment, soit l'introduction de cette protéine conduit à l'obtention de plusieurs sous groupes. Dans ce dernier cas, seuls les sous groupes de taille supérieure à 2 et ne contenant que des protéines provenant du même génome seront conservés comme LCSs.

Chaque protéine de chaque LSC dont la taille était supérieure à 2 a été comparée par PSI-BLAST (Altschul *et al.*,1996) contre la totalité de la banque nr (non-redundant protein database). Les annotations fonctionnelles ont alors été récupérées à ce stade. La recherche de domaines structuraux connus a été réalisée en comparant chaque protéine contre la banque de domaines CDD (Conserved

Domain Database) développée par Stephen Bryant (Marchler-Bauer *et al.*, 2002).

2.3.2. Principaux Résultats.

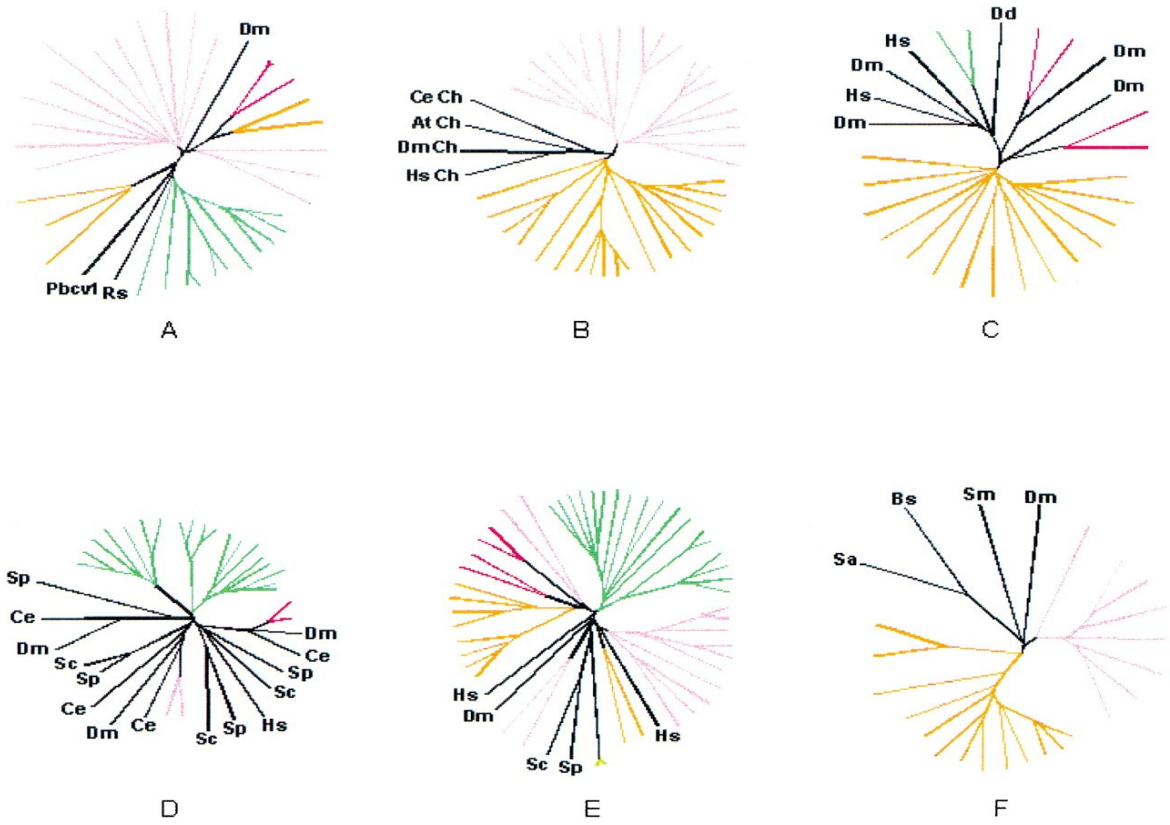


Figure 4 : *Arbres phylogénétiques de quelques LSEs. Les groupes pour lesquels la valeur de bootstrap est supérieure à 70 sont colorés : en rose pour Drosophila melanogaster, rouge pour Homo sapiens, orange pour Caenorhabditis elegans, vert pour Arabidopsis thaliana et jaune pour Schizosaccharomyces pombe. (A) Prolyl hydroxylases. (B) Small molecule kinases (Ch pour choline kinase). (C) Patched-like protein. (D) MAP-Kinases. (E) P450 family hydroxylases. (F) MBOAT membrane acyltransferases. (At) Arabidopsis thaliana, (Bs) Bacillus subtilis, (Ce) Caenorhabditis elegans, (Dd) Dictyostelium discoideum, (Dm) (Drosophila melanogaster, (Hs) Homo sapiens, (Pbcv1) Paramecium bursaria Chlorella virus 1, (Rs) Ralstonia solanacearum, (Sa) Staphylococcus aureus, (Sc) Saccharomyces cerevisiae, (Sm) Sinorhizobium meliloti, (Sp) Schizosaccharomyces pombe (d'après Lespinet et al., 2002b).*

Pour chacune des espèces inclues dans notre étude, 10 LSCs ont été choisi pour construire un arbre phylogénétique. Six des 50 arbres obtenus sont représentés sur la Figure 4. On observe que pour chacun de ces arbres il s'agit bien d'expansions géniques postérieures à la séparation d'avec l'espèce la plus proche. On observe également que pour certaines des protéines présentées ici, des expansions indépendantes se sont produites dans plusieurs espèces (Figure 4).

L'analyse détaillée des LSCs obtenus par notre procédure montre qu'il existe une corrélation linéaire entre le nombre de protéines déduites d'un génome et la fraction de ces protéines constituées en groupes de paralogie récents. Le pourcentage de protéines appartenant à un LSC varie de 20% (*Saccharomyces cerevisiae*) à 80% (*Arabidopsis thaliana*). En outre, plus le nombre de groupes de paralogie est important pour une espèce donnée et plus le nombre moyen de gènes

constituant chacun de ces groupes est grand. Enfin, nous avons calculé un coefficient d'expansion (CE) au sein de chaque famille de paralogues. Ce CE correspond au ratio entre le nombre de gènes inclus dans un LSC sur le nombre total de gènes présent dans la famille. Pour chaque famille de paralogues, ce CE représente donc la fraction de la famille qui résulte uniquement de LSE (Lineage-Specific gene family Expansion). La distribution des CE montre qu'environ 40% des LSC présentent un CE supérieur à 0,9 indiquant que la plupart de ces familles multigéniques sont apparues quasi-exclusivement par LSE (Figure 5).

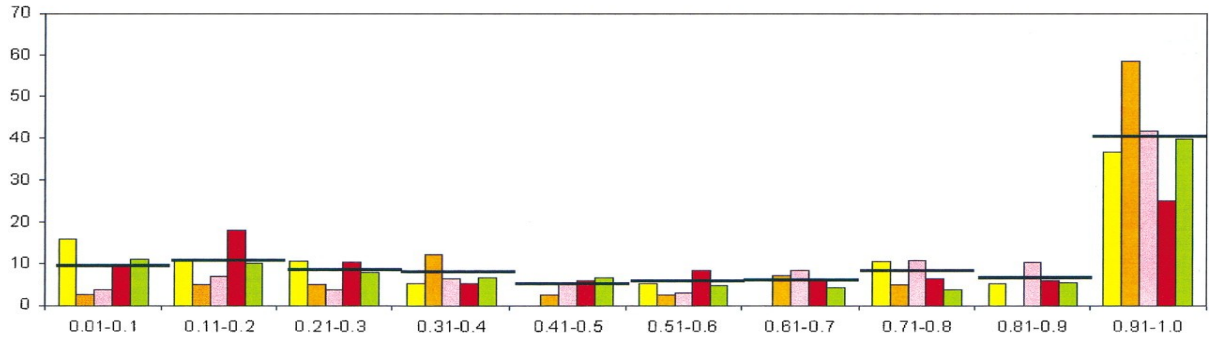


Figure 5 : *Distribution des LSCs en fonction de la valeur du coefficients d'expansion (CE). Les classes de CE sont indiquées en abscisse et le pourcentage de LSCs en ordonné. (jaune) Schizosaccharomyces pombe, (orange) Saccharomyces cerevisiae, (rose) Drosophila melanogaster, (rouge) Caenorhabditis elegans et (vert) Arabidopsis thaliana. Pour chaque classe, la moyenne des valeurs des cinq espèces est indiquée par un trait horizontal (d'après Lespinet et al., 2002b).*

Si l'on s'intéresse aux groupes de LSE les plus importants au sein de chaque génome, on constate qu'ils présentent une grande diversité de fonctions. Les groupes présentant les LSE les plus abondant semblent être constitués de facteurs de transcriptions, de protéines de réponses au stress et aux pathogènes, de protéines impliquées dans les voies de dégradation, de protéines impliquées les voies de transduction du signal ainsi que de protéines de type chémo-récepteurs. L'ensemble de ces protéines constituent des facteurs d'interaction et de réponse aux modifications des conditions du milieu dans lequel vive les organismes étudiés.

Les résultats détaillés de cette analyse ont été publiés en 2002 dans Genome Research (Lespinet *et al.*, 2002b).

2.4. Evolution des Génomes et du Métabolisme : Annotation, Analyse et Exploration de données. *(Orsay : 2002 à aujourd'hui)*

Depuis mon recrutement comme maître de conférences et mon intégration en 2002 à l'Institut de Génétique et Microbiologie, j'ai développé ou participé au développement de plusieurs projets au sein de l'équipe Evolution Moléculaire et Bioinformatique des Génomes dirigée par Bernard Labedan. J'ai également collaboré avec plusieurs équipes de l'Institut sur des projets que je ne décrirai pas ici car ma contribution y est somme toute assez marginale (Bidard *et al.* 2010; Poitelon *et al.*, 2009; Bergemann *et al.*, 2008).

Dans les paragraphes qui vont suivre je me suis donc attaché à décrire cinq des projets que je trouve les plus représentatifs de mes travaux au cours de ces huit dernières années. Les projets que je vais détailler s'articulent autour de trois grands thèmes : (1) Evolution des génomes, (2) Annotation et Analyse du Génome, (3) Exploration de données.

2.4.1. Vers la construction d'un atlas des modules protéiques

Au milieu des années 1990 Monica Riley et Bernard Labedan ont développé le concept de segment protéique structural homologue ou module protéique (Riley et Labedan, 1997). Ces modules sont les briques évolutives élémentaires qui constituent les protéines modernes (figure 6).

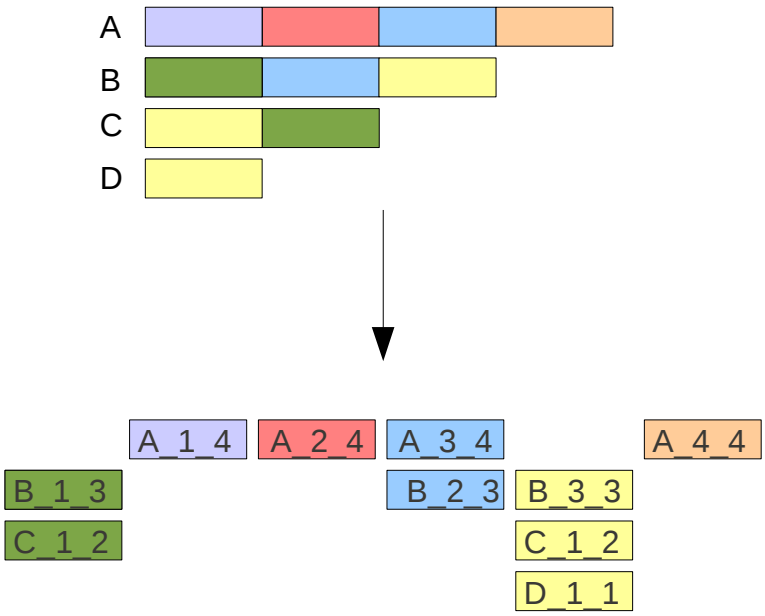


Figure 6 : *Les protéines A, B, C et D sont décomposées en 10 modules structuraux homologues. Les 6 différents types de modules sont repérés par leurs couleurs, les modules homologues étant de la même couleur.*

Bernard Labedan et ses collaborateurs allaient par la suite développer une approche originale pour reconstruire la dynamique et l'évolution de ces modules protéiques (de Rosa et Labedan, 1998; Le Bouder-Langevin *et al.*, 2002). Leur approche était basée sur l'utilisation du logiciel DARWIN (Gonnet *et al.*, 2000) pour la réalisation d'une comparaison exhaustive des protéomes déduits des génomes de bactéries et d'archées. Le programme DARWIN réalise un alignement local de toutes

les paires de protéines possibles en utilisant l'algorithme de Smith et Waterman (Smith et Waterman, 1981). En dehors de la qualité de l'alignement obtenu, DARWIN permet également de calculer la distance PAM entre les deux protéines comparées. Cette distance reflétant la distance évolutive entre les deux protéines est obtenue par maximum de vraisemblance en recherchant la matrice PAM pour laquelle le score de similarité obtenu pour un alignement sera maximal (Gonnet *et al.*, 2000).
La distance PAM et la longueur de l'alignement sont les deux critères qui ont été choisis pour la détection des modules structuraux homologues. Ainsi la longueur minimale d'un module structural homologue doit être au minimum de 80 acides aminés et la distance PAM entre deux séquences protéiques doit être inférieure à 250 unités PAM. En dessous de 80 acides aminés les régions ressemblantes risquent de correspondre à des motifs ou des domaines fonctionnels et non plus à des segments structuraux homologues (modules). En dessous de 250 unités PAM la ressemblance des séquences est trop incertaine pour que l'on puisse conclure de façon certaine à leur homologie.

Lors de mon arrivée dans l'équipe j'ai tout naturellement rejoint ce projet et j'ai en particulier ré-écrit un certain nombre des scripts qui péchaient par leur lenteur. J'ai également travaillé à élaborer une définition plus précise des modules et à améliorer l'étape de construction des familles de modules. L'objectif étant de permettre à la chaine des programmes utilisée pour détecter les modules protéiques de fonctionner avec des jeux de données beaucoup plus importants que ce qui avait été réalisé jusque là, c'est-à-dire quelques dizaines voire une centaine de génomes au lieu de quelques génomes.

La constitution des familles de modules est basée sur un algorithme de lien simple. Ainsi dans l'exemple de la figure 6 les résultats de DARWIN ont mis en évidence l'existence de 10 modules constitutifs des protéines A, B, C et D. Au sein de chacune des protéines, les différents modules sont identifiés par une nomenclature composée du nom de la protéine qui possède ce module, suivi de deux numéros qui indiquent à quelle position relative se trouve ce module dans la séquence de la protéine et quelle fraction de la protéine est représentée par ce module. Ainsi le module A_3_4 est constitutif de la protéine A et correspond au troisième quart de cette protéine (figure 6). Sur la figure 6 seuls les modules verts, bleus et jaunes sont communs à plusieurs protéines. C'est-à-dire que leurs séquences possèdent un niveau de similarité tel qu'ils seront considérés comme de probables homologues. Les modules mauve, rose et orange ne présentent pas d'homologues ils ne répondent donc pas à la définition des modules structuraux homologues et correspondent à ce que l'on appelle des modules déduits dont on postule l'existence car un autre module a été identifié dans la protéine où ils sont présents. L'algorithme de lien simple appliqué à ses résultats permet donc de construire trois premières familles constituées respectivement des modules B_1_3 et C_1_2, des modules A_3_4 et B_2_3 et des modules B_3_3, C_1_2 et D_1_1. On peut également proposer trois autres familles constituées chacune d'un seul module déduit (A_1_4, A_2_4 et A_4_4). Au final, notre algorithme de détection des modules et de construction des familles permet donc sur cet exemple de proposer l'existence de 6 familles distinctes (Figure 6).

Les résultats obtenus en comparant les résultats obtenus à partir des protéomes déduits de bactéries et d'archées ont montré que la plupart des protéines étaient multi-modulaires. La taille moyenne d'un module est de 220 acides aminés ce qui est une valeur proche de la taille des domaines structuraux qui elle est estimée à 150±50 acides aminées (Wheelan *et al.*, 2002).

D'un point de vue de l'évolution des génomes, chaque famille de module provient soit d'un module ancestral soit d'un module spécifique à une espèce donnée. L'étude de la distribution taxonomique des modules nous permet donc de décrire 4 classes de modules : les modules spécifiques d'une espèce présents en une copie unique (uni-sp), les modules spécifiques d'une espèce présents en plusieurs copies (para-sp), les modules communs à plusieurs espèces qui pour certaines espèces

peuvent être présents en plusieurs copies (para-ortho), et enfin les modules communs à plusieurs espèces mais pour lesquels une seule copie par espèce a été détectée (uni-ortho).

L'étude des différentes classes de modules (Figure 7) nous permet de dire que leur distribution semble varier avec le mode de vie de chaque espèce. Les espèces pathogènes qui possèdent un génome réduit (*H. influenzae et V. cholerae*) présentent une plus faible proportion de modules dans la classe para-ortho alors que la classe uni-sp semble être plus représentée pour ces mêmes espèces (Figure 7). Les bactéries non pathogènes présentent quant à elle une plus forte proportion des classes para-ortho et para-sp (Figure 7A). Les modules spécifiques des espèces pathogènes semblent donc être présents en un petit nombre de copies au sein de chacun de ces protéomes.

Lorsque l'on compare un plus grand nombre de génomes (Figure 7B), on constate que pour les même espèces, le nombre de gènes uni-sp et para-sp diminue, alors que le nombre de gènes uni-ortho et para-ortho semble augmenter.

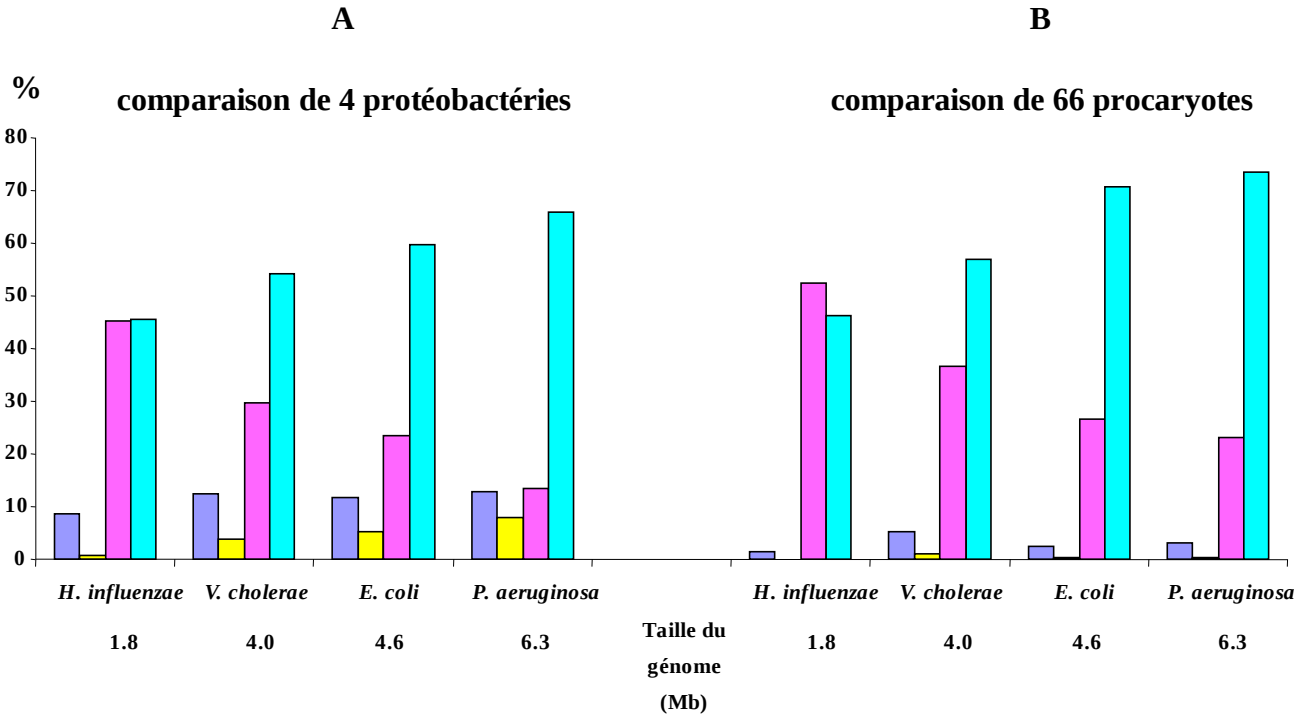


Figure 7 : *Distribution (en %) des 4 classes de familles de modules après comparaison intra et inter génomique de 4 protéobactéries (A) et de 66 procaryotes (B). La classe uni-sp est représentée en violet, la classe para-sp en jaune, la classe uni-ortho en rose et la classe para-ortho en bleu. Les espèces représentées sur ce graphique sont Haemophilus influenzae (H. influenzae), Vibrio cholerae (V. cholerae), Escherichia coli (E. coli) et Pseudomonas aeruginosa (P. aeruginosa) (d'après Descorps-Declère et al., 2008).*

Les groupes de modules obtenus ont également été utilisés pour faire de l'annotation fonctionnelle (Sculo *et al.*, 2003; Descorps-Declère *et al.*, 2008) et pour mesurer les distances évolutives relatives séparant les espèces comparées (Sculo *et al.*, 2005).

L'étude de la distribution taxonomique des familles de modules nous permet également de reconstituer l'histoire des protéines et d'identifier les évènements anciens de duplication (Figure 8) (Sculo *et al.*, 2003 ; Descorps-Declère *et al.*, 2008).

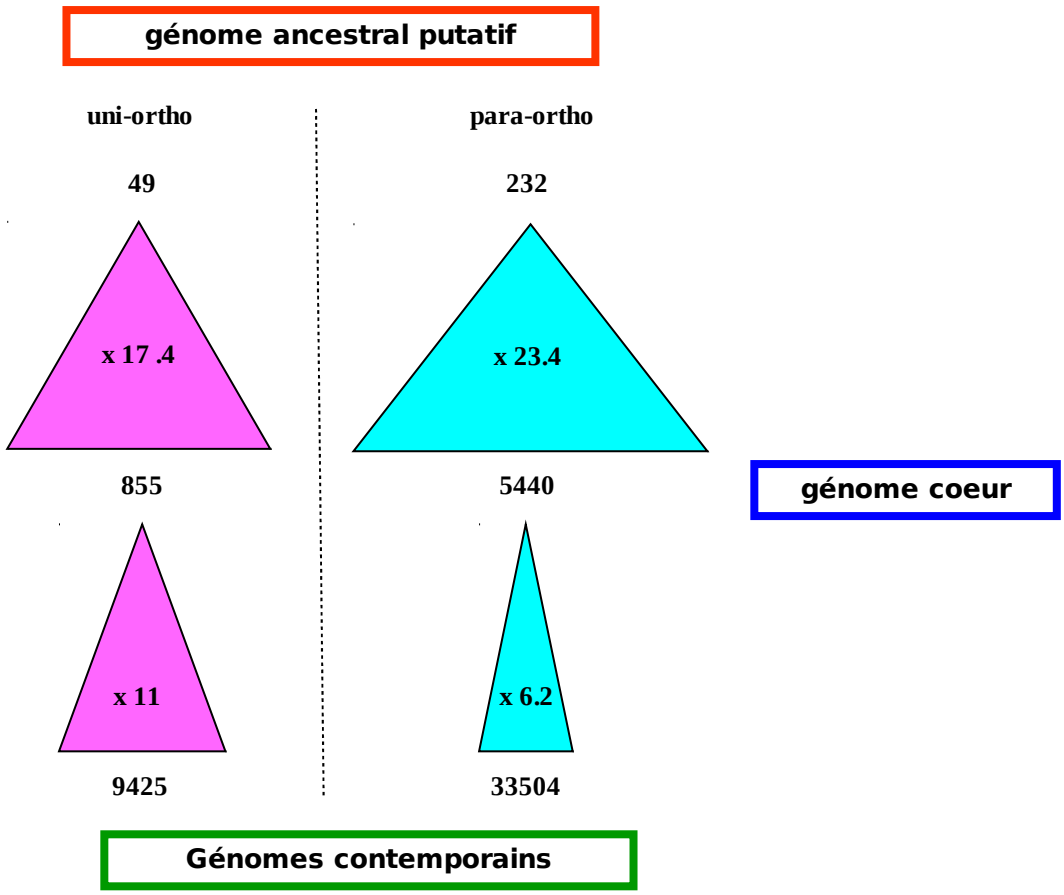


Figure 8 : *Evolution des modules de classes uni-ortho et para-ortho chez 15 gammaprotéobactéries (d'après Descorps-Declère et al., 2008).*

L'analyse des modules protéiques de 15 gammaprotéobactéries (Figure 8) montre que les classes uni-ortho et para-ortho ont subit un important taux d'amplification depuis le génome ancestral constitué de 281 modules. On notera que la phase d'amplification est moins importante entre le génome coeur et les génomes contemporains qu'entre le génome coeur et le génome ancestral. Ces duplications importantes se sont accompagnées de nombreux réarrangements via des phénomènes de fusion de modules ancestraux (Sculo *et al.*, 2003 ; Descorps-Declère *et al.*, 2008).

L'ensemble des données obtenues depuis ma participation à ce projet ont été publiées dans les revues Genomes Letters (Sculo *et al.*, 2003), Biochimie (Descorps-Declère *et al.*, 2008), dans un chapitre du livre Microbial Proteomics: Functionnal Biology of Whole Organisms édité par Ian Humphery-Smith et Michael Hecker (Labedan et Lespinet, 2006) ainsi que dans les actes de JOBIM (Sculo *et al.*, 2005).

L'analyse de nos résultats a également fait l'objet d'un projet ANR porté par Bernard Labedan. Dans le cadre de ce projet intitulé « microbiogenomics », j'ai participé à la construction d'un atlas des modules protéiques. Cet outil devrait, à terme, nous permettre de retracer l'évolution des modules et de leurs fonctions. Ce projet s'intègre dans la cadre d'une collaboration avec les équipes de Jean-François Gibrat (INRA, Jouy-en-Josas), de Christine Froidevaux et de Jean-Daniel Fekete (LRI/INRIA, Orsay).

2.4.2. Evolution et synténie chez les procaryotes

Lorsque l'on étudie les familles de gènes dupliqués (LSE ou autres) ou bien les familles de modules protéiques des classes para-sp et para-ortho on peut raisonnablement se poser la question de la localisation sur le chromosome de ces gènes ou de ces modules. Nous manquions cependant d'outils pour pouvoir comparer d'aussi grande quantité de données que celles auxquelles nous avions accès. Il n'existait en effet pas de visualiseurs de la synténie qui permettent de comparer simultanément et aisément plusieurs dizaines de génomes.

Dans le cadre du stage de Master 2 de Frédéric Lemoine, puis de sa thèse, nous avons entrepris de construire un tel outil de visualisation. Si d'un point de vue officiel la thèse de Frédéric était co-encadrée par Bernard Labedan et Christine Froidevaux, d'un point de vue officieux j'ai en ce qui me concerne supervisé son travail d'obtention et de visualisation des données de synténie.

Quelle que soit l'échelle à laquelle on se place, l'obtention des informations de synténie requiert d'obtenir un certain nombre d'informations. Si la première d'entre elles est l'ordre des gènes sur le génome, la seconde est l'obtention de groupes d'orthologues pour les génomes sur lesquels nous ferons porter notre étude.

Or, si la construction de groupes d'orthologues est un processus clé de la génomique comparée il n'en reste pas moins qu'il s'agit d'un exercice difficile. Bien que de nombreuses méthodes différentes ont été développées pour détecter les orthologues, aucune n'offre encore aujourd'hui la garantie de la qualité des résultats (Altenhoff et Dessimoz, 2009). La méthode la plus connue est celle basée sur les BLAST Reciprocal Best Hit (BRH) (Overbeek *et al*., 1999). Cette méthode est cependant connue pour générer un nombre important d'erreurs (Koski et Golding, 2001 ; Fulton *et al.*, 2006). Nous avons donc entrepris de développer des méthodes alternatives dont nous pensions qu'elles seraient plus efficaces.

Dans le cadre de ce projet nous avons développé deux méthodes différentes d'obtention des groupes d'orthologues. L'une d'entre elles est assez ressemblante à celle de Reciprocal Smallest Distance (RSD) (Wall *et al.*, 2003) et l'autre consiste en une nouvelle approche basée sur l'analyse d'arbres phylogénétiques.

Dans les deux cas, le préalable était de réaliser une comparaison exhaustive de toutes les séquences des 107 génomes (92 bactéries et 15 archées) que nous avions décidé d'analyser. Ces comparaisons ont été réalisées à l'aide du programme DARWIN (Gonnet *et al.*, 2000). Pour considérer que deux gènes sont orthologues nous avons imposé que leur distance PAM soit inférieure à 250 et que la région alignée entre les deux séquences représente au moins 80% de la longueur de la plus petite des deux protéines alignées. La région alignée doit en outre être d'une taille supérieure à 80 acides aminés. A l'issue de cette première étape nous avons obtenu des couples de protéines dont on pense qu'il s'agit d'homologues et dont on va essayer de déterminer par la suite s'il s'agit d'orthologues.

Dans le cas de la méthode RSD si la distance PAM entre les deux protéines A et B, appartenant respectivement aux génomes a et b, est d'une part inférieure aux distances PAM obtenues entre A et toutes les autres protéines du génome b et d'autre part inférieure aux distances PAM obtenues entre B et toutes les autres protéines du génome a, alors nous en concluons que A et B forment réellement un couple d'orthologues. Cette méthode a néanmoins le défaut de ne proposer que des relations d'orthologie de type 1-1 alors que dans le cas ou pour l'un des gènes (ou les deux) il existe des paralogues récents postérieurs à l'évènement de spéciation (in-paralogues) (Remm *et al.*, 2001), il devrait être possible d'envisager des relations d'orthologie de type 1-n ou même de type n-n. Afin

d'essayer de régler ce problème, nous avons donc développé une seconde méthode basée sur la phylogénie.

Dans le cas de la méthode basée sur l'analyse d'arbres phylogénétiques, les paires de protéines retenues à l'issu de l'étape de comparaison des génomes sont regroupées en familles multigéniques par une approche de lien simple. L'utilisation du regroupement par lien simple présente néanmoins l'inconvénient de produire quelques très grosses familles constituées de sous groupes homogènes reliés entre eux par un petit nombre de protéines pour constituer un groupe finalement hétérogène. L'application de l'algorithme MCL développé par Stijn von Dongen au cours de sa thèse (van Dongen, 2000 ; Enright *et al.*, 2002) nous a permis de pallier à ce problème et de décomposer les plus grosses familles en sous familles homogènes. Les séquences de chacune des familles ont alors été alignées à l'aide du logiciel MUSCLE (Edgar, 2004), puis des arbres phylogénétiques ont été obtenus à l'aide du logiciel PhyML (Guindon et Gascuel, 2003). L'analyse des arbres a alors été réalisée pour obtenir la liste des groupes d'orthologues. A la différence d'autres approches basées sur la phylogénie (Hubbard *et al.*, 2007) notre méthode d'analyse n'impose pas de disposer d'un arbre des espèces pour pouvoir obtenir les groupes d'orthologues (Figure 8).

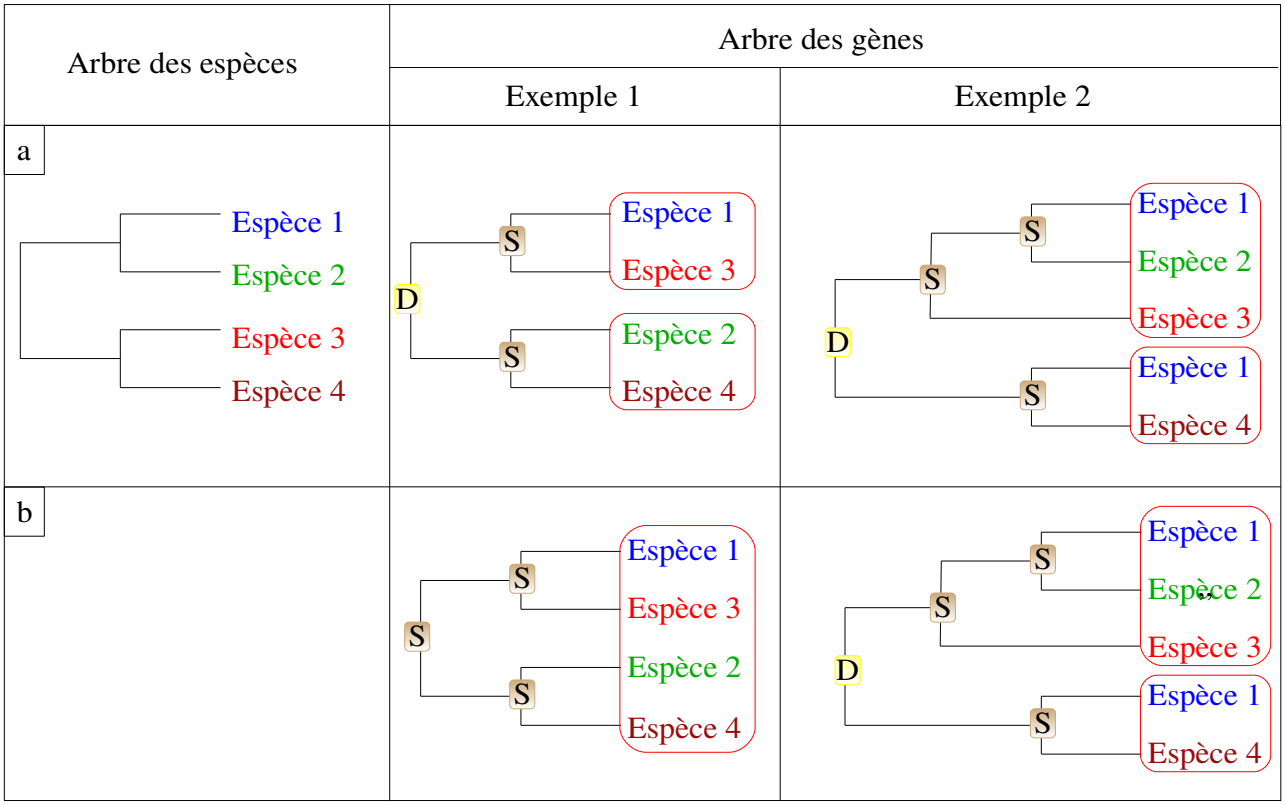


Figure 8 : *Construction de groupes orthologues par analyse phylogénique. a - Analyse phylogénique se basant sur l'arbre des espèces. b - Analyse phylogénique n'utilisant pas l'arbre des espèces. Les chiffres et les couleurs représentent différentes espèces. Les évènements évolutifs sont représentés par les lettres S pour spéciation (carré marron) et D pour duplication (carré jaune). Les rectangles au contour rouge représentent les groupes d'orthologues déduits après analyse (d'après Grossetête, 2010).*

Dans le cas où l'arbre des gènes est comparé avec l'arbre des espèces, les événements de duplication sont détectés lorsqu'apparaissent des incongruences entre les deux arbres (Figure 8a). Dans le cas de la méthode que nous proposons, les duplications sont détectées lorsque deux gènes appartenant à la

même espèce appartiennent à un même groupe. Il suffit alors d'analyser tous les groupes en partant des feuilles terminales et en remontant dans les branches les plus profondes de l'arbre pour pouvoir détecter tous les évènements de duplication.

Les deux méthodes que nous avons développées pour la détection d'orthologues donnent des résultats différents mais présentent cependant environ 68% de recouvrement.

Une fois que les orthologues ont été identifiés, nous avons recherché ceux dont le voisinage est conservé. On peut ainsi définir le niveau de conservation des blocs de synténie comme une fonction de la distance taxonomique qui séparent les espèces à travers l'ensemble de l'arbre du Vivant des microorganismes (Tableau 1).

Tableau 1 *: La moyenne des distances PAM entre paire d'orthologues et la taille moyenne des blocs de synténie dépendent de la distance taxonomique des deux espèces comparées (d'après Descorps-Declère et al., 2008).* [a] *Taille du protéome de E. coli K12: 4279.*

Rang	**Espèce 1 (*E. coli*)[a] taxonomie**	**Espèce 2**			**Moyenne des distances PAM**	**Blocs de Synténie**	
		Nom d'espèce	**Taxonomie**	**Taille du Protéome**		**Taille moyenne**	**Plus long bloc**
Famille	Enterobacteriaceae	*S. enterica*	Enterobacteriaceae	4318	18.7	3.47	20
Ordre	Enterobacteriales	*V. cholerae*	Vibrionales	3835	69.5	2.96	10
		P. aeruginosa	Pseudomonadales	5567	85.6	2.94	12
Classe	Gammaproteobacteria	*M. loti*	alphaproteobacteria	6746	110.8	2.66	9
Phylum	Proteobacteria	*B. subtilis*	Firmicutes	4112	112.8	2.44	9
		M. tuberculosis	Actinobacteria	3995	129.9	2.47	6
		C. tepidum	Bacteroidetes/Chlorobi	2252	117.8	2.70	9
		R. baltica	Planctomycetes	7325	127.1	2.39	
Domaine	Bacteria	*M. acetivorans*	Archaea (Euryarchaeota)	4540	139.8	2.19	3
		S. solfataricus	Archaea (Crenarchaeota)	2977	145.3	2.09	3

L'analyse de ces données de synténie nous a permis de démontrer que les gènes orthologues positionnels sont plus contraints (distance PAM plus petite) que les autres orthologues (Tableau 2). Une telle différence de pression de sélection renforce le concept de contexte génétique qui propose que des gènes restent voisins parce que leurs produits interagissent physiquement et/ou participent à un même processus cellulaire.

Tableau 2 : *Comparaison de la distance PAM moyenne et du nombre de paires d'orthologues présents dans les blocs de synténie ou en dehors de ces blocs (d'après Descorps-Declère et al., 2008).*

***Comparaison E. coli* contre**	**Distance PAM moyenne**		**Nombre de paires d'orthologues**	
	Bloc Synténie	**Hors bloc Synténie**	**% Bloc Synténie**	**% Hors Bloc Synténie**
S. enterica	11.93	23.53	27	73
B. subtilis	86.27	109.48	23	77
B. thetaiotaomicron	109.14	114.42	16	84
M. acetivorans	113.01	124.53	14	86

L'ensemble des données obtenues dans ce projet ont été intégrées dans une base de données SynteBase qui peut être interrogée avec l'outil SynteView (http://www.synteview.u-psud.fr). L(ensemble SynteBase/SynteView permet de comparer un grand nombre de génomes à la fois (Lemoine *et al.*, 2008). A ce jour SynteBase propose des données de synténie pour 598 organismes (550 bactéries et 48 archées). On notera également que le visualiseur Synteview est un outil flexible qui permet également de visualiser la synténie de données autres que celles contenues dans SynteBase à condition toutefois que les données soient structurées de façon analogue.

L'ensemble de ces travaux ont fait l'objet de trois publications dans les revues BMC Evolutionary Biology (Lemoine *et al.*, 2007), BMC Bioinformatics (Lemoine *et al.*, 2008) et Biochimie (Descorps-Declère *et al.*, 2008).

2.4.3. La découverte des activités enzymatiques orphelines

Dans tous les génomes, il existe des gènes spécifiques pour lesquels il n'est pas possible de trouver d'orthologues. Les quelques 1374 génomes actuellement publiés (source GOLD du 08/10/2010) (Liolios *et al.*, 2010) sont truffés de gènes orphelins codant des protéines hypothétiques de fonction inconnue. A l'inverse, comme nous allons le voir par la suite, il existe de nombreuses activités enzymatiques pour lesquelles il est extrêmement difficile de trouver dans les banques de données les séquences des gènes correspondants. C'est ce que nous avons appelé les activités enzymatiques orphelines.

Les différentes activités enzymatiques sont identifiées à l'aide d'un code attribué par le comité de nomenclature (Nomenclature Committee) de l'International Union of Biochemistry and Molecular Biology (NC-IUBMB). Ce code à 4 nombres séparés par des points est communément appelé EC number et sert à identifier et à classer les enzymes en fonction des réactions qu'elles catalysent. Le premier nombre représente la classe de l'enzyme. Il existe actuellement 6 classes numérotées de 1 à 6 qui correspondent respectivement aux activités de type : Oxidoréductases, Transférases, Hydrolases, Lyases, Isomérases et Ligases. Les trois nombres suivants correspondent à la sous classe, à la sous-sous classe et enfin à la spécificité de substrat. Ainsi, le EC number 1.1.1.1 correspond à une Oxidoréductase (classe) agissant sur un groupe CHOH qui sert de donneur (sous-classe) et dont l'accepteur est le NAD ou le NADP (sous-sous classe) et dont le substrat est un alcool. Cette activité correspond donc à l'alcool déhydrogénase. La réaction d'oxydation catalysée par cette enzyme étant la suivante : 1 alcool + NAD^+ → 1 aldéhyde ou 1 cétone + NADH + H^+.

Le NC-IUBMB reconnaît à ce jour (08/2010) 4256 activités enzymatiques différentes, qui sont toutes identifiées par un EC number différent. Les EC numbers sont donc précis et évitent les problèmes d'homonymie que l'on peut rencontrer avec les noms usuels des enzymes. Par exemple l'alcool déhydrogénase est, selon les auteurs, appelée alcool déhydrogénase, aldéhyde réductase, NADH-alcool déhydrogénase, *etc.*. L'usage des EC numbers est donc largement encouragé lorsque l'on réalise l'annotation fonctionnelle d'une protéine et cette information est largement répandue dans les bases de données de séquences protéiques.

Cependant, en 2004, Bernard Labedan m'a alerté sur la difficulté qu'il rencontrait à retrouver dans les banques de données la séquence de la putrescine carbamoyl transférase (EC 2.1.3.6). En effet, toutes les séquences qu'il obtenait en utilisant la requête « putrescine carbamoyl transférase » correspondaient en réalité à des séquences d'ornithine carbamoyl transférase (EC 2.1.3.3) qui avaient été incorrectement annotées. Or sur ce cas particulier Bernard savait qu'une partie de la séquence de ce gène avait été déterminée par Victor Stalon chez *Enterococcus faecalis*. En outre le génome complet de cet organisme avait été séquencé et publié en 2003 (Paulsen *et al.*, 2003). La séquence de la putrescine carbamoyl transférase aurait donc dû se trouver dans les banques de données.

J'ai donc entrepris une analyse systématique des principales banques de données protéiques (UniProtKB (UniProt Consortium, 2010), PDB (Berman HM *et al.*, 2003)) pour rechercher s'il existait d'autres activités enzymatiques pour lesquelles nous ne trouvions aucune séquence associée.

A notre grande surprise presque la moitié (42% en 2004) des EC numbers définis par le NC-IUBMB étaient orphelins de toute séquence. La même constatation était faite indépendamment aux Etats-Unis par Peter Karp (Karp, 2004). Le même type d'analyse fait en août 2010 montre que 1207 des 4256 EC numbers (28.6%) sont encore orphelins.

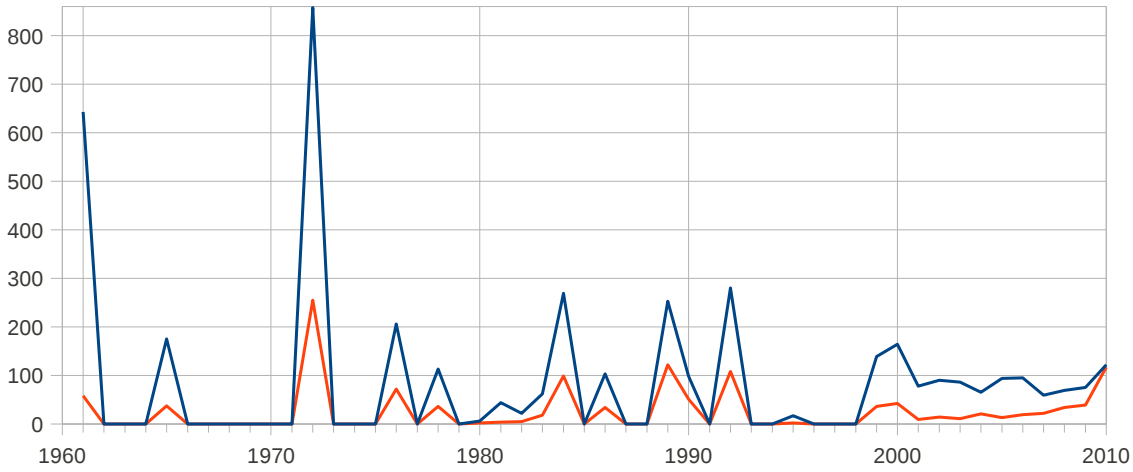


Figure 9 : *Distribution des EC numbers (bleu) et des EC numbers orphelins (rouge) en fonction de leur année de description. Les années sont indiquées sur l'axe des abscisses et les EC numbers sur l'axe des ordonnées. L'analyse a été réalisée en août 2010.*

Si les activités enzymatiques nouvellement décrites (4 dernières années) présentent en proportion plus d'orphelins que celles décrites dans les années 70, 80 ou 90 (Figure 9), on constate que même pour les années les plus anciennes un certain nombre d'activités enzymatiques sont toujours

orphelines de séquences. Ainsi, presque un demi siècle plus tard, 9% des activités enzymatiques décrites en 1961 restent orphelines (Figure 9). Ce qui est assez paradoxal à une époque où de nouveaux génomes sont continuellement séquencés et annotés.

Si toutes les classes enzymatiques présentent des activités orphelines, la classe des ligases est celle qui en présente le moins (21.2%) alors que celles des transférases en présente le plus (33.5%). On retrouve des activités enzymatiques orphelines aussi bien en ce qui concerne des enzymes du métabolisme que pour des enzymes qui n'interviennent pas dans le métabolisme (Lespinet et Labedan, 2006a). Si l'on s'intéresse au métabolisme, tel qu'il est défini dans la base de données KEGG (Kanehisa *et al.*, 2010), on constate là encore que tous les types de voies métaboliques possèdent des activités enzymatiques orphelines (Tableau 3). On remarque également qu'en 5 ans la fraction des activités enzymatiques orphelines impliquées dans le métabolisme est passé de 22.9 à 15.6%, indiquant que les voies métaboliques sont de mieux en mieux connues. A titre de comparaison la fraction totale d'activités enzymatiques orphelines est passée de 39.3% en août 2005 (Lespinet et Labedan, 2006a) à 28.6% en août 2010.

Tableau 3 *: Distributions des EC orphelins dans les principales voies métaboliques définies par KEGG (Kanehisa et al., 2010)). Les données en bleu datent de août 2005 et celles en rouge de août 2010.*

Type de voies métaboliques KEGG	**Nombre de EC numbers**	**Nombre d'orphelins**	**pourcentage d'orphelins**	**Nombre de voies complètes**
Cofactors and Vitamins	231 269	54 35	23.4 13.0	1 sur 11 4 sur 12
Carbohydrate	669 521	156 87	23.3 16.7	1 sur 17 2 sur 15
Amino acid	755 541	169 72	22.4 13.3	2 sur 25 4 sur 20
Lipid	241 270	51 39	21.2 14.4	2 sur 10 4 sur 15
Nucleotide	156 142	28 17	17.9 11.9	0 sur 2 0 sur 2
Glycan	153 125	26 9	17.0 7.2	6 sur 14 10 sur 15
Energy	156 161	21 13	13.4 8.1	4 sur 8 5 sur 8
Secondary metabolites	195 203	73 74	37.4 36.5	3 sur 14 4 sur 17
Biodegradation of xenobiotics	200 212	55 44	27.4 20.8	2 sur 18 9 sur 25
Polyketides and nonribosomal peptides	13	2	15.4	3 sur 5
Terpenoids and Polyketides	98	18	18.7	8 sur 17
Biosynthetic pathways	299	36	12,0	0 sur 7
Total	**2769** **2841**	**635** **444**	**22.9** **15.6**	**24 sur 124** **50 sur 153**

Plusieurs hypothèses sont envisagées pour expliquer un nombre aussi important d'activités

enzymatiques orphelines. La première serait que les EC numbers ne sont pas assez utilisés par les annotateurs des génomes et les curateurs des banques de données. Cela était certainement vrai en 2004 mais à mon avis ca n'est plus le cas aujourd'hui, ce qui doit expliquer en grande partie la baisse de la fraction de ces activités orphelines. Une autre hypothèse serait que les activités enzymatiques auraient été décrites chez des espèces « exotiques » pour lesquelles on dispose de très peu de gènes séquencés. Il est effectivement à noter que 42% des activités enzymatiques n'ont été observées que chez une seule espèce. A l'inverse certaines activités enzymatiques ont été décrites chez un grand nombre d'organismes et restent malgré cela toujours orphelines, comme par exemple la « Vanillin Synthase » (EC 4.1.2.41) qui est décrite chez neuf espèces dont, entre autre, *Bacillus subtilis* et *Escherichia coli*. Si l'on s'intéresse aux espèces pour lesquelles le *corpus* de données est très abondant, on observe malgré tout de l'ordre de 5 à 10% des activités enzymatiques qui demeurent orphelines, comme par exemple chez *Escherichia coli*, *Homo sapiens* ou encore *Saccharomyces cerevisiae*.

Afin d'inciter la communauté à participer à l'identification de toutes les activités enzymatiques orphelines, j'ai construit ORENZA (ORphan ENZyme Activities : http://www.orenza.u-psud.fr) une base de données ou l'on retrouve la liste de toutes les activités enzymatiques orphelines. Il s'agit en fait d'un entrepôt où sont stockées et périodiquement mise à jour des données provenant des bases UniProtKB (Uniprot Consortium, 2010), IntEnz (Fleischmann *et al.*, 2004), PDB (Berman *et al.*, 2003), ENZYME (Bairoch, 2000), BRENDA (Chang *et al.*, 2009), KEGG (Kanehisa *et al.*, 2010) ainsi que la taxonomie du NCBI (Sayers *et al.*, 2009). Grâce à un ensemble de scripts, ces données sont analysées, croisées et présentées via une série de pages Web. La base de données ORENZA est actuellement la base de données de référence dans le domaine des activités enzymatiques orphelines.

L'ensemble de ces travaux a donné lieu à plusieurs publications dans les revues : Science (Lespinet et Labedan, 2005), Cellular and Molecular Life Sciences (Lespinet et Labedan, 2006a), Drug Discovery Today (Lespinet et Labedan, 2006b) et BMC Bioinformatics (Lespinet et Labedan, 2006c).

2.4.4. Annotation structurale et fonctionnelle du génome du champignon filamenteux Podospora anserina

Podospora anserina est un champignon ascomycète qui pousse sur le crottin des herbivores. Depuis les années 40 (Rizet G, 1941), c'est également un système modèle en génétique pour étudier notamment la méiose, le vieillissement, les prions ou bien encore la reproduction sexuée des champignons.

Les 35,7 Mb du génome nucléaire de la souche S *mat+* de *Podospora anserina* ont été séquencées en 2002 par le Genoscope avec un niveau de couverture de 10x. L'assemblage des séquences a été réalisée par le Genoscope à l'aide du logiciel Arachne (Jaffe *et al.*, 2003). A la demande de Philippe Silar (Institut de Génétique et Microbiologie), je me suis intégré à ce projet en 2003 et en collaboration avec lui j'ai entrepris de réaliser l'annotation structurelle et fonctionnelle de ce génome.

L'annotation structurale a été réalisée en utilisant deux types d'approches. Une première approche par homologie a été employée en comparant, par tblastn (Altschul *et al.*, 1990), toutes les ORF (de plus de 20 acides aminés) de Podospora contre les génomes des ascomycètes *Neurospora crassa* (*Galagan et al.*, 2003) et *Chaetomium globosum* (données non publiées). La recherche des introns a

été réalisée en se basant sur le modèle d'introns développé par Philippe Silar dans un projet pilote publié en 2003 (Silar *et al.*, 2003). Les ORF présentant un hit avec Neurospora ou Chaetomium pour laquelle la e-value était inférieure à 10^{-18} ont été retenues comme des exons putatifs. Les exons distants de moins de 200 bp et qui possédaient des pieds d'introns compatibles ont été fusionnés pour créer des CDS. La recherche des exons 5' et 3' a été affinée en réduisant les contraintes sur la taille des ORF et sur la valeur de la e-value (10^{-5}).
Dans un second temps, les prédicteurs de gènes *ab initio* FGENESH (Softbury) et GeneID (Blanco *et al.*, 2007) ont également été utilisés afin d'identifier des gènes putatifs dans les régions où l'approche par homologie ne faisait aucune prédiction.
Les CDS prédites par homologie et par les prédicteurs *ab initio* ont été comparées aux 51759 EST séquencées par le Genoscope afin de confirmer ou d'infirmer leur existence.

Enfin une vérification manuelle de chaque CDS a été réalisée, en utilisant Artemis (Rutherford *et al.*, 2000), afin de confirmer sa prédiction et le cas échéant de l'éditer manuellement.

J'ai recherché les tRNA à l'aide de tRNAscan (Lowe et Eddy, 1997) et les ARN non codants ont été recherché par Daniel Gautheret à l'aide d'une combinaison des programmes Erpin (Gautheret et Lambert, 2001), Blast (Altschul *et al.*, 1990) et Yass (Noé et Kucherov, 2005).

L'ensemble de cette phase d'annotation structurale a permis de prédire 10545 CDS, 361 tRNA, 75 unités de rDNA, 87 5S rRNA, 14 snRNA et 13 snoRNA (Tableau 4). Si l'on compare le nombre de CDS avec celui de *Neurospora crassa*, qui est l'espèce la plus proche, la capacité codante de Podospora semble environ être de 10% supérieure. Il faudra à l'avenir déterminer si cette différence est réelle ou seulement due à une différence de stratégie d'annotation. A titre d'information, la version 6.28 du génome de Podospora (13/07/10) contenait 10640 CDS soit seulement 95 CDS de plus que ce que nous avions initialement prédit.

Tableau 4 : *Principales caractéristiques du Génome de Podospora anserina.*

Génome nucléaire	
Taille	35.5-36 Mb
Nombre de Chromosomes	7
Pourcentage de GC (génome total)	52.02%
Pourcentage de GC (séquences codantes)	55.87%
Pourcentage de GC (régions non codantes)	48.82%
Nombre de tRNA	361
Nombre d'unités de rDNA	75
Taille de l'unité rDNA consensus	8192 pb
Nombre de 5S rRNAs	87
Nombre de snRNA	14
Nombre de snoRNA	13
Nombre de Protein (CDS)	10545
Fraction codante du génome	44.75%
Taille moyenne des CDS (min; max)	496.4 codons (10; 8,070)
Nombre moyen d'introns/CDS (max)	1.27 (14)
Taille moyenne des introns (max)	79.32 nucléotides (2,503 nucléotides)
Fraction du génome constituée d'éléments transposables	3.5%
Génome Mitochondrial	
Taille	94,197 bp
Nombre de chromosomes	1 chromosome circulaire
Pourcentage de GC	30%

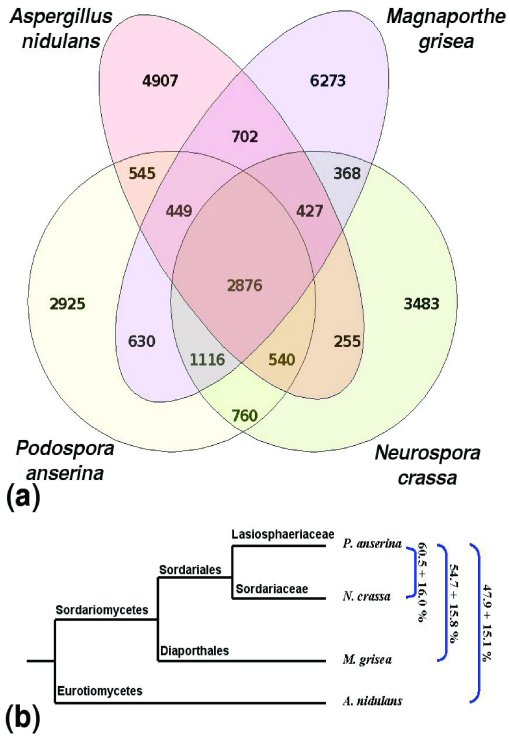


Figure 10 : *(a) Diagramme de Venn de 4 ascomycètes. (b) Pourcentage d'identité ± écart type entre les orthologues des 4 ascomycètes comparés en (a) (d'après Espagne et al., 2008).*

L'annotation fonctionnelle des CDS a été réalisée en comparant par blastp (Altschul *et al.*, 1990) les séquences protéiques contre les banques UniProtKB (Uniprot Consortium, 2010) et nr du NCBI. InterProScan (Mulder et Apweiler, 2007) a également été utilisé pour prédire les motifs fonctionnels. L'ensemble de ces données ont été intégrées dans une banque de données relationnelle consultable par les différents annotateurs (une dizaine environ) qui se sont répartis les 10545 CDS prédites initialement. L'ensemble du travail des annotateurs a été supervisé par un super annotateur afin de disposer à la fin d'une annotation qui soit la plus homogène possible.

L'analyse des catégories fonctionnelles a permis de mettre en évidence que Podospora possédait un large répertoire d'enzymes spécialisées dans l'utilisation des différentes sources de carbone (Espagne *et al.*, 2008).

Des groupes d'orthologues entre *Podospora anserina*, *Neurospora crassa*, *Magnaporthe grisae* et *Aspergillus nidulans* ont été réalisés par la méthode des BRH (e-value inférieure à 10^{-3}) afin de construire un diagramme de Venn (Figure 10) et de comparer le pourcentage d'identité entre les orthologues. Sans surprise ces résultats confirment que Podospora et Neurospora sont les deux espèces les plus proches.

En ce qui concerne la dynamique de son génome, Podospora possède de nombreux éléments transposables, quelques duplications segmentales et des gains et pertes de gènes ont été identifiés (Espagne *et al.*, 2008).

L'ensemble de ces résultats a fait l'objet d'une publication dans Genome Biology (Espagne *et al.*, 2008). Les outils créés dans le cadre de ce projet ont permis de développer par la suite des approches à haut débit (transcriptome) et facilitent l'analyse génétique de cet organisme.

2.4.5. Evolution du métabolisme des champignons : FUNGIpath.

Les champignons se caractérisent par un métabolisme secondaire riche et adapté à leur mode de vie. En conséquence, on trouve chez ces organismes une très grande variété d'enzymes aux nombreuses applications potentielles (Peláez, 2005). Le projet FUNGIpath vise à étudier l'ensemble de la diversité métabolique des champignons afin de comprendre comment elle a pu prendre place et se maintenir au cours de l'évolution.

Nous avons mené l'étude de la diversité métabolique des champignons en nous basant sur la capacité métabolique codante d'un certain nombre d'espèces pour lesquelles le génome avait été complètement séquencé. Avec plus de 100 génomes déjà disponibles (Cuomo et Birren, 2010), les champignons constituent en effet, après les animaux, le second groupe eucaryotes pour lequel le plus grand nombre de génomes est séquencé. Il est à noter que les champignons devrait, dans les années à venir, devenir le groupe eucaryote qui possède le plus d'espèces séquencées (Figure 11).

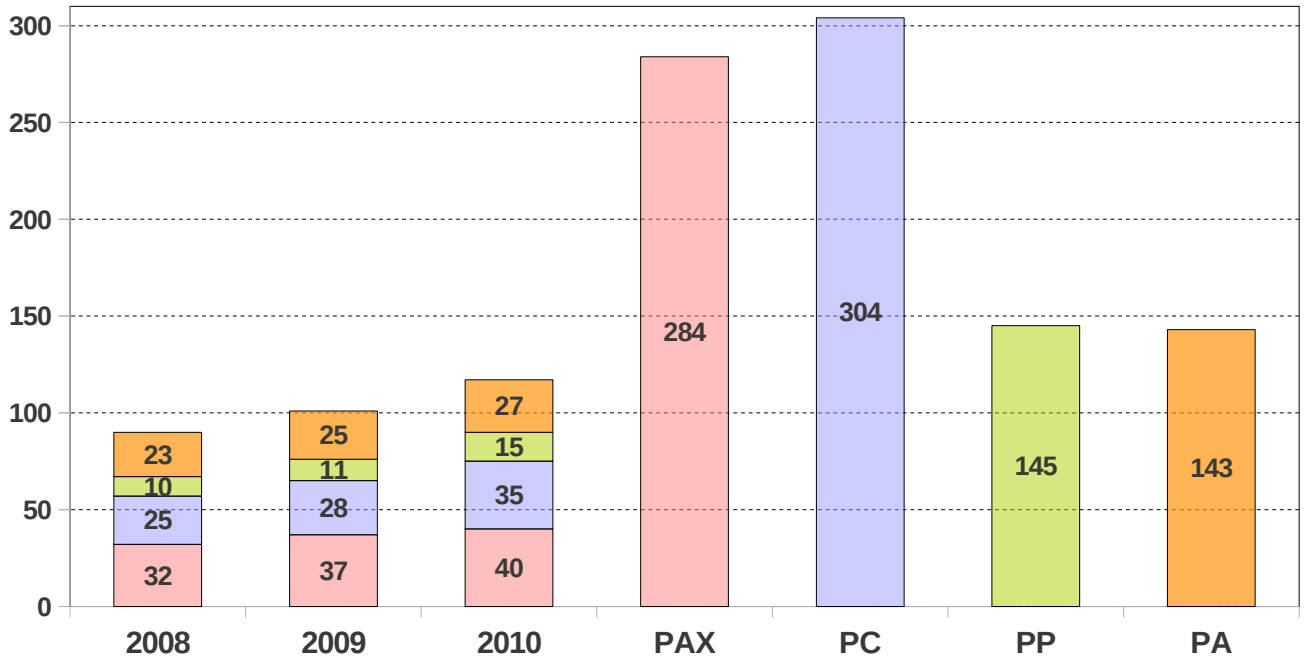


Figure 11 : *Nombre de génomes eucaryotes publiés. Bilan au 1er janvier des années 2008, 2009 et 2010. Les projets en cours sont indiqués dans les 4 colonnes les plus à droite. PAX : projets de génomes animaux (rose), PC : projets de génomes de champignons (violet), PP : Projets de génomes de plantes (vert), PA : projets d'autres génomes eucaryotes (orange) (d'après Liolios et al., 2009).*

On peut définir la capacité métabolique codante comme étant le nombre de CDS qui dans un génome codent pour une activité enzymatique. Dans la suite de notre propos, nous utiliserons le terme ID-EC comme étant la relation entre une protéine ou plus exactement son identifiant dans le génome ou dans les bases de données et son activité enzymatique décrite sous forme de EC number. Le nombre total d'ID-EC reflète la capacité métabolique codante d'un organisme.
Si l'on s'intéresse aux capacités métaboliques codantes des champignons, relativement peu de données sont disponibles dans les principales banques de données où l'on retrouve ce type d'informations comme par exemple KEGG (Kanehisa *et al.*, 2010), MetaCyc (Caspi *et al.*, 2010) ou Swiss-Prot (UniProt consortium, 2010). Plus exactement, si le nombre d'espèces pour lesquelles des données sont disponibles peut sembler important, un faible nombre d'informations de type ID-EC sont disponibles pour chacune de ces espèces, soulignant ainsi le caractère lacunaire des

informations contenues dans ces bases de données (Tableau 5). Ainsi, avec 344 espèces de champignons mais seulement 2 ID-EC par espèce en valeur médiane, on constate qu'en se basant sur les données de Swiss-Prot, il sera difficile de réaliser une comparaison du métabolisme des champignons. La situation est plus ou moins la même en ce qui concerne MetaCyc à la différence que le nombre de champignons présents dans cette base est très faible (Tableau 5).

Tableau 5 : Distribution des informations métaboliques dans les principales bases de données du métabolisme. D'aprèsc des données issues de KEGG version55.0 (Kanehisa et al., 2010), MetaCyc version 13.6 (Caspi et al., 2010) et Swiss-Prot version 15.14 (UniProt Consortium, 2010).

	Nombre d'espèces avec au moins 1 ID-EC			Nombre médian d'ID-EC par espèce		
	KEGG	MetaCyc	Swiss-Prot	KEGG	MetaCyc	Swiss-Prot
Animaux	47	2	1290	2021	2	1
Champignons	43	34	344	855	3	2
Plantes	10	133	928	1051	2	1

La base de données KEGG est celle pour laquelle on trouve le plus d'informations avec 43 espèces présentes et une médiane de 855 ID-EC par espèce. Il pourrait donc apparaître tentant d'utiliser les données de KEGG comme cela a été fait pour d'autres organismes (Wylie *et al.*, 2008; Whitaker *et al.*, 2009). En réalité la situation est beaucoup moins idéale avec certaines espèces qui possèdent jusqu'à 1464 ID-EC alors que d'autres n'en ont qu'un peu plus de 200.

Devant un telle hétérogénéité, nous avons entrepris d'annoter tous les génomes sur lesquels nous voulions faire porter notre étude. Nous avons pour cela utilisé une stratégie en 3 étapes. La première étape a consisté à construire des groupes d'orthologues entre toutes les espèces qui nous intéressent. Dans un second temps nous avons réalisé une annotation enzymatique automatique de bonne qualité des groupes d'orthologues générés lors de la première étape. La dernière étape a consisté à reporter sur des cartes métaboliques de référence les informations obtenues à l'issue des étapes précédentes.

Comme nous l'avons vu dans la partie traitant de l'étude de la synténie chez les microorganismes, la construction de groupes d'orthologues est en pratique un processus difficile à réaliser. Afin d'essayer d'y voir un peu plus clair entre les différentes méthodes proposées pour construire des groupes d'orthologues, nous en avons essayé et comparé 4 différentes, qui ont été retenues en raison de leur efficacité supposée et de la facilité de leur implémentation. Il s'agit des méthodes BRH (Overbeek *et al.*, 1999), Inparanoid (Remm *et al.*, 2001), OrthoMCL (Li *et al.*, 2003), ainsi que la méthode d'analyse d'arbres phylogénétiques décrite précédemment. Ces 4 méthodes ont permis d'analyser le même jeux de données constitué dans un premier temps d'un ensemble test composé de 20 génomes (Grossetête *et al.*, 2010). Par la suite notre approche a été étendue à 30 puis à 50 génomes.

Comme attendus, les résultats obtenus sont assez différents d'une méthode à l'autre (Tableau 6). On observe que le pourcentage de groupes trouvés à l'identique entre les méthodes est relativement faible avec une moyenne de 9% et un maximum de 22,4% entre les méthodes Inparanoid et OrthoMCL. *A contrario* le pourcentage de groupes trouvés spécifiquement par l'une des deux méthodes est assez élevé avec une moyenne de 21%. On constatera que de ce point de vue c'est la méthode de Phylogénie qui se montre la plus atypique.

Tableau 6 : *Comparaison des groupes d'orthologues prédits par 4 méthodes différentes pour 20 génomes de champignons. La partie haute (rouge) indique le pourcentage de groupes d'orthologues trouvés à l'identique entre les méthodes. La partie basse (bleu) indique le pourcentage de groupe spécifique de chacune des méthodes (d'après Grossetête at al., 2010).*

Nombre de groupes obtenus	Comparaison des méthodes en pourcentages de groupes d'orthologues				
		BRH	Inparanoid	OrthoMCL	Phylogénie
52292	BRH	-	4.8	3.7	5.8
18235	Inparanoid	8	-	22.4	8.5
20379	OrthoMCL	12.4	16.3	-	8.3
12676	Phylogénie	32.4	25.9	32.6	-

Afin d'essayer de tirer partie de ces résultats et d'essayer de produire un résultat robuste nous avons choisi de proposer une méta-méthode de prédiction qui combine les résultats provenant des 4 méthodes précédentes. Pour se faire, pour chaque groupe d'orthologues, nous avons dans un premier temps conservé uniquement les séquences communes à l'ensemble des méthodes. Les séquences conservées ont constitué ce que nous avons appelé les graines des groupes d'orthologues. Des profils HMM de ces graines ont été établis à l'aide du logiciel hmmer (Eddy, 1998). Les séquences qui se sont trouvées exclues des graines, soit parce qu'elles n'étaient trouvées que par un sous ensemble des méthodes, soit par aucune, ont été comparées une à une, à l'aide du logiciel hmmer (Eddy, 1998), contre une base de données contenant tous les profils HMM construits précédemment. Celles pour lesquelles le résultat présentait une e-value inférieure ou égale à 10^{-10} ont été ajoutées à la graine de départ, les autres ont été définitivement écartées. Cette seconde phase dite d'enrichissement a permis d'augmenter la taille des groupes d'orthologues tout en conservant une relativement bonne homogénéité au sein du groupe.

Tableau 7 : *Comparaison des groupes d'orthologues prédits par FUNGIpath avec ceux prédits par les méthodes BRH, Inparanoid, OrthoMCL et Phylogénie (d'après Grossetête et al., 2010).*

	BRH	Inparanoid	OrthoMCL	Phylogénie	**Moyenne**
Pourcentage de groupes identiques à FUNGIpath	2.8	18.6	18.8	10.7	**12.7**
Pourcentage de groupes spécifiques à FUNGIpath	10.6	10.6	23.4	24.6	**17.3**

Si l'on compare les groupes d'orthologues prédits par FUNGIpath avec ceux prédits par les 4 méthodes initiales, on constate que pour chaque méthode en moyenne 12.7% des groupes sont identiques et que 17.3% d'entre eux sont spécifiques (Tableau 7).
Pour la version à 50 génomes, FUNGIpath contient des groupes d'orthologues dont la taille varie de 2 à 446 séquences. 20 % des groupes ne contiennent que 2 séquences. Selon les espèces, de 49% à 95% des gènes sont attribués à un groupe d'orthologues (Grossetête, 2010).

L'annotation fonctionnelle des différents groupes d'orthologues a été réalisée en développant une approche originale consistant à comparer le profil HMM produit à partir de chaque groupe d'orthologues contre la totalité des séquences des banques Swiss-Prot (Uniprot Consortium, 2010) et MetaCyc (Caspi *et al.*, 2010) possédant une fonction enzymatique connue, c'est-à-dire associée à au moins un EC number complet. Les banques de données Swiss-Prot et MetaCyc contiennent des

informations qui ont été vérifiées et sont donc à priori de meilleure qualité que celles présentes dans d'autres banques de données où l'analyse est pour l'essentiel automatique comme c'est le cas par exemple de KEGG (Kanehisa *et al.*, 2010).
La procédure d'annotation des groupes d'orthologues nous garantit une annotation homogène utilisant le même protocole pour chacun des génomes. Ce qui, je le pense, est essentiel lorsque l'on veut utiliser une approche de génomique comparée. L'application de cette procédure a permis d'annoter 10% des groupes d'orthologues avec une activité enzymatique (2610 sur 26708). Ce qui est significativement plus que ce que l'on aurait pu prédire si nous nous étions contentés de transférer à l'ensemble du groupe les annotations qui pouvaient exister pour l'un ou plusieurs des gènes du groupe. Dans ce cas, nous aurions annoté seulement 1193 groupes d'orthologues (Grossetête, 2010).

Avant de poursuivre et de nous lancer dans la comparaison des capacités codantes des différents champignons sur lesquels nous avons fait porter notre étude, nous avons voulu essayer d'estimer la qualité de nos annotations. Pour se faire, nous les avons comparées aux annotations prédites pour *Saccharomyces cerevisiae* dans les banques KEGG, MetaCyc, Swiss-Prot ainsi que dans la banque spécialisée SGD (Christie *et al.*, 2004). Nous avons choisi *Saccharomyces cerevisiae* car c'est à priori le champignon pour lequel nous disposons du plus grand nombre d'informations dans les différentes banques de données (Tableau 8).

Tableau 8 : *Comparaison des prédictions ID-EC de Saccharomyces cerevisiae entre les bases de données KEGG, MetaCyc, SGD, Swiss-Prot et FUNGIpath (d'après Grossetête et al., 2010).*

Base de données	Nombre total de ID-EC	Nombre d'ID-EC annotés dans FUNGIpath	Parmi les ID-EC annotés dans FUNGIpath		Nombre et position des différences dans le code EC			
			identiques	différents	1ère	2ième	3ième	4ième
KEGG	1101	878	844 (96.1%)	34 (3.9%)	1	1	5	27
MetaCyc	155	142	134 (94.4%)	8 (5.6%)	2	0	1	5
SGD	527	451	419 (92.9%)	32 (7.1%)	5	2	3	22
Swiss-Prot	1261	1051	1024 (97.4%)	27 (2.6%)	1	0	3	23
FUNGIpath	**1261**	-	-	-	-	-	-	-

On observe que le nombre de prédictions ID-EC, faites par FUNGIpath, est supérieur à ce qui est fait par les autres bases de données à l'exception de Swiss-Prot pour laquelle le nombre de prédictions est aussi élevé que pour FUNGIpath (Tableau 8). Dans l'ensemble 83 à 97% des prédictions faites par FUNGIpath sont identiques à celles faites par les autres bases de données. En outre les différences, lorsqu'il y en a, portent plutôt sur le quatrième nombre du code EC, c'est-à-dire uniquement sur la spécificité de substrat. L'ensemble de ses résultats tend à penser que la qualité des prédictions faites par FUNGIpath est non seulement bonne, mais de qualité comparable à ce qui est fait par les autres bases de données. Deux interprétations possibles peuvent être faites en ce qui concerne le nombre plus élevé de prédictions faites par FUNGIpath. Soit FUNGIpath aurait tendance à prédire des annotations pour un trop grand nombre de gènes. Soit les prédictions faites par les autres bases de données sont lacunaires et FUNGIpath fait un nombre correct de prédictions. Afin d'essayer de trancher entre ces deux hypothèses nous avons comparé les prédictions faites par FUNGIpath avec celles faites par KEGG pour les douze espèces de champignons présents dans les deux bases de données (Tableau 9). Si nous avons choisi de ne réaliser la comparaison qu'avec KEGG c'est parce qu'après FUNGIpath c'est la base de données qui possède le plus d'informations

de type ID-EC (Tableau 5).

Tableau 9 : *Comparaison des prédictions ID-EC des banques KEGG et FUNGIpath pour 12 génomes de champignons (d'après Grossetête et al., 2010).*

Génome	Nombre de ID-EC		Nombre de			
	KEGG	FUNGIpath	ID-EC identiques	Même ID avec un EC différent	ID-EC spécifiques de KEGG	ID-EC spécifiques de FUNGIpath
Aspergillus nidulans	967	1890	675 (31%)	30 (1%)	262 (12%)	1185 (55%)
Aspergillus oryzae	1142	2148	853 (36%)	45 (2%)	244 (10%)	1250 (52%)
Fusarium graminearum	725	1786	535 (27%)	26 (1%)	164 (1%)	1225 (63%)
Laccaria bicolor	684	1536	472 (27%)	31 (2%)	181 (11%)	1033 (60%)
Magnaporthe grisea	1070	1801	749 (36%)	39 (2%)	282 (14%)	1013 (49%)
Neurospora crassa	852	1407	658 (42%)	26 (2%)	168 (11%)	723 (46%)
Podospora anserina	665	1594	473 (27%)	18 (1%)	174 (10%)	1103 (62%)
Saccharomyces cerevisiae	1101	1261	844 (57%)	35 (2%)	222 (15%)	382 (26%)
Schizosaccharomyces pombe	1009	1073	752 (58%)	33 (3%)	224 (17%)	288 (22%)
Sclerotinia sclerotiorum	651	1601	493 (28%)	16 (1%)	142 (8%)	1092 (63%)
Ustilago maydis	772	1206	546 (39%)	35 (3%)	191 (3%)	625 (45%)
Yarrowia lipolytica	909	1311	710 (48%)	27 (2%)	172 (12%)	574 (39%)
Moyenne	879	1551	647 (38%)	30 (2%)	202 (12%)	874 (48%)

L'analyse de ces résultats (Tableau 9) nous indique que si le nombre d'ID-EC est toujours plus élevé pour FUNGIpath, les résultats semblent néanmoins plus cohérents que ceux prédits par KEGG. En effet, si pour *Saccharomyces cerevisiae* et *Schizosaccharomyces pombe* les deux outils donnent des prédictions du même ordre de grandeur ce n'est plus le cas pour *Neurospora crassa* et *Podospora anserina*. KEGG prédit plus de relations ID-EC pour *Neurospora crassa* que pour *Podospora anserina*, à l'inverse de FUNGIpath. Bien que cela ne soit pas impossible, il serait toutefois surprenant que le nombre d'ID-EC soit supérieur pour *Neurospora crassa* alors que le nombre de CDS est sensiblement inférieur, chez cette espèce, au nombre de CDS prédites pour *Podospora anserina*. Ayant eu une grande responsabilité dans l'annotation du génome de Podospora, j'ai en outre plus tendance à croire en mes prédictions pour cet organisme qu'en celles faites par nos collègues de KEGG, surtout lorsqu'elles divergent autant. Enfin, si l'on compare la capacité codante globale (nombre de CDS prédites) des différents génomes, il est étonnant de constater un nombre très inférieur de prédictions faites par KEGG pour *Sclerotinia sclerotiorum* alors que le génome de cet organisme contient moitié plus de CDS que celui de Podospora. L'ensemble de ces résultats laisse donc penser que les données de KEGG sont lacunaires et rien ne semble prouver que celles de FUNGIpath sont surnuméraires.

Afin de pouvoir comparer les voies métaboliques des différents champignons, les annotations métaboliques produites par FUNGIpath ont été reportées sur des cartes métaboliques de référence. Ces cartes appartiennent à deux systèmes différents que sont MetaCyc et KEGG. Les voies de MetaCyc sont très spécialisées, petites (valeur médiane de 3 EC numbers par voie) et nombreuses alors que celles de KEGG sont des voies générales de plus grande taille (valeur médiane de 18 EC numbers par voie) où sont réunies les informations provenant de plusieurs espèces.

L'ensemble de ces données ont été regroupées au sein d'une base de données relationnelle interrogeable *via* une interface web dénommée FUNGIpath (http://www.fungipath.u-psud.fr). FUNGIpath permet de naviguer de façon intuitive dans les différentes cartes métaboliques de champignons et d'en tirer des informations, qui nous l'espérons, seront pertinentes pour les

biologistes.

Au travers d'un exemple, je vais essayer d'illustrer l'usage que l'on peut faire de cet outil.

Le galactose peut être utilisé de différentes façons (Figure 12). La voie de dégradation la plus répandue est la voie de Leloir (Frey, 1996). Elle permet de dégrader le galactose en α-D-glucose-1-phosphate ou en UDP-galactose. Cette voie constituée de 5 enzymes (2.7.1.6, 2.7.7.12, 2.7.7.9, 2.7.7.10 et 5.1.3.2) est plutôt bien conservée chez les champignons à l'exception de l'EC numbers 2.7.7.10 qui ne semble pas être présente (Figure 13). On remarque également que les activités enzymatiques 2.7.1.6 et 2.7.7.12 semblent absentes chez *A. gossypii*, *C. glabrata*, *C. cinereus*, *A. macrogynus et E. cuniculi* (Figure 13). Ces 5 champignons ne devraient donc pas dégrader le galactose par la voie de Leloir car il semble qu'il leur manque les enzymes correspondant aux 2 premières étapes de la voie.

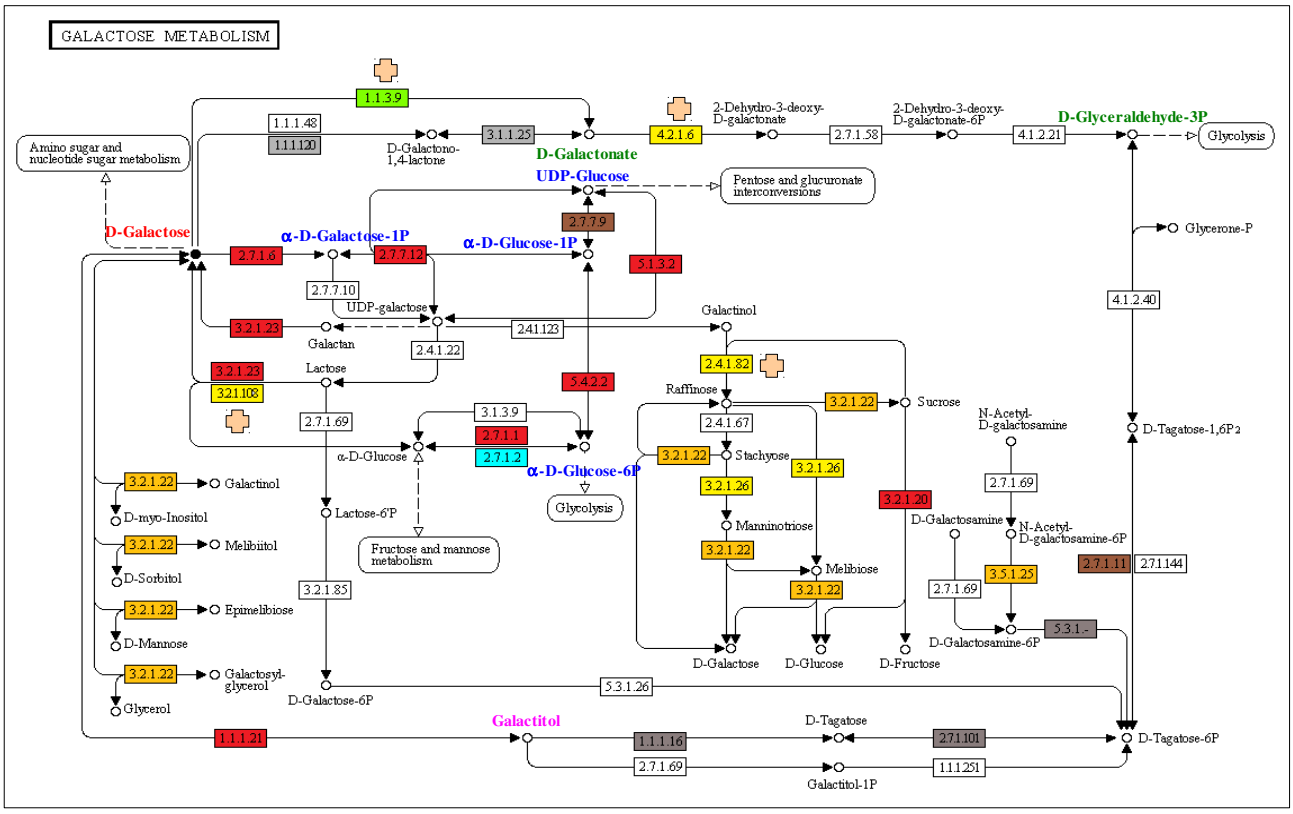


Figure 12 : *Conservation du métabolisme du galactose. Les EC numbers ont été colorés en fonction de leur niveau de conservation : blanc (0%), bleu (entre 0 et 20%), vert (entre 20 et 40%), jaune (entre 40 et 60%), orange (entre 60 et 80%), rouge (entre 80 et 100%), marron (100%). Les EC numbers en gris sont incomplets ou absent des banques Swiss-Prot et MetaCyc. Les EC numbers spécifiques de cette voie sont identifiés par des croix oranges (d'après Grossetête, 2010).*

Une étude récente réalisée chez 80 *Eumycota* (Slot *et al.*, 2010) a montré que la grande majorité des espèces étudiées possédaient les enzymes de la voie de Leloir. Cependant, ces auteurs ont observé que la voie avait été perdue chez plusieurs espèces. Les prédictions faites par FUNGIpath sont tout à fait en accord avec leurs résultats (Slot *et al.*, 2010).

La voie de Leloir n'est pas la seule façon d'utiliser le galactose. Ce dernier peut aussi être dégradé en D-glycéraldéhyde-3-phosphate *via* le galactonate (Figure 12). Cette voie alternative comporte 7 activités enzymatiques (1.1.3.9, 1.1.1.48, 1.1.1.120, 3.1.1.25, 4.2.1.6, 2.7.1.58 et 4.1.2.21) et a été mise en évidence expérimentalement chez *A. niger* (Elshafei *et al*., 2001). Trois des sept activités enzymatiques de cette voie ne semblent pas présente chez les champignons étudiés ici (1.1.1.48, 2.7.1.58 et 4.1.2.21). Deux autres actvités enzymatiques (1.1.1.120 et 3.1.1.25) n'étant associées à aucune séquence dans Swiss-Prot et MetaCyc, il est donc difficile de conclure sur leur présence chez les champignons. En ce qui concerne l'activité 3.1.1.25 une séquence d'*A. niger*, n'est présente que dans TrEMBL. Enfin les deux dernières activités enzymatiques (1.1.3.9 et 4.2.1.6) sont retrouvées chez certains champignons mais pas chez *A. gossypii*, *C. glabrata*, *C. cinereus*, *A. macrogynus et E. cuniculi*.

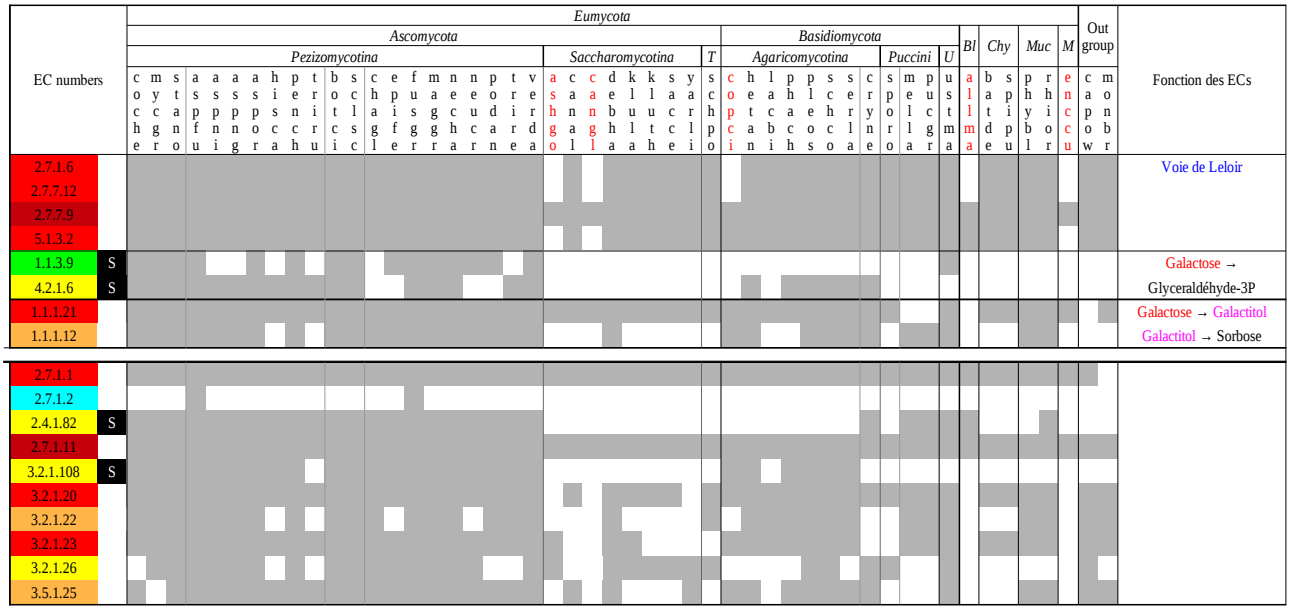


Figure 13 : Profil phylogénétique de la conservation du métabolisme du galactose. Les EC numbers ont été colorés en fonction de leur conservation (même code que la figure 12). Les EC numbers suivis d'une case noire avec un S indiquent qu'il s'agit de EC numbers spécifiques à la voie. Les cases sont grises si l'espèce possède l'activité enzymatique, blanches sinon. La dernières colonnes précisent le rôle de l'EC number dans la voie. ashgo : Ashbya gossypii, cangl : Candida glabrata, copci : Coprinus cinereus, allma : Allomyces macrogynus, enccu : Encephalitozoon cuniculi (d'après Grossetête, 2010).

Une autre manière d'utiliser le galactose est de le dégrader en galactitol (grâce à l'oxydoréductase 1.1.1.21, Figure 12). Cet EC, non spécifique de la voie, est trouvé dans la plupart des espèces à l'exception de *M. larici-populina*, *P. graminis*, *A. macrogynus* et *C. owczarzaki*. Les enzymes utilisant le galactitol comme substrat n'ont pas été trouvées chez les champignons (Figure 12) on peut donc se demander si cette voie est utilisée par les champignons. Ce qui ne figure pas sur la voie KEGG, et qui a été démontré chez *A. nidulans*, c'est que le galactose pouvait être dégradé en sorbose via le galactitol (Fekete *et al.*, 2004). La première étape est catalysée par le même EC que précédemment et le galactitol est ensuite oxydé en sorbose par une autre oxydoréductase (EC 1.1.1.12). Cet EC est présent dans la plupart des *Pezizomycotina* (à l'exception de *H. capsulatum* et *T. rubrum*) et des *Basidiomycota* (exceptés *L. bicolor*, *P. chrysosporium* et *S. roseus*), dans une levure *Ascomycota* (*D. hansenii*) et dans les deux *Mucoromycotina*. Il est à noter que *C. cinereus*, qui ne pouvait réaliser la voie de Leloir, semble posséder cette voie alternative pour dégrader le

galactose.

En définitive, quatre *Eumycota* (*A. gossypii*, *C. glabrata*, *A. macrogynus* et *E. cuniculi*) paraissent incapables d'utiliser le galactose. Il est cependant possible qu'ils possèdent une voie spécifique non identifiée à ce jour. Des résultats expérimentaux réalisés chez *A. gossypii* et *C. glabrata* (Kurtzman et Fell, 1998), montrent cependant que ces deux levures sont comme nous le prédisons incapables d'assimiler le galactose

L'ensemble de ce travail a constitué le sujet de la thèse de Sandrine Grossetête que j'ai officieusement et concrètement encadrée sous la direction officielle de Bernard Labedan. Ce travail des 3 dernières années a pour l'instant donné lieu à une publication dans BMC Genomics (Grossetête *et al.*, 2010). L'analyse détaillée des résultats devrait, je l'espère, conduire à d'autres publications.

- Chapitre 3 -

Projet de Recherche

3.0. Principaux objectifs

Le projet que je propose de développer dans les années à venir concerne pour partie l'analyse de l'énorme quantité de données produites au cours du projet FUNGIpath. Cette analyse visera à mieux comprendre la dynamique et l'évolution d'un réseau biologique, le métabolisme. L'objectif sera *in fine* de comprendre comment le métabolisme riche et complexe des champignons a pu se mettre en place au fil du temps et se maintenir dans sa diversité. S'il est difficile de dire dans quelle proportion l'étude du métabolisme permettra d'expliquer l'adaptation fonctionnelle des champignons à leur niche écologique, je pense néanmoins qu'avec FUNGIpath nous disposons des données et des outils qui pourraient nous permettre de répondre à cette question importante.

Je souhaiterais également généraliser ce type d'étude à d'autres groupes d'organismes. A titre d'exemples, les bactéries et les archées constituent également des groupes taxonomiques pour lesquels l'étude comparée du métabolisme permettrait d'obtenir des informations intéressantes.

Enfin la comparaison pourra également être étendue à d'autres type de réseaux. Je pense par exemple aux réseaux de signalisation ou aux réseaux de régulations génétiques. L'étude de la conservation à grande échelle des gènes impliquées chez les eucaryotes dans les processus de biologie du développement pourraient notamment constituer un sujet de recherche qui je le pense serait particulièrement intéressant.

3.1. Poursuivre l'analyse des données obtenues dans le cadre du projet FUNGIpath

Le projet FUNGIpath a permis de produire 127 cartes métaboliques KEGG et 858 cartes MetaCyc regroupant chacune des informations pour 50 organismes (figure 12). En outre, chaque carte est accompagnée d'un profil phylogénétique qui, pour chaque type d'enzyme de la voie représenté par un EC number, indique la liste des espèces possédant ou pas cette activité enzymatique (figure 13). L'exemple de la conservation du métabolisme du galactose présenté au chapitre précédent montre à quel point les deux types d'informations (carte et profil) sont indispensables si l'on veut exploiter et interpréter pleinement les informations produites par FUNGIpath.

La principale difficulté à laquelle nous devons actuellement faire face pour analyser nos données réside dans la complexité et le caractère multidimensionnel du problème. En effet, pour chaque voie métabolique, il faut non seulement tenir compte de l'information de la présence ou de l'absence de chaque enzyme, mais pour être pertinente l'analyse doit aussi tenir compte des relations entre les enzymes. S 'agit-il par exemple d'activités enzymatiques agissant de concert ou bien agissent-elles de façon ordonnée ? Ont-elles des produits ou des substrats communs ? Les relations taxonomiques entre les espèces sont également importantes à prendre en compte. Les espèces présentant un même profil phylogénétique appartiennent-elles à un même groupe taxonomique ou bien s'agit-il de groupes apparentés ? Les informations concernant la biologie des organismes considérés doit également être au coeur de l'analyse. Les organismes partageant des profils similaires ont-ils des milieux et modes de vie communs ? Sont-il pathogènes ?

Il faut également avoir en tête que la complexité de l'analyse ne fera que croitre au fur et à mesure que de nouvelles espèces seront ajoutées à FUNGIpath.

Jusqu'à présent l'analyse des données a principalement été réalisée manuellement par Sandrine Grossetête (Grossetête, 2010). Afin de pouvoir espérer mener à bien l'analyse nous l'avons réduite aux 65 voies KEGG qui présentent un nombre de EC numbers supérieur à 10 et pour lesquelles au

moins 20% des EC numbers de la voie sont retrouvés chez au moins une même espèce de champignon (Figure 14).

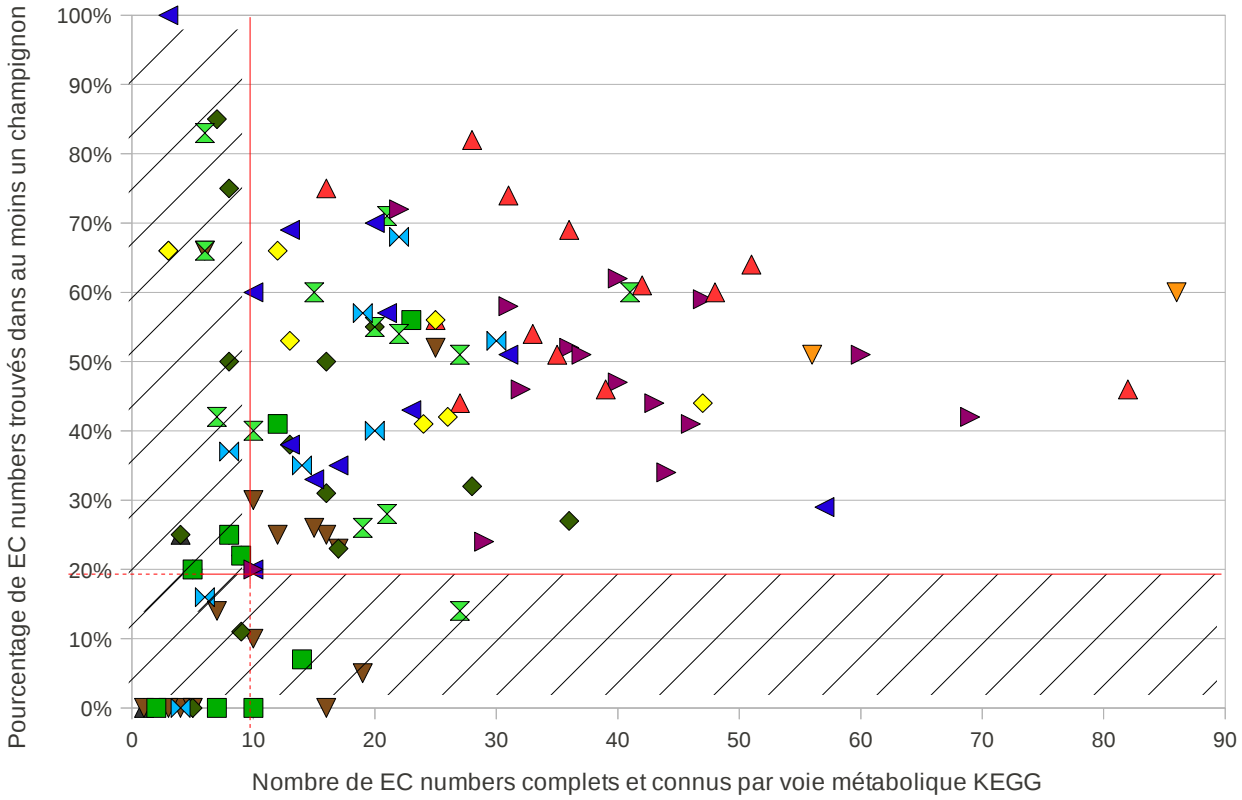


Figure 14 : *Pour chaque voie KEGG le pourcentage de EC numbers retrouvés chez au moins un champignon est indiqué. Les différentes classes de voies KEGG sont représentées par des symboles et des couleurs différents.*

L'analyse manuelle est non seulement difficile mais elle est répétitive, fastidieuse et consommatrice de temps. Nous envisageons donc de développer des stratégies d'analyse qui permettraient d'exploiter nos données de façon à réduire au minimum le temps humain consacré à l'analyse.

Pour essayer de réduite ce processus long et complexe nous prévoyons de faire porter conjointement nos efforts sur l'analyse des cartes métaboliques ainsi que sur l'analyse des profils phylogénétiques.

En ce qui concerne les cartes métaboliques, il s'agira de simplifier l'analyse en essayant d'extraire pour chaque carte les points les plus pertinents nécessitant une l'analyse plus détaillée. Parmi les points importants, l'analyse de la voie métabolique du galactose a permis de voir que par exemple la première étape de phosphorylation du galactose (EC 2.7.1.6) était cruciale pour que le reste de la voie puisse se poursuivre (Figure 12 et 13). Les activités enzymatiques 1.1.3.9, 4.2.1.6, 2.4.1.82 et 3.2.1.108 qui sont spécifiques de la voie du métabolisme du galactose pourraient également constituer des étapes importantes de cette voie.

L'analyse automatique des différentes voies ne sera possible que si chacune de ces voies est au préalable codée sous la forme d'un graphe orienté qui pourra ensuite être analysé en utilisant les algorithmes de la théorie des graphes. La recherche du plus long chemin conservé permettra par exemple de trouver la plus longue liste de EC numbers connectés de façon ordonné (étapes successives) présente pour une espèce ou un groupe taxonomique donné. La recherche des cliques

conservées (sous-ensemble des sommets tous connectés les uns aux autres) permettra de retrouver les modules taxonomiques ou fonctionnels constitutifs des différentes voies métaboliques. Ces modules sont des groupes de EC numbers qui présentent entre eux un degré (nombre d'arcs entrants et sortants) plus important qu'avec les EC numbers extérieurs au groupe. Il s'agit donc des EC les plus interconnectés (Kovacs *et al.*, 2010). La taille, la nature des EC les composant seront des éléments à prendre en compte lors de cette analyse. Les EC numbers spécifiques d'une voie donnée devront également être codés différemment car ils jouent certainement un rôle important dans le maintien de cette voie ou dans sa spécificité pour certains organismes. Pour chaque voie, la recherche du ou des noeuds présentant le plus haut degré sera également un point intéressant à regarder.

La détection de modules fonctionnels conservés et la détermination de modules fonctionnels spécifiques devraient nous permettre d'appréhender les mécanismes régulant la dynamique d'évolution des systèmes biologiques étudiés ici et de comprendre comment d'un système ancestral nous pouvons passer à une variété de systèmes dérivés spécifiques. Cette approche devrait également nous permettre de faire des propositions pour réaliser des modifications expérimentales ciblées pouvant changer le fonctionnement de certains de ces réseaux. Il va de soit que cette partie du projet ne pourra se faire que dans le cadre d'une collaboration avec une équipe de biologistes qui maîtrise l'aspect bio-technologique du projet.

L'analyse de graphes devrait nous permettre de répondre à un certain nombre de questions parmi lesquelles : Y a t-il un lien entre le nombre de voies dans laquelle une activité enzymatique est présente et son niveau de conservation ? Y a t-il un lien entre le nombre de fois où une activité enzymatique apparaît dans une même voie et son niveau de conservation ? Y a t-il un lien entre le degré de connexion d'une activité enzymatique et son niveau de conservation ?

Il est communément admis que les protéines agissant dans une même voie métabolique ou participant à un même complexe fonctionnel vont co-évoluer de la même façon et doivent donc présenter des profils phylogénétiques similaires (Pellegrini *et al.*, 1999; Vert, 2002). Les profils phylogénétiques sont des chaînes de caractères binaires constituées d'une succession de 0 (absence du caractère) et de 1 (présence du caractère). Un moyen de quantifier les différences entre profils consiste à compter le nombre de caractères différents entre les deux chaînes de caractères (Pellegrini *et al.*, 1999). Plus ce nombre sera faible et plus les profils seront à priori ressemblants. Cette approche naïve ne tient pas compte de l'information taxonomique présente dans les profils or cette information est essentielle si les questions posées sont de nature évolutive (Vert, 2002).

Dans notre analyse, les voies métaboliques (réseaux) et les profils phylogénétiques sont étroitement corrélés. Notre type d'analyse diffère cependant légèrement de ce qui se fait habituellement lorsque l'on utilise les profils phylogénétiques dans l'étude du métabolisme. En effet, la recherche des activités enzymatiques partageant le même profil permet en général de retrouver les modules constitutifs des voies métaboliques (Yamada *et al.*, 2006). Notre objectif n'est pas ici de reconstruire des voies *de novo* mais plutôt d'étudier quelle partie de la voie est conservée entre différents taxons. C'est donc plus le niveau de conservation que la définition des modules qui nous intéresse. Il n'en reste pas moins que l'analyse des profils phylogéniques est pour nous primordiale et constitue une part importante de notre analyse. Afin d'optimiser l'étude des profils phylogénétiques, j'aimerai développer une approche qui aille au delà de la recherche des profils les plus ressemblants. Par exemple la détection de l'absence d'une activité enzymatique pour une espèce ou un groupe d'espèces taxonomiquement reliés. Si un outil d'aide à l'analyse des profils phylogénétiques a récemment été conçu par Miklós Csurös (Csurös, 2010), à ma connaissance il ne permet pas d'automatiser l'analyse de nombreux profils. En outre je souhaiterai que l'analyse tienne compte du

Tableau 10 : *Profils phylogénétiques les plus fréquents. Chaque colonne représente un profil phylogénétique. Le nombre d'EC numbers présentant ce profil est indiqué sur la première ligne. Une case grise signifie qu'un génome est présent dans le profil et une case blanche qu'il est absent du profil. Seules les profils communs à plus de 4 EC numbers sont présentés ici (d'après Grossetête, 2010).*

Nombre d'ECs partageant le même profil					189	155	23	19	16	8	7			6		5			4			
Eumycota	*Ascomycota*	*Pezizomycotina*	*Dothideomyceta*	*C. heterostrophus*	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
				M. graminicola	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
				S. nodorum	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
			Eurotiomycetes	*A. fumigatus*	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
				A. nidulans	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
				A. niger	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
				A. oryzae	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
				H. capsulatum	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
				P. chrysogenum	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
				T. rubrum	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
			Leotiomycetes	*B. cinerea*	■	■	■		■	■	■	■	■	■	■	■			■	■	■	
				S. sclerotiorum	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
			Sordariomycetes	*C. globosum*	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
				E. festucae	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
				F. graminearum	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
				M. grisea	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
				N. haematococca	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
				N. crassa	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
				P. anserina	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
				T. reesei	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
				V. dahliae	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
		Saccharomyceta		*A. gossypii*	■	■	■		■	■	■	■		■	■	■	■		■	■	■	■
				C. albicans	■	■	■		■	■	■	■		■	■	■	■		■	■	■	■
				C. glabrata	■	■	■		■	■	■	■		■	■	■	■		■	■	■	■
				D. hansenii	■	■	■		■	■	■	■		■	■	■	■		■	■	■	■
				K. lactis	■	■	■		■	■	■	■		■	■	■	■		■	■	■	■
				K. thermotolerans	■	■	■		■	■	■	■		■	■	■	■		■	■	■	■
				S. cerevisiae	■	■	■		■	■	■	■		■	■	■	■		■	■	■	■
				Y. lipolytica	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	■
		Taphrinomycotina		*S. pombe*	■	■	■			■	■	■		■	■	■	■		■	■	■	
	Basidiomycota	*Agaricomycotina*	*Homobasidiomycetes*	*C. cinereus*	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
				H. annosum	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
				L. bicolor	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
				P. chrysosporium	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
				P. ostreatus	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
				S. commune	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
				S. lacrymans	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
			Tremel.	*C. neoformans*	■	■	■		■	■	■	■	■	■	■	■	■		■	■		
		Puccinomycotina	*Microb.*	*S. roseus*	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
			Pucciniomycetes	*M. larici-populina*	■	■	■		■	■	■	■	■	■	■	■	■		■	■		
				P. graminis	■	■	■		■	■	■	■	■	■	■	■	■		■			
		Ustilaginomycotina		*U. maydis*	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	
	Blastocladiomycota			*A. macrogynus*	■	■	■		■	■		■	■	■	■	■	■	■	■			
	Chytridiomycota			*B. dendrobatidis*	■	■	■		■	■	■		■	■		■		■	■			
				S. punctatus	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■		
	Mucoromycotina			*P. blakesleeanus*	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■		
				R. oryzae	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■		
	Microsporidia			*E. cuniculi*		■				■												
Ichthyospore				*C. owczarzaki*	■	■	■	■	■	■	■		■	■		■	■	■				
Choanoflagellida				*M. brevicollis*	■	■		■	■		■		■	■	■	■		■		■		

fait que les activités enzymatiques sont ou non liées directement au sein de la voie analysée.

En dehors des analyses détaillées réalisées pour chaque voie métabolique, il sera également intéressant d'effectuer une analyse globale des données. Nous pourrons par exemple déterminer le profil enzymatique complet de chaque espèce, c'est-à-dire établir la liste des activités enzymatiques présentes pour chacune des espèces étudiées. Il sera particulièrement intéressant de regarder si l'on peut corréler ces profils enzymatiques avec les spécificités biologiques de certaines espèces de champignons.

Une analyse préliminaire des profils phylogénétiques des différentes activités enzymatiques a montré que bien évidemment un certain nombre d'activités enzymatiques présentaient le même profil (Tableau 10). Il sera intéressant de regarder la répartition au sein des différentes voies des activités enzymatiques qui présentent un profil identique. Sont-elle retrouvées dans les mêmes voies ? Au sein d'une même voie, sont-elles retrouvées dans les même modules ? Sont-elles plus connectées entre elles ?

3.2. Comparer le métabolisme d'autres organismes.

Les champignons ne sont pas les seuls organismes à présenter un métabolisme riche et varié. Les bactéries et les archées constituent également des groupes taxonomiques sur lesquels une approche de type FUNGIpath pourrait être réalisée.

En ce qui concerne les bactéries, la tâche est assez ardue car le nombre de génomes à comparer est vraiment important (plus d'un millier). En outre des projets dont je subodore qu'ils sont du même type sont actuellement en cours de réalisation comme par exemple le projet Microme (http://www.microme.eu) qui est supporté par la communauté européenne. Il est cependant difficile, à la simple lecture de la page web de ce projet, de savoir s'il réalisera le même genre d'applications que ce que nous avons déjà réalisé pour les champignons.

En ce qui concerne les archées, à ma connaissance, aucun groupe ne prévoit de développer de projets similaires à FUNGIpath. Même si une analyse détaillée du métabolisme des archées halophiles a été publiée récemment (Falb *et al.*, 2008), on peut cependant espérer que l'analyse de plus de génomes incluant des organismes de différente nature devrait permettre de faire ressortir de nouvelles informations.

3.3. Comparer d'autres réseaux.

Les voies métaboliques constituent un type de réseaux biologiques mais il en existe d'autres qu'il pourrait être intéressant d'étudier avec le même type d'approche que celles utilisées dans FUNGIpath.

Les réseaux de régulation étudiés en biologie du développement sont particulièrement bien décrits chez certains organismes et comme nous l'avons vu dans le chapitre précédent, une partie de la discipline se consacre à vérifier la conservation de ces réseaux dans différentes espèces choisies pour leur intérêt taxonomique. Il serait assez tentant d'envisager de développer un outil de génomique comparée permettant de faire de l'évolution du développement *in silico*.

Bibliographie

Adachi J, Hasegawa M. (1996). **MOLPHY: programs for molecular phylogenetics.** Institute of Statistical Mathematics, Tokyo.

Adoutte A, Balavoine G, Lartillot N, Lespinet O, Prud'homme B, and de Rosa R. (2000) **The new animal phylogeny : Reliability and Implications.** PNAS USA ; 97(9):4453-4456.

Aguinaldo AM, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, and Lake JA. (1997). **Evidence for a clade of nematodes, arthropods and other moulting animals.** *Nature 387,489-93.*

Akimaru H, Hou DX, and Ishii S (1997). **Drosophila CBP is required for dorsal-dependent twist gene expression.** Nat Genet 17, 211-4.

Alberga A., Boulay JL, Kempe E, Dennefeld C, and Haenlin M. (1991). **The snail gene required for mesoderm formation in Drosophila is expressed dynamically in derivatives of all three germ layers.** Development 111, 983-92.

Altenhoff AM, Dessimoz C. (2009). **Phylogenetic and Functional Assessment of Orthologs Inference Projects and Methods.** PloS Computational Biology. 5(1):e1000262.

Altschul SF , Madden TL , Schaffer AA , Zhang J , Zhang Z , Miller W , and Lipman DJ. (1996). **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** Nucleic Acids Res. 25(17):3389-402.

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. (1990). **Basic local alignment search tool.** J. Mol. Biol. 215:403-410.

Anderson KV, and Nusslein-Volhard C. (1984). **Information for the dorsal--ventral pattern of the *Drosophila* embryo is stored as maternal mRNA.** Nature 311, 223-7.

Bairoch A. (2002). **The ENZYME databse in 2000.** Nucleic Acids Res. 28(1):304-5.

Batlle E, Sancho E, Franci C, Dominguez D, Monfar M, Baulida J, and Garcia De Herreros A. (2000). **The transcription factor snail is a repressor of E-cadheNucleic Acids Res. 2000 Jan 1;28(1):304-5.rin gene expression in epithelial tumour cells.** Nat Cell Biol 2, 84-9.

Bergemann M, Lespinet O, M'Barek SB, Daboussi MJ and Dufresne M. (2008) **Genome-wide analysis of the *Fusarium oxysporum* mimp family of MITEs and mobilization of both native and de novo created mimps**. Journal of Molecular Evolution; Dec;67(6):631-42.

Berman HM, Henrick K, Nakamura H. (2003). **Announcing the worldwide Protein Data Bank.** Nature Structural Biology. 10:980.

Bidard F, Imbeaud S, Reymond N, Lespinet O, Silar P, Clavé C, Delacroix H, Berteaux-Lecellier V and Debuchy R. (2010). **A general framework for optimization of probes for gene expression microarray and its application to the fungus *Podospora anserina*.** BMC Research Notes. 3:171.

Blanco E, Parra G, Guigó R. (2007). **Using geneid to identify genes.** Curr Protoc Bioinformatics. Chapter 4:Unit 4.3.

Cano A, Perez-Moreno MA, Rodrigo I, Locascio A, Blanco MJ, del Barrio MG, Portillo F, and Nieto MA. (2000). **The transcription factor snail controls epithelial-mesenchymal transitions by repressing E-cadherin expression.** Nat Cell Biol 2, 76-83.

Casal J, and Leptin M. (1996). **Identification of novel genes in Drosophila reveals the complex regulation of early gene activity in the mesoderm.** Proc Natl Acad Sci U S A 93, 10327-32.

Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, Kaipa P, Karthikeyan AS, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Paley S, Popescu L, Pujar A, Shearer AG, Zhang P, Karp PD. (2010). **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases.** Nucleic Acids Res. 38(Database issue):D473-9.

Chang A, Scheer M, Grote A, Schomburg I, Schomburg D. (2009). **BRENDA, AMENDA and FRENDA the enzyme information system : new content and tools in 2009.** Nucleic Acids Res. 37(Database issue):D588-92.

Christie KR, Weng S, Balakrishnan R, Costanzo MC, Dolinski K, Dwight SS, Engel SR, Feierbach B, Fisk DG, Hirschman JE, Hong EL, Issel-Tarver L, Nash R, Sethuraman A, Starr B, Theesfeld CL, Andrada R, Binkley G, Dong Q, Lane C, Schroeder M, Botstein D, Cherry JM. (2004). **Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms.** Nucleic Acids Res. 32(Database issue):D311-4.

Costa M, Wilson ET, and Wieschaus E. (1994). **A Putative Cell Signal Encoded by the folded gastrulation Gene Coordinates Cell Shape Changes during Drosophila Gastrulation.** Cell 76, 1075-1089.

Csurös M. (21010). **Count: evolutionary analysis of phylogenetic profiles with parsimony a and likelihood .** Bioinformatics. 26(15) : 1910–1912.

Cuomo CA, Birren BW. (2010). **The Fungal Genome Initiative and lessons learned from genome sequencing.** Methods in Enzymology, Guide to Yeast Genetics and Molecular Cell Biology. 470:833-855.

de Rosa R, Grenier JK, Andreeva T, Cook CE, Adoutte A, Akam M, Carroll SB, Balavoine G. (1999) **Hox genes in brachiopods and priapulids and protostome evolution.** Nature ; 399(3678):772-6.

de Rosa R, Labedan B. (1998). **The evolutionary relationships between the two bacteria Escherichia coli and Haemophilus influenzae and their putative last common ancestor.** Mol Biol Evol. 15(1):17-27

Descorps-Declère S, Lemoine F, Sculo Q, Lespinet O and Labedan B. (2008). **The multiple facets of homology and their use in comparative genomics to study the evolution of genes, genomes, and species.** Biochimie. 90(4):595-608.

Eddy SR. (1998). **Profile Hidden Markov models.** Bioinformatics. 14(9)**:**755-63.

Edgar C. (2004). **MUSCLE : a multiple sequence alignment method with reduced time and space complexity.** BMC Bioinformatics. 5:113.

Elshafei AM, Abdel-Fatah OM. (2001). **Evidence for a non-phosphorylated route of galactose breakdown in cell-free extracts of *Aspergillus niger*.** Enzyme Microb Technol. 29(1):76-83.

Enright AJ, Van Dongen S, Ouzounis CA. (2002). **An efficient algorithm for large-scale detection of protein families.** Nucleic Acids Res. 30(7):1575-84.

Espagne E, Lespinet O, Malagnac F, Da Silva C, Jaillon O, Porcel BM, Couloux A, Aury JM, Ségurens B, Poulain J, Anthouard V, Grossetête S, Khalili H, Coppin E, Déquard-Chablat M, Picard M, Contamine V, Arnaise S, Bourdais A, Berteaux-Lecellier V, Gautheret D, de Vries RP, Battaglia E, Coutinho PM, Danchin EGJ, Henrissat B, El Khoury R, Sainsard-Chanet A, Boivin A, Pinan-Lucarré B, Sellem CH, Debuchy R, Wincker P, Weissenbach J and Silar P. (2008). **The Genome Sequence of the Model Ascomycete Fungus Podospora anserina.** Genome Biology. 9(5):R77.

Essex LJ, Mayor R, and Sargent MG. (1993). **Expression of Xenopus Snail in Mesoderm and Prospective Neural Fold Ectoderm.** Developmental Dynamics 198, 108-122.

Falb M, Muller K, Konigsmaier L, Tanja Oberwinkler T, Horn P, von Gronau S, Gonzalez O, Pfeiffer F, Bornberg-Bauer E, Oesterhelt D. (2008). **Metabolism of halophilic archaea .** Extremophiles. 12:177–196 .

Fleischmann A, Darsow M, Degtyarenko K, Fleischmann W, Boyce S, Axelsen KB, Bairoch A, Schomburg D, Tipton KF, Apweiler R. (2004). **IntEnz, the integrated relational enzyme database.** Nucleic Acids Res. 32(Database issue):D434-7.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, and Postlethwait J. (1999). **Preservation of duplicate genes by complementary, degenerative mutations.** Genetics 151, 1531-45.

François V, Solloway M, O'Neill JW, Emery J, and Bier E. (1994). **Dorsal-ventral patterning of the Drosophila embryo depends on a putative negative growth factor encoded by short gastrulation gene.** Genes & Development 8, 2602-2616.

Frey PA. **The Leloir pathway: a mechanistic imperative for three enzymes to change the stereochemical configuration of a single carbon in galactose.** FASEB J. 1996 Mar;10(4):461-70.

Fulton DL, Li YY, Laird MR, Horsman BG, Roche FM, Brinkman FS. (2006). **Improving the specificity of high-throughput ortholog prediction.** BMC Bioinformatics. 7:270

Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S, Rehman B, Elkins T, Engels R, Wang S, Nielsen CB, Butler J, Endrizzi M, Qui D, Ianakiev P, Bell-Pedersen D, Nelson MA, Werner-Washburne M, Selitrennikoff CP, Kinsey JA, Braun EL, Zelter A, Schulte U, Kothe GO, Jedd G, Mewes W, Staben C, Marcotte E, Greenberg D, Roy A, Foley K, Naylor J, Stange-Thomann N, Barrett R, Gnerre S, Kamal M, Kamvysselis M, Mauceli E, Bielke C, Rudd S, Frishman D, Krystofova S, Rasmussen C, Metzenberg RL, Perkins DD, Kroken S, Cogoni C, Macino G, Catcheside D, Li W, Pratt RJ, Osmani SA, DeSouza CP, Glass L, Orbach MJ, Berglund JA, Voelker R, Yarden O, Plamann M, Seiler S, Dunlap J, Radford A, Aramayo R, Natvig DO, Alex LA, Mannhaupt G, Ebbole DJ, Freitag M, Paulsen I, Sachs MS, Lander ES, Nusbaum C, Birren B. (2003). **The genome sequence of the filamentous fungus Neurospora crassa.** Nature. 422(6934):859-68.

Gautheret D, Lambert A. (2001). **Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles.** J Mol Biol. 313(5):1003-11.

Gonnet GH, Hallett MT, Korostensky C, and Bernardin L. (2000). **Darwin v.2.0 : an interpreted computer language for the biosciences.** Bioinformatics. 16(2):101-3.

Graïa F, Lespinet O, Rimbault B, Dequard-Chablat M, Coppin E and Picard M. (2001) **Genome quality control : RIP (Repeat Induced Point mutation) comes to *Podospora*.** *Molecular Microbiology ; 40(3):586-95.*

Grau Y, Carteret C, and Simpson P. (1984). **Mutations and chromosomal rearrangement affecting the expression of Snail, a gene involved in embryonic patterning in Drosophila melanogaster.** Genetics 108, 347-360.

Grossetête S, Labedan B and Lespinet O. (2010). **FUNGIpath: a tool to assess fungal metabolic pathways predicted by orthology.** BMC Genomics; Feb 1;11(1):81.

Grossetête S. (2010). **Génomique comparée et développement de nouveaux outils bioinformatiques permettant l'analyse de la diversité métabolique des Eumycota.** Thèse de l'Université Paris-Sud 11.

Guidon V, Gascuel O. (2003). **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** Syt Biol. 52:696-704.

Halanych KM, Bacheller JD, Aguinaldo AMA, Liva SM, Hillis DM, and Lake JA. (1995) **18S rDNA evidence that the lophophorates are protostome animals.** Science 267, 1641-1643.

Hammerschmidt M, and Nusslein-Volhard C. (1993). **The expression of a zebrafish gene homologous to Drosophila snail suggests a conserved function in invertebrate and vertebrate gastrulation.** Development 119,1107-18.

Hasegawa M, Kishino H. (1994). **Accuracies of the simple methods for estimating the bootstrap probability of a maximum-likelihood tree.** Mol Biol Evol 11:142-145.

Hopwood ND, Pluck A, and Gurdon JB. (1989). **A Xenopus mRNA related to Drosophila twist is expressed in response to induction in the mesoderm and the neural crest.** Cell 59, 893-903.

Hubbard TJP, Aken BL, Beal K, Ballester B, Caccamo M, et al. (2007). **Ensembl 2007.** Nucleic Acids Res 35: D610–D617.

Ip YT, Levine M, and Bier E. (1994). **Neurogenic expression of snail is controlled by separable CNS and PNS promoter elements.** Development 120, 199-207.

Ip YT, Park RE, Kosman D, Bier E, and Levine M. (1992). **The dorsal gradient morphogen regulates stripes of rhomboid expression in the presumptive neuroectoderm of the Drosophila embryo.** Genes Dev 6, 1728-39.

Jaffe DB, Butler J, Gnerre S, Mauceli E, Lindblad-Toh K, Mesirov JP, Zody MC, Lander ES. (2003). **Whole-genome sequence assembly for mammalian genomes: Arachne 2.** Genome Res. 13(1):91-6.

Jones DT, Taylor WR, Thornton JM. (1992). **The rapid generation of mutation data matrices from protein sequences.** Comput Appl Biosci 8:275-282.

Jordan IK, Makarova KS, Spouge JL, Wolf YI, and Koonin EV. 2001. **Lineage-specific gene expansions in bacterial and archaeal genomes.** Genome Res. 11: 555-565.

Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. (2010). **KEGG for representation and analysis of molecular networks involving diseases and drugs.** Nucleic Acids Res. 38(Database issue):D355-60.

Karp PD. (2004). **Call for an enzyme genomics initiative.** Genome Biol. 5(8):401.

Kasai J, Nambu JR, Lieberman PM, and Crews ST. (1992). **Dorsal-ventral patterning in Drosophila : DNA binding of snail protein to the single-minded gene.** Proc. Natl. Acad. Sci. USA 89, 3414-3418.

Koski B, Golding GB. (2001). **The closest BLAST hit is often not the nearest neighbor.** J Mol Evol. 52:540-542

Kosman D, Ip YT, Levine M, and Arora K. (1991). **Establishment of the mesoderm-neuroectoderm boundary in the Drosophila embryo.** Science 254, 118-22.

Kovacs,IA, Palotai R, Szalay MS, Csermely P. (2010). **Community Landscapes: An Integrative Approach to Determine Overlapping Network Module Hierarchy, Identify Key Nodes and Predict Network Dynamics.** PLoS One. 5(9) : e12528.

Kurtzman CP, Fell JW. (1998). **The Yeasts, a Taxonomic Study.** *Elsevier,Amsterdam.*

Labedan B and Lespinet O. (2006). **Interspecies and intraspecies comparison of microbial proteins: learning about gene ancestry, protein function, and species life style.** WHILEY, Microbial Proteomics: Functionnal Biology of Whole Organisms. Editeurs Ian Humphery-Smith, Michael Hecker.

Lai Z, Fortini ME, and Rubin GM. (1991). **The embryonic expression patterns of zfh-1 and zfh-2, two *Drosophila* genes encoding novel zinc-finger homeodomain proteins.** Mechanisms of Development 34, 123-134.

Lartillot N, Lespinet O, Vervoort M, van den Biggelaar JAM, and Adoutte A. (2002). Expression pattern of Brachyury in the mollusc Patella vulgata suggests a conserved role in the establishment of the A-P axis in Bilateria. Development; 129(6):1411-21.

Le Bouder-Langevin S, Capron-Montaland I, de Rosa R, and Labedan B. (2002). **A strategy to retrieve the whole set of protein modules in microbial proteomes.** Genome Research. 12(12):1961-73

Lemoine F, Labedan B and Lespinet O. (2008). **SynteBase/SynteView: a tool to visualize gene order conservation in prokaryotic genomes.** BMC Bioinformatics; Dec 16;9:536.

Lemoine F, Lespinet O and Labedan B. (2007). **Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data.** BMC Evolutionary Biology; 7(1):237.

Leptin M. (1999). **Gastrulation in Drosophila: the logic and the cellular mechanisms.** Embo J 18, 3187-92.

Leptin M, Casal J, Grunewald B, Reuter R. (1992) **Mechanism of early Drosophila mesoderm formation.** Dev. Suppl. 1992:23-31

Leptin M. (1991). **twist and snail as positive and negative regulators during *Drosophila* mesoderm development.** Genes Dev 5, 1568-76.

Leptin M, and Grunewald B. (1990). **Cell shape changes during gastrulation in Drosophila.** Development 110,73-84.

Lespinet O and Labedan B. (2006a). **Puzzling over orphan enzymes.** Cellular and Molecular Life Science; 63(5):517-23.

Lespinet O and Labedan B. (2006b). **Orphan enzymes could be an unexplored reservoir of new drug targets.** Drug Discovery Today; 11(7-8):300-5.

Lespinet O and Labedan B. (2006c). **ORENZA: a web resource for studying ORphan ENZyme Activities.** (2006) - BMC Bioinformatics; 7:436.

Lespinet O and Labedan B. (2005). **Orphan enzymes?** Science; 307(5706):42.

Lespinet O, Nederbragt AJ, Cassan M, Dictus WJAG, van Loon AE, and Adoutte A. (2002a). **Characterisation of two Snail genes in the gastropod mollusc *Patella vulgata*. Implications for understanding the ancestral function of the Snail-related genes in *Bilateria*.** Development, Gene and Evolution; 212(4):186-95.

Lespinet O, Wolf YL, Koonin EV and Aravind L. (2002b). The role of lineage-specific gene family expansion in the evolution of eukaryotes. Genome Research; 12(7):1048-59.

Lespinet O. (2001). **La famille des genes *Snail*. Caractérisation de deux nouveaux membres chez le mollusque *Patella vulgata*. Hypothèse sur leur fonction ancestrale chez les *Bilateria*.** Thèse de l'université Paris-Sud 11.

Li L, Stoeckert CJ Jr, Roos DS. (2003). **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res.* **13:**2178-2189.

Liolios K, Chen IM, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM, Kyrpides NC. (2010). **The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata.** 38(Database issue):D346-54.

Lowe TM, Eddy SR. (1997). **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** Nucleic Acids Res. 25(5):955-64.

Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH. (2002). **CDD : a database of conserved domain alignments with links to domain three-dimentional structure.** Nucleic Acids Res. 30(1):281-3.

Morisato D, and Anderson KV. (1995). **Signaling pathways that establish the dorsal-ventral pattern of the Drosophila embryo.** Annu Rev Genet 29, 371-99.

Mulder N, Apweiler R. (2007). **InterPro and InterProScan: tools for protein sequence classification and comparison.** Methods Mol Biol. 396:59-70.

Nederbragt AJ, Lespinet O, van Wageningen S, van Loon AE, Adoutte A, and Dictus WJAG. (2002). A lophotrochozoan ***twist*** gene is expressed in the ectomesoderm of the gastropod mollusc ***Patella vulgata.*** Evolution & Development; 4(5):334-43.

Nibu Y, Zhang H, Bajor E, Barolo S, Small S, and Levine M. (1998). **dCtBP mediates transcriptional repression by knirps, Kruppel and snail in the drosophila embryo.** Embo J 17, 7009-20.

Noé L, Kucherov G. **YASS: enhancing the sensitivity of DNA similarity search.** Nucleic Acids Res. 33(Web Server issue):W540-3.

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. (1999). **The use of gene clusters to infer functional coupling.** Proc Natl Acad Sci U S A 96: 2896–2901.

Pan DJ, Huang JD, and Courey AJ. (1991). **Functional analysis of the Drosophila twist promoter reveals a dorsal- binding ventral activator region.** Genes Dev 5, 1892-901.

Paulsen IT, Banerjei L, Myers GS, Nelson KE, Seshadri R, Read TD, Fouts DE, Eisen JA, Gill SR, Heidelberg JF, Tettelin H, Dodson RJ, Umayam L, Brinkac L, Beanan M, Daugherty S, DeBoy RT, Durkin S, Kolonay J, Madupu R, Nelson W, Vamathevan J, Tran B, Upton J, Hansen T, Shetty J, Khouri H, Utterback T, Radune D, Ketchum KA, Dougherty BA, Fraser CM. (2003). **Role of mobile DNA in the evolution of vancomycin-resistant *Enterococcus faecalis*.** Science. 299(5615):2071-4.

Peláez F. (2005). **Biological activities of fungal metabolites.** Handbook of Industrial Mycology (ed. Zhiqiang An). 54-104 (Marcel Dekker, New York).

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. (1999). **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** Proc Natl Acad Sci U S A. 96(8):4285-8.
Poitelon JB, Joyeux M, Welté B, Duguet JP, Prestel E, Lespinet O and DuBow MS. (2009). **Assessment of phylogenetic diversity of bacterial microflora in drinking water using serial analysis of ribosomal sequence tags.** Water Research; Sep;43(17):4197-206.

Ray RP, Arora K, Nüsslein-Volhard C, and Gelbart WM. (1991). **The control of cell fate along the dorsal-ventral axis of the Drosophila embryo.** Development 113, 35-54.

Remm M, Storm CE, Sonnhammer EL. (2001). **Automatic clustering of orthologs and in paralogs from pairwise species comparisons.** J Mol Biol. 314:1041-1052.

Riley M, and Labedan B. (1997) **Protein evolution viewed through Escherichia coli protein sequences : introducing the notion of a structural segment of homology, the module.** J Mol Bio. 268(5):857-68.l

Rizet G. (1941). **Sur l'analyse génétique des asques du *Podospora anserina*.** C R Acad Sci. 212:59-61.

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. (2000). **Artemis: sequence visualization and annotation.** Bioinformatics. 16(10):944-5.

Sargent MG, and Bennett MF. (1990). **Identification in Xenopus of a structural homologue of the Drosophila gene snail.** Development 109, 967-73.

Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Mizrachi I, Ostell J, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Yaschenko E, Ye J. (2009). **Database resources of the National Center for Biotechnology Information.** Nucleic Acids Res. 37(Database issue):D5-15.

Sculo Q, Lespinet O, and Labedan B. (2005). **New approaches to improve the soundness of the deep evolutionary relationships in genomic trees of microorganisms.** JOBIM, Lyon 6-8 Juillet, Editeurs G Perrière, A Guénoche, C Geourjon.

Sculo Q, Lespinet O, and Bernard Labedan B. (2003). **Retrieving the whole set of protein modules of campylobacter jejeuni and Helicobacter pylori.** Genome Letters. 2:8-15.

Sefton M, Sanchez S, and Nieto MA. (1998). **Conserved and divergent roles for members of the Snail family of transcription factors in the chick and mouse embryo.** Development 125, 3111-21.

Silar P, Barreau C, Debuchy R, Kicka S, Turcq B, Sainsard-Chanet A, Sellem CH, Billault A, Cattolico L, Duprat S, Weissenbach J. (2003). **Characterization of the genomic organization of the region bordering the centromere of chromosome V of Podospora anserina by direct sequencing.** Fungal Genet Biol. 39(3):250-63.

Simpson P. (1983). **Maternal-zygotic gene interactions during formation of the dorsoventral pattern in drosophila embryos.** Genetics 105, 615-632.

Slot JC, Rokas A. **Multiple GAL pathway gene clusters evolved independently and by different mechanisms in fungi.** Proc Natl Acad Sci U S A. 2010 Jun 1;107(22):10136-41.

Smith JC, Price BM, Green JBA, Weigel D, and Herrmann BG. (1991). **Expression of a Xenopus Homolog of Brachyury (T) Is an Immediate-Early Response to Mesoderm Induction.** Cell 67, 79-87.

Smith TF, Waterman MS. (1981). **Identification of common molecular subsequences. Journal of Molecular Biology.** 147(1):195-7

Sokal R and Michener C. (1958). **A statistical method for evaluating systematic relationships.** University of Kansas Science Bulletin, 38:1409-1438, 1958.

Thisse C, Thisse B, and Postlethwait JH. (1995). **Expression of snail2, a second member of thezebrafish snail family, in cephalic mesendoderm and presumptive neural crest of wild-type andspadetail mutant embryos.** Dev Biol 172, 86-99.

Thisse C, Thisse B, Schilling T., and Postlethwait JH. (1993). **Structure of the zebrafish snail1 gene and its expression in wild-type, spadetail and no tail mutant embryos.** Development 119, 1203-15.

Thisse C, Perrin-Schmitt F, Stoetzel C, Thisse B. (1991) **Sequence-specific transactivation of the *Drososphila* twist gene by the dorsal gene product.** Cell ; 68(7):1191-201

Thisse B, Stoetzel C, Gorostiza-Thisse C, and Perrin-Schmitt F. (1988). **Sequence of the twist gene and nuclear localization of its protein in endomesodermal cells of early Drosophila embryos.** Embo J 7, 2175-83.

UniProt Consortium.(2010). **The Universal Protein Ressource (UniProt) in 2010**. Nucleic Acids Res. 38(Database issue):D142-8.

van Dongen S. (2000). **Graph Clustering by Flow Simulation.** PhD thesis, University of Utrecht.

Vert JP. (2002). **A tree kernel to analyse phylogenetic profiles.** Bioinformatics. 18 Suppl 1:S276-84.

Wall DP, Fraser HB, Hirsh AE (2003) **Detecting putative orthologs.** Bioinformatics 19: 1710–1711.

Wheelan SJ, Marchler-Bauer A, Bryant SH. (2000). **Domain size distribution can predict domain boundaries.** Bioinformatics. 16:613-618.

Whitaker JW, Letunic I, McConkey GA, Westhead DR. (2009). **metaTIGER: a metabolic evolution resource.** Nucleic Acids Res. 37 (Database issue):D531-8.

Wylie T, Martin J, Abubucker S, Yin Y, Messina D, Wang Z, McCarter JP, Mitreva M. (2008). **NemaPath: online exploration of KEGG-based metabolic pathways for nematodes.** BMC Genomics. 9:525.

Yamada T, Kanehisa M and Goto S. (2006). **Extraction of phylogenetic network modules from the metabolic network.** BMC Bioinformatics. 7:130

- Annexes -

5 Publications Significatives

1. **The multiple facets of homology and their use in comparative genomics to study the evolution of genes, genomes, and species.** (2008) *S Descorps-Declère, F Lemoine, Q Sculo, O Lespinet and B Labedan* - Biochimie; Apr;90(4):595-608.

2. **SynteBase/SynteView: a tool to visualize gene order conservation in prokaryotic genomes.** (2008) *F Lemoine, B Labedan and O Lespinet* - BMC Bioinformatics; Dec 16;9:536.

3. **ORENZA: a web resource for studying ORphan ENZyme Activities.** (2006) *O. Lespinet and B Labedan* - BMC Bioinformatics; 7:436.

4. **The Genome Sequence of the Model Ascomycete Fungus *Podospora anserina*.** (2008) *E Espagne*, O Lespinet*, F Malagnac*, C Da Silva, O Jaillon, BM Porcel, A Couloux, JM Aury, B Ségurens, J Poulain, V Anthouard, S Grossetête, H Khalili, E Coppin, M Déquard-Chablat, M Picard, V Contamine, S Arnaise, A Bourdais, V Berteaux-Lecellier, D Gautheret, RP de Vries, E Battaglia, PM Coutinho, EGJ Danchin, B Henrissat, R El Khoury, A Sainsard-Chanet, A Boivin, B Pinan-Lucarré, CH Sellem, R Debuchy, P Wincker, J Weissenbach and Philippe Silar* - Genome Biology; 9(5):R77.

5. **FUNGIpath: a tool to assess fungal metabolic pathways predicted by orthology.** (2010) *S Grossetête, B Labedan and O Lespinet* - BMC Genomics; Feb 1;11(1):81.

_____

Articles n°4 et 5 : Ces deux articles ont obtenu l'étiquette ”Highly Accessed” décernée par les journaux du groupe BMC

Article n°4 : Les trois premiers auteurs (*) ont contribué à part égale à ce travail.

Available online at www.sciencedirect.com
ScienceDirect
Biochimie 90 (2008) 595–608
BIOCHIMIE
www.elsevier.com/locate/biochi

Review

The multiple facets of homology and their use in comparative genomics to study the evolution of genes, genomes, and species

Stéphane Descorps-Declère [a], Frédéric Lemoine [a,b], Quentin Sculo [a], Olivier Lespinet [a], Bernard Labedan [a,*]

[a] *Institut de Génétique et Microbiologie, CNRS UMR 8621, Bâtiment 400, Université Paris Sud XI, 91405 Orsay Cedex, France*
[b] *Laboratoire de Recherche en Informatique, CNRS UMR 8623, Bâtiment 490, Université Paris Sud XI, 91405 Orsay Cedex, France*

Received 14 June 2007; accepted 14 September 2007
Available online 22 September 2007

Abstract

The incredible development of comparative genomics during the last decade has required a correct use of the concept of homology that was previously utilized only by evolutionary biologists. Unhappily, this concept has been often misunderstood and thus misused when exploited outside its evolutionary context. This review brings back to the correct definition of homology and explains how this definition has been progressively refined in order to adapt it to the various new kinds of analysis of gene properties and of their products that appear with the progress of comparative genomics. Then, we illustrate the power and the proficiency of such a concept when using the available genomics data in order to study the evolution of individual genes, of entire genomes and of species, respectively. After explaining how we detect homologues by an exhaustive comparison of a hundred of complete proteomes, we describe three main lines of research we have developed in the recent years. The first one exploits synteny and gene context data to better understand the mechanisms of genome evolution in prokaryotes. The second one is based on phylogenomics approaches to reconstruct the tree of life. The last one is devoted to reminding that protein homology is often limited to structural segments (SOH = segment of homology or module). Detecting and numbering modules allows tracing back protein history by identifying the events of gene duplication and gene fusion. We insist that one of the main present difficulties in such studies is a lack of a reliable method to identify genuine orthologues. Finally, we show how these homology studies are helpful to annotate genes and genomes and to study the complexity of the relationships between sequence and function of a gene.

Keywords: Homology; Paralogy; Orthology; Xenology; Positional orthologous gene; Segment of homology; Module; Modular structure; Phylogenetic profile; Phylogenomics; Tree of life

1. Introduction

Comparing the infinite variety of living beings and of their features has long been a central task in Biology. In this context, Owen was the first to define homology as the property of *the same organ that is found in different animals under every variety of form and function* [1].

For a while, this concept of homology was used only by evolutionary biologists while the majority of the other life scientists did not consider it or − maybe worse − misused it [2,3].

Since we entered the Genomics Era twelve years ago, comparative biology and use of the notion of homology have become again fashionable, but many present-day biologists are still incorrectly using this important concept [2,3].

In this review, we first remind what the correct definition of homology is and explain how this definition has been progressively refined in order to adapt it to the various new kinds of analysis of gene properties that appear with the progress of comparative genomics.

We further show how the correct use of the different facets of homology is a powerful and proficient tool for understanding

* Corresponding author. Tel.: +33 1 6915 3560; fax: +33 1 6915 7296.
E-mail address: bernard.labedan@igmors.u-psud.fr (B. Labedan).

0300-9084/$ - see front matter © 2007 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.biochi.2007.09.010

gene function and genome evolution. Using this conceptual framework, it becomes possible to: (i) trace back protein history by identifying all events of gene duplication and gene fusion; (ii) define the core of conserved genes at internal nodes of the tree of life; and (iii) reconstructing the ancestral genomes and the evolution of present-day microbial species.

Finally, this review emphasizes how crucial are homology approaches to annotate and/or reannotate genes and genomes and to assess the molecular and/or cellular function of a gene product.

2. Homology: variations on the same theme

2.1. Defining the concept of homology

Although the conceptual definition of homology has been historically worked out in a more complex way (see Ref. [3] for comprehensive discussion), we only insist on what its basic and widely acknowledged definition is: *two items are defined as homologues if they share a common ancestry*. Such a definition has two decisive implications. (1) Homology is an all-or-none property (see Ref. [4] for thorough discussion). (2) Homology is *always* a hypothesis (see Refs. [4,5] for detailed discussion of this important point). Thus, another empirical criterion must be used to assess experimentally that two objects are homologues. A two-step procedure is generally used [6]: in a first step, proteins will be labelled (*primary*) homologues if their level of sequence similarity is higher than an imposed threshold (see, for example, Refs. [7,8]). But only a tree topology based on a protein multiple alignment will help to determine if the primary homologues share a common ancestry (*secondary homologues*).

It must be underlined that the concept of homology (evolution by divergence from a common ancestor) is contrasted with that of analogy where structures have evolved separately to perform similar functions (convergent evolution). Thus, similarity in itself might not be sufficient to distinguish between divergent (homology) and convergent (homoplasy = non-homology [9]) evolution. In other words, the observation that structural similarity among proteins is found greater than might be anticipated by chance is a criterion to *detect* homology but it not enough to *define* it [3,9]. Fitch [3] further insists: (i) that homology resides in the characters, not in their states; (ii) homoplasy is a relation of two character states (e.g. for an amino acid, say glycine and leucine) in a tree; whereas (iii) analogy is a relation of two characters (say amino acid), independent of any tree. Thus, in the case of molecules (*but not* of morphological structures), high level of identity means the sequences are homologous – especially if the ancestral sequences are significantly more alike than today's sequences as demonstrated for the alpha and beta hemoglobins [3,10]. If the reverse is true, the sequences are analogous [3,10]. In other words, the method is not circular; it does detect analogy when analogy occurred [3].

2.2. Distinguishing three main classes of homology

As early as 1970, Fitch [10] made an essential distinction between different classes of homology.

- *Orthologous* genes are homologous genes which diverged by speciation. Therefore, orthologues are the pertinent objects to use when reconstructing species trees and/or comparing gene order and evolution of genomes. Moreover, orthologous sequences provide useful information to assess organism's classification (taxonomic studies). Note that many people comparing genomes used the term *orthologue* as meaning functionally equivalent genes in different species. We, as others [3,11], insist that this usage is evolutionarily incorrect, although we acknowledge that, in many cases, orthologues display similar or identical functions.
- *Paralogous* genes descend from an ancestral duplication, independently of speciation. Thus, paralogues are helpful for understanding the course of protein (gene) evolution as long as the changes to sequences over time by processes of mutation, recombination and repair have not blurred the similarities.
- *Xenologous* genes have been laterally transmitted from one organism to another one by an external vector (phage, plasmid). Xenologues may blurr the interpretation of tree topology in phylogenetic studies [12], especially when the two organisms are evolutionarily distant. To complicate even more, it must be underlined that it is often difficult in comparative genomics to differentiate xenologues from ancient paralogous genes which have been separated by speciation and further submitted to an asymmetrical loss of the paralogous copies.

2.3. Refining the distinction of homologous entities inside paralogous and orthologous classes

More recently, the group of Sonnhamer [13] has distinguished two kinds of paralogues when comparing the genomes of various organisms. The so-called *inparalogues* correspond to genes that duplicated recently in one species *after* the last speciation event. This underlines that orthology between individual genes (one-to-one relationship) rarely exist when comparing completely sequenced genomes. Rather, we have one-to-many or many-to-many relationships making useless to search for "the true" orthologue of a gene (see below for detailed discussion). Instead, it is more appropriate to think in terms of group of orthologues (sometimes called co-orthologues). This concept of *group orthology* [14,15] allows to assemble genes that have a single representative in the last common ancestor of the compared species, independently of their present-day status of orthologue or inparalogue.

However, if one-to-one orthologous relationships are required, it is advisable to combine contextual information with protein sequence information to distinguish between the different putative co-orthologues detected in a genome. With the publication of an increasing number of genomes of species that are more or less distant in the taxonomic space, it was observed that a small proportion of orthologous genes, hereafter called Positional Orthologous Genes (POGs), conserve their local order [15–17] despite genome fluidity. Such

observations gave birth to the concept of genomic context [18,19]. Accordingly, this rare conservation of gene neighbourhood is interpreted as the signature of functional relationships between the products of these stably associated genes [20–26].

It must be further underlined that such positional homologous genes are very helpful when studying genome evolution and species differentiation since they mirror the ancestral gene order as detailed below (Lemoine et al., in press).

3. Searching strict synteny blocks: a blueprint of the putative architecture of the ancestral genome

3.1. Searching genuine orthologues

As underlined above, the search for *bona fide* orthologues is not a trivial task. During the immediate last years a consensus has been reached that the so-called bidirectional best hit method which have been very popular for a while [15,24,27,28], is giving erroneous results occasionally. Indeed, it has been shown that BLAST searches often return as the highest scoring hit a protein that is *not* the nearest phylogenetic neighbour of the query sequence [28] leading to striking errors. To cope with this difficulty, many groups have outlined new approaches, their diversity showing that none is satisfying in itself (see for instance Refs. [29–32]).

To obtain better and sounder data, we have recently proposed to use two different and complementary approaches (Lemoine et al., in press). First, we adapted a Reciprocal Smallest Distance (RSD) [29,32] approach by taking advantage of specific properties of the DARWIN AllAll program [33,34] that set it apart from BLAST algorithm. Indeed, the AllAll program uses a maximum likelihood (ML) approach to estimate the evolutionary distances separating homologous genes. This makes trivial to retaining the shortest distance – computed in PAM (point accepted mutation) units – for each pair of proteins that are analyzed by this ML approach. Accordingly, RSD orthologous pairs are easily determined when comparing two proteomes. However, we found that this one-to-one approach is missing a significant number of more complex orthologous relationships, especially when gene duplications arose after a recent speciation event. To handle this problem, we developed an *ad hoc* algorithm that analyses the leaves of a rooted phylogenetic tree to determining for each node whether its descendants arose from a gene duplication (sharing several sequences in the same species) or a speciation event (no common species). This phylogeny approach allows us to identify group orthology in trees, including the inparalogues, when analysing a family of homologous genes.

Although the total number of orthologous pairs detected by both methods (RSD and phylogeny) in comparing 107 species (supplementary Table 1) is found similar, it is striking that they have only two thirds of their predictions in common, namely 67% of the RSD orthologues and 69% of the phylogeny ones, respectively (Table 1). Both approaches, thus, yield valid but incomplete data. Phylogeny approach is occasionally missing some genuine orthologues, especially when the tree is too large and complex. However, it is very helpful to identify those RSD detected genes that are actually pseudoorthologues (actual paralogues that appear to be orthologous due to differential lineage-specific gene loss). Analysing the union of the result sets provided a total number of pairs of orthologues that can be used to further study the evolution of gene order in genomes comparison.

Table 1
Identifying sets of syntenic blocks using two different approaches to detect orthologues after comparing the proteomes of 107 microbial species

Detection method	Number of pairs of orthologues	Blocks of two adjacent orthologues	Syntenic blocks	
			Number	Mean size
RSD	2,332,248	290,108	182,289	2.59 genes
Phylogeny	2,255,324	302,468	186,744	2.62 genes
Union	3,014,995	377,256	235,519	2.60 genes
Intersection	1,572,577	226,799	149,058	2.62 genes

Note, however, that if one search genuine orthologues with a very high degree of confidence, it might be wiser to use only the intersection of these two complementary approaches. This will exclude the false positives returned by both methods, but also, unhappily, authentic cases including inparalogues.

3.2. Detecting all synteny blocks

Once all *bona fide* orthologues are uncovered, we further determine how many ones form reciprocal pairs of adjacent genes in at least two genomes. Then, we extend the detected neighbourhood relationships to larger syntenic blocks. The quadruplets were analyzed with *synblock*, an algorithm we designed to map synteny blocks of size 2 and to merge them as soon as they share a common pair of orthologous genes. Since our goal is to recover the ancestral gene order, we required strict gene adjacency, forbidding any insertion in a synteny block of a gene which would be unique to one of the pair of compared genomes.

Table 1 shows that we obtained a total of 149,058 synteny blocks with a mean size of 2.62 genes per block. These data confirm how fluid prokaryotic genomes are when they are studied at any taxonomic distance.

3.3. Positional Orthologous Genes (POGs) have a low evolution rate

Analysing in details the full set of POGs in 107 genomes allowed us to further observe that these peculiar orthologues display specific features. Table 2 summarizes the data obtained when comparing the model organism *E. coli* with various bacteria and archaea separated from it by increasing ranks of the taxonomy hierarchy such as family, order, class, phylum and Domain, respectively. Such comparisons show that the mean PAM distance separating the orthologues of *E. coli* from those of other organisms increases rapidly when moving far away in the taxonomic space from family to class ranks, reaching a plateau value of around 120 PAM units when comparing the different bacterial phyla and around 140 PAM units when comparing Domains Bacteria and Archaea. Table 2 further

Table 2
The mean PAM distance separating two orthologues and the average size (number of genes) of synteny blocks are dependent on the taxonomic distance separating the two genomes that have been compared at the level of their genetic context

Rank	Species 1 (*E. coli*)[a] taxonomy	Species 2			Mean PAM distance	Synteny block	
		Species name	Taxonomy	Proteome size		Mean size	Longest size
Family	Enterobacteriaceae	*S. enterica*	Enterobacteriaceae	4318	18.7	3.47	20
Order	Enterobacteriales	*V. cholerae*	Vibrionales	3835	69.5	2.96	10
		P. aeruginosa	Pseudomonadales	5567	85.6	2.94	12
Class	Gammaproteobacteria	*M. loti*	alphaproteobacteria	6746	110.8	2.66	9
Phylum	Proteobacteria	*B. subtilis*	Firmicutes	4112	112.8	2.44	9
		M. tuberculosis	Actinobacteria	3995	129.9	2.47	6
		C. tepidum	Bacteroidetes/Chlorobi	2252	117.8	2.70	9
		R. baltica	Planctomycetes	7325	127.1	2.39	8
Domain	Bacteria	*M. acetivorans*	Archaea (Euryarchaeota)	4540	139.8	2.19	3
		S. solfataricus	Archaea (Crenarchaeota)	2977	145.3	2.09	3

[a] Proteome size of *E. coli* K12: 4279.

shows that the size of conserved synteny blocks also depends on the phylogenetic (taxonomic) distance between species. Indeed, the mean maximum size of a synteny block is nearly 3.5 genes when comparing two closely related bacteria such as the Enterobacteriaceae *E. coli* and *Salmonella enterica*, but goes down to the minimal size of 2 when comparing a bacterium (*E. coli*) and an archaeon (*M. acetivorans*), although these genomes encode a similar range of proteins. Likewise, the longest synteny block is only 3 when comparing domains Bacteria and Archaea, but goes up to around 9 when comparing organisms of the same class or phylum, and around 11 within the same order. The longest block for the two studied Enterobacteriaceae contains 20 adjacent genes. The record up to now is 30 strictly adjacent genes for the pair *Bacillus subtilis − B. halodurans*.

We found that the evolutionary distance separating POGs levels off at less than 150 PAM units, even for extremely distant species (Table 2). This suggests that some strong evolutionary constraint is exerted on these peculiar genes. To check this model, we further compared the PAM distance distributions for orthologous genes located either inside synteny blocks (the so-called POGs) or outside these blocks. Table 3 shows that PAM distances appear to be lower inside than outside synteny blocks. This difference in sequence conservation (already occasionally observed [20]) appears to be independent of the taxonomic distance separating species, even though there are fewer and fewer synteny blocks as the taxonomic distance increases. We further used a bootstrap sampling approach to confirm that the difference between the PAM distances separating orthologous genes located inside and outside synteny blocks is statistically significant for any pair of species in the whole set of compared genomes. Fig. 1 shows that the overwhelming majority of the statistical tests reject the null hypothesis H_0 which assumes that proteins encoded by POGs evolve at the same evolutionary rate as those encoded by orthologous genes located outside the synteny blocks.

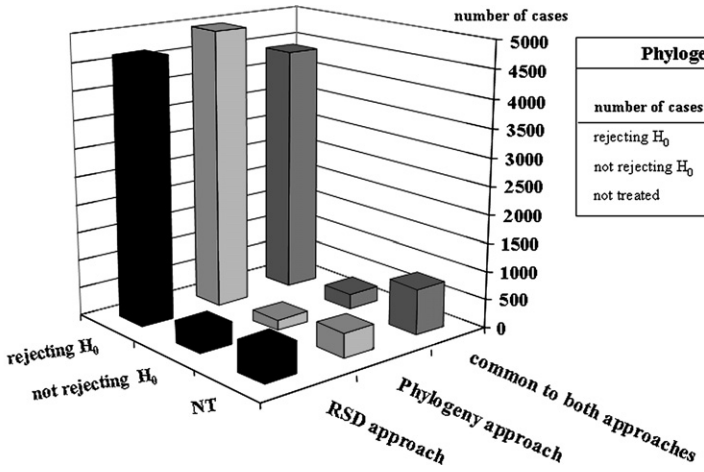
Table 3
Comparing the means of Pam distances and the total number of orthologues located inside or outside synteny blocks and shared by four pair of genomes

E. coli proteome compared with	Mean PAM distance		Number of orthologous pairs		
	Inside	Outside	Total	% inside	% outside
S. enterica	11.93	23.53	2592	27	73
B. subtilis	86.27	109.48	994	23	77
B. thetaiotaomicron	109.14	114.42	802	16	84
M. acetivorans	113.01	124.53	431	14	86

Thus, there is a universal trend in the evolution of prokaryotic genomes: genes located in regions with preserved gene order (POGs) evolve more slowly than genes found outside. This would fit well with the concept of genomic context in which it is assumed that neighbouring genes are participating to the same biological process. However, in most cases it is not yet clear what the actual cause of low evolutionary rate is (Lemoine et al., in press).

4. Families of orthologues help to reconstructing the tree of life

Microbiology has long suffered from the lack of sound taxonomic relationships between prokaryotic species (see Ref. [35] for a review). The comparison of 16S ribosomal RNA sequences has been an immense progress in the (re)definition of the main groups of bacteria and the foundation of the concept of archaea [35,36]. However, the phylogenetic tree built using these RNA sequences for the whole set of available species is unresolved at the level of its deeper nodes, whichever the method used to reconstruct this tree. Various approaches have been used to reconstruct a tree of life by using the maximum of available genomic data (see Ref. [37] for a review). However, these different phylogenomic trees differed in their topology and were also unresolved at the level of their deeper nodes. Moreover, it has been suggested by several groups that such tree reconstruction is meaningless due to a high rate of horizontal transfer between prokaryotic species [12]. Since we set up a sound method to find out genuine orthologues (see above), we further used them to meet this topology problem of the prokaryotic tree.



Phylogeny approach	tree complexity	
number of cases	0	1
rejecting H_0	4099	4201
not rejecting H_0	174	79
not treated	1283	1276

Fig. 1. Bootstrap analysis of the distribution of PAM distances separating pairs of orthologues located inside and outside synteny blocks. The number of cases rejecting and not rejecting hypothesis H_0, and those corresponding to cases too small to be included in this statistical test are shown for each method (RSD and phylogeny) and their intersection. The inset table details the data for trees of complexity 0 (orthologues only) or 1 (orthologues and in-paralogues).

4.1. Reconstructing distance trees using a phylogenomics approach

The families of homologues obtained as described in Section 2.1 were filtered to eliminate non-orthologous links (28.7% of the total), and further used to calculating the evolutionary distance between a pair of genomes as the mean of the PAM distances separating each pair of orthologues common to these two genomes in each filtered family. These evolutionary distances were further used to build a matrix and to derive a distance tree using the Neighbor-Joining algorithm [38].

In order to get genomic trees reflecting the diversity of the set of studied genomes and to minimize the impact of possible lateral transfer, we used only the most ubiquitous families. Two examples are shown in Supplementary Figs. 1 and 2 where we asked for the presence in each family of members belonging to at least 20 and 80 genomes, respectively. Although only 9.1% and 0.8% of the total families were respectively used, the phylogenetic signal was clearly increased.

Moreover, these family trees (Supplementary Figs. 1 and 2) display a topology where the positions of many groups of organisms, as well as that of their most external nodes are very similar to their accepted taxonomic distribution. We note however two differences. (i) The epsilon proteobacteria are not grouping with the other proteobacteria. Such an unexpected position has been already observed in other trees based on a gene content approach [37]. (ii) The mesophilic methanogen *M. acetivorans* and the two halophiles *Halobacterium species NRC-1* and *Haloarcula marismortui* are emerging at the basis of the archaeal subtree and are not grouped with the other (extremophile) euryarchaeota. This may reflect an adaptation to a non-thermophilic life style and a few cases of lateral transfers could not be excluded.

4.2. Reliable deep nodes of the Domain Bacteria

The increase of the number of genomes per family enlarged the distance separating the node common to the grouping of archaea and eukaryotes from bacteria and the distances between the main bacterial branches. After collapsing their external branches, it was possible to compare the simplified topologies of the trees shown in Supplementary Figs. 1 and 2 (and others not shown trees), respectively. Fig. 2 emphasizes several of the remarkable results obtained with this approach — that set it apart from previous works [37] — by underlining the species groupings which are found constantly whatever the number of families used to reconstruct trees: One of the most interesting point is the grouping of the two hyperthermophiles *Aquifex* and *Thermotoga* together and with the Firmicutes and *Fusobacterium* (shaded rectangle in Fig. 2).

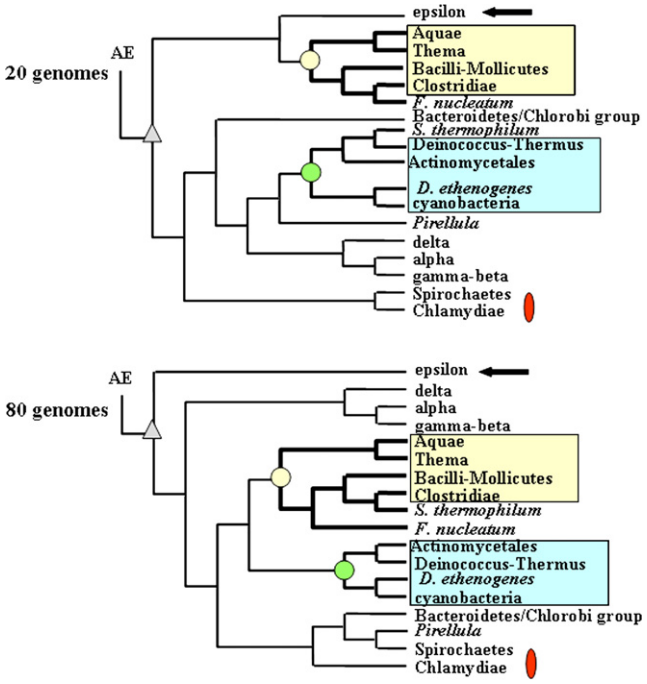


Fig. 2. Comparing the core topologies of the Domain Bacteria in phylogenomic trees. All external nodes have been collapsed. Bacterial trees are rooted (triangle) by the outgroup AE made of archaea and eukaryotes.

Note that the clustering of *Fusobacterium* (*Fusobacteria*) within the *Firmicutes* has been recently described in the case of a specific protein [39]. The grouping of both hyperthermophilic bacteria far from the bacterial root and thus far from the hyperthermophilic archaea is clearly differentiating our genomic trees from the canonical 16S RNA tree and other phylogenomic trees [37] and fits better with the biology of these organisms [40]. It also contradicts previous hypotheses about a hyperthermophilic last common ancestor to all extant life (see Ref. [40] and references inside). Grouping of Spirochaetes and Chlamydiae, generally as a clade (oval in Fig. 2) has been occasionally observed [37]. There is a strong tendency to clustering the following branches: *Deinococcus-Thermus*, cyanobacteria and chloroflexi (*Dehalococcoides ethenogenes*) and Actinomycetales and/or *Symbiobacterium thermophilum* (open rectangle in Fig. 2) either in a monophyletic or a paraphyletic way. Remarkably, the actinobacterium *S. thermophilum* is never grouping with the other Actinobacteria (which are all belonging to the order of Actinomycetales).

5. Refining homology detection to study gene evolution

5.1. Concept of module, a structural Segment of Homology (SOH)

In many cases, homology between two genes/proteins was found to be limited to long structural segments with a mean size of 220 amino acids [41]. Such segments were called modules to reflect they play a role in the mechanism of combinatorial construction of a prokaryotic gene from ready-made basic components [41]. In support of this model, it is striking that, for many prokaryotic genomes, the mean size of proteins with (known) homologues (∼450 residues) is about twice the size of proteins without any (known) homologues (∼250 residues), this last size being close to the module size. It may be more than a coincidence that this mean size of 220 amino acids we found for prokaryotic modules is rather close to typical fold (=structural domain) size, 150 ± 50 determined by a completely different approach ([42]; see also Ref. [43]). Such a figure appears as a universal unit, supporting our model of combinatorial construction of a protein from ready-made basic components, the modules.

Furthermore, we must insist that these SOH that we call modules are conceptually different from shorter segments of similarity which have been registered as domains or motifs in various specialised databases [44–47]. Domains and motifs are at a different level, being entities such as binding sites for cofactors and prosthetic groups. Domains/motifs are common features of many proteins that otherwise have no sequence similarity over a significant fraction of their total length. Although domains/motifs are important elements of the specificity of binding and the chemistry of action of a protein, as well as important features of any tertiary structure, it may be misleading to use their sequences to attempt to trace back protein history. Similar domains/motifs are found within proteins and modules that as a whole are not members of evolutionarily defined families. This is the case for example of the so-called ''$p21^{ras}$ family'' of GTP-binding proteins [48]. Besides sharing the ''P-loop'' functional motif, the majority of these proteins are not homologues. For example, we have shown that the adenylosuccinate synthase is clearly unrelated to the ras protein or the elongation factor TU from *E. coli* [49]. It has been suggested that the structural correspondence between the adenylosuccinate synthase and the ras protein at the level of their P-loop is a consequence of convergent evolution of two distinct families of proteins which bind and hydrolyze GTP [50,51]. Therefore, the SOH approach helps to differentiate between homology (*divergent* evolution) and analogy (*convergent* evolution).

5.2. Finding out all segments of homology in a proteome

One of our main objectives of the last decade has been and continues to be the study of the evolutionary mechanisms which were used to build up the present-day proteins (and their encoding genes). The module (SOH) concept we have presented in detail above is crucial to understanding protein history by taking into account two major mechanisms occurring at the gene level: duplication and fusion. As a matter of fact, comparison of modern-day proteins and identification of all modules will help to number the events of duplication and fusion, and thus, after grouping all homologous modules in families, to trace back to the ancestral genes which were at the origin of each family.

We are constantly updating an exhaustive comparison of the whole sets of proteins (hereafter called proteomes) encoded by completely sequenced genomes of microbial species. Such an exhaustive approach allows collecting – in one step – all paralogous (intragenomic comparison) and all orthologous (intergenomic comparison) pairs of aligned protein sequences. To retrieve the whole set of SOHs we adapted, as already described [41,52], the Darwin *AllAll* program [33,34] which is based on a maximum likelihood approach. We found that this *AllAll* program is very powerful in detecting all SOHs of interest including distant homologues. We have built a suite of automatic programs [52] in order to cope with the present deluge of data released by the whole-genome sequencing projects. (i) The AllAll program [33,34] detects SOHs using thresholds for evolutionary distance (less than 250 PAM units) and alignment length (at least 80 residues). (ii) Another program classifies these SOHs according to their length and location inside the aligned proteins. For example, an alignment between the first third of protein A and the last third of protein B will be interpreted as a module A_1_3 matching with a module B_3_3. (iii) We gather automatically into one family all SOHs that are related by a chain of similarities, collecting all relatives of both members of each *AllAll* pair until no further pairwise relationship is found. (iv) We further group families which are related by a chain of neighbouring unrelated SOHs. (v) Automatic analysis of these groups of families allows to split into their component parts many fused modules and/or to deduce by logic more distant modules. (vi) All detected and inferred modules are reassembled in refined

families. These two last steps (v–vi) are made by the program SortClust [52].

5.3. Grouping segments of homology in families to number their ancestral genes

5.3.1. The concept ''one family – one ancestral gene''

By definition, all SOHs belonging to the same family have a unique ancestor. Thus, counting the number of SOHs families is operationally equivalent to counting the number of ancestral genes that were at the origin of these families. To meet our different experimental needs, various kinds of simple-link families were assembled as described above, grouping either paralogues of each species, or orthologues for each pair of genomes or all homologues for different groupings of species. For instance, Fig. 3 shows 9 segments 1_1 (entire proteins) corresponding to various periplasmic binding proteins of *E. coli* that are involved in transport of amino acids from the outer to the inner membrane. This family contains also two unknown proteins, YfhD and YhdW. Since this family looks homogeneous in terms of function, one can propose that YfhD and YhdW, although they appear to be more distant, are putative periplasmic binding proteins for amino acids. Thus, the more parsimonious hypothesis is that these nine SOHs descend from an ancestral gene (located as root in the tree shown on Fig. 3) which coded a periplasmic binding protein with a broad specificity for amino acids.

5.3.2. Distinguishing biologically pertinent families in a sea of looking alike modules

It must be noted that the transitive approach we used to gather SOHs in families (see above) is very fast and rather efficient in delivering a large array of biologically pertinent families [41,52,53]. However, it has two major drawbacks. First, it often occurs that two or more groups of highly-connected proteins are linked by a few edges, often corresponding to paralogous relationships. Although many of these observed bridges are biologically meaningful, they appear detrimental when one is searching for consistent and valid families required to identify genuine ancestral genes (as detailed below). Secondly, a significant part (which may reach 60% as soon as four or more genomes are compared) of the SOHs which look alike are progressively amassed into a huge ''super-cluster''. For example, a large majority of membrane proteins are progressively and indistinctly put in the same bulky bag uniquely because, in order to cross the membrane, they have to harbour hydrophobic chains that look alike independently of the history and other specific features of these proteins. Therefore, in such a case it is hard to decide if the observed similarities are due to homology or to analogy. To cope with these two drawbacks, we have recently improved the pipeline designed to make consistent families of homologous entities. We added a graph algorithm for bridge detection to break any unwanted bridge and further used the MCL algorithm [54,55] to resolve huge and heterogeneous clusters in pertinent small families with success (Lemoine et al., in press).

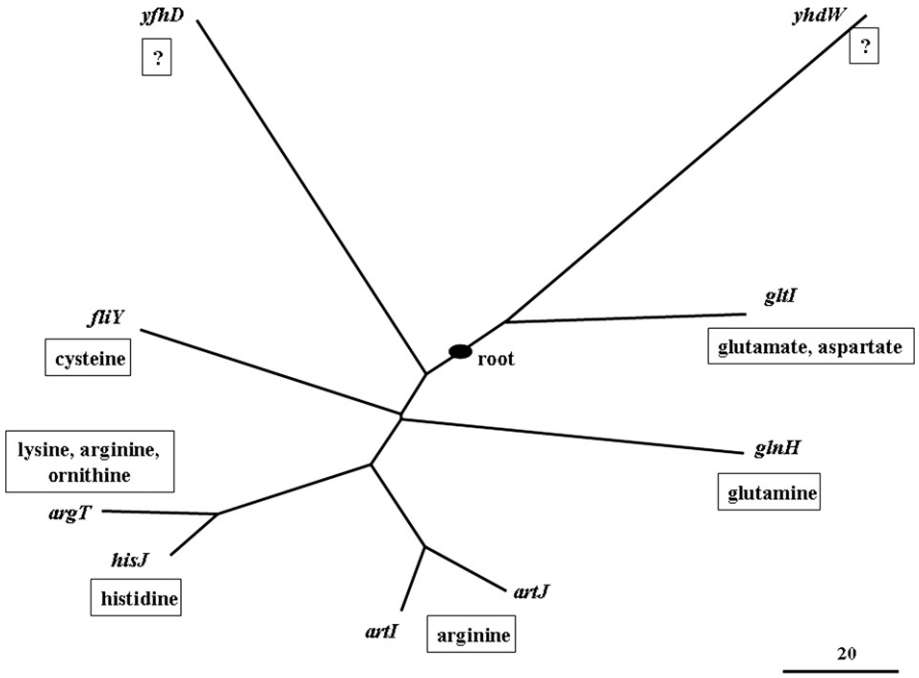


Fig. 3. Phylogenetic tree based of nine periplasmic binding proteins of *E. coli*. The PhyloTree program of the DARWIN package [33,34] was used to make an exhaustive measure of the evolutionary distances (PAM distances) separating each sequence from all its homologues, to build a multiple alignment and to derive a distance tree which is an approximation to maximum likelihood tree since the deduced evolutionary distances are weighted by computing the variance of the respective PAM distance when reconstructing the tree. The branch lengths (in PAM units, weighted by the respective variance) are drawn to scale. The black dot indicates the location of the putative root as it is computed by PhyloTree.

5.4. Applying the SOH approach to study protein history

Once the correct families have been obtained, and further refined by using the SortClust program [52] as described above, it was possible to compile, for each protein of each analyzed organism, all homology data in order to define the three following features:

- The *modular structure* of the present-day protein allows to trace back the successive events of ancestral gene duplication and fusion of evolutionarily unrelated genes which occurred at different periods and were at the origin of the present-day SOHs.
- The *phylogenetic profiles* (listing of all species containing at least one orthologue) for each SOH and for the entire protein link the protein history with the different speciation events that occurred during the evolution of the present-day species.
- The *class* of each SOH and protein is computed as previously described [52]. We defined four different *classes*: the first two categories correspond to proteins which are found in only one species (sp) and which either have a paralogue (*para-sp*) or are unique to their species (*uni-sp*). The last two categories correspond to orthologous proteins which either have a paralogue (*para-ortho*) or are unique to their species (*uni-ortho*). The distribution of these different classes display two main features as illustrated on Fig. 4A for a set of selected organisms. (1) We observed a good correlation between the proteome size and the gene content in the two orthologous classes, para-ortho and uni-ortho. The Pearson correlation values obtained are 0.87 (associated probability $P = 0.005$) and -0.92 ($P = 0.001$) respectively. (2) The distribution of these different classes appears to vary with the life style of each species. There is a clear difference between the pathogenic species (*H. influenzae, V. cholerae*) which have smaller genomes and the facultative pathogens (*E. coli*, *P. aeruginosa*) which have larger genomes. The pathogens display a low to very low proportion of *para-ortho* class while the *uni-sp* class is remarkably high. On the contrary, non-pathogenic bacteria display a large excess of *para-ortho* proteins and also a larger size for the *para-sp* class. Indeed, *E. coli* and *P. aeruginosa* maintain a strikingly high number (e.g. 411 in the case of *E. coli*) of small families (from two to six paralogous members) coding for putative transcriptional regulators, resistance to various substances (including antibiotics), or putative membrane proteins involved in various stages of transport of metallic ions and other rare environmental substances. These different functions may be important for survival of free-living bacteria in adverse conditions. The contrasted distributions observed on Fig. 4A confirm several of our precedent findings [52,56,57], which suggested that many pathogens have reduced their genome size by preferentially diminishing the size of their families of paralogous genes.

5.5. Applying the SOH approach to determine the core of ubiquitous genes in analyzed genomes

5.5.1. Comparing close and distant species

As expected, the probability to find a homologue for any gene of a genome increased as the total number of genomes to which it is compared is raised. Fig. 4B shows the distribution of the four different classes of genes for the same set of gamma proteobacterial species (*E. coli*, *H. influenzae*, *P. aeruginosa*, *V. cholerae*) but in the comparison of a larger and more diverse set of procaryotic genomes. Comparing panels A and B shows that the relative proportions of *uni-sp* and *para-sp* are decreasing and that of *uni-ortho* and *para-ortho* are increasing since the addition of more and more genomes helps to find out homologues to genes which seemed previously specific to their species. It seems that we are approaching a sort of plateau due to the presence in each genome of a significant proportion of enduring orphans [58] (genes we called *uni-sp* in our nomenclature).

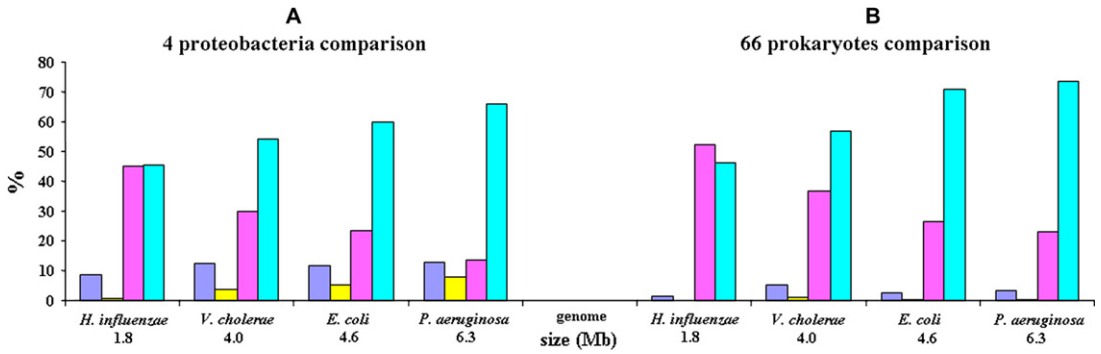


Fig. 4. Comparison of distribution (in percentage) of modules classes of homology resulting from inter and intra genomic comparisons of 4 gamma proteobacteria and inter and intra genomic comparisons of 66 microbial genomes. Results are shown for 4 selected species: *Haemophilus influenzae*, *Vibrio cholerae*, *Escherichia coli*, *Pseudomonas aeruginosa*. Defined classes of homology are: module unique to one species (uni-sp, violet), paraloguous module only found in one species (para-sp, yellow), paralogous modules found in several species (para-ortho, blue), orthologous modules (without paralogues) found in several species (uni-ortho, pink).

Table 4
Defining the core of ubiquitous genes is improved when increasing the number of compared genomes

	Four gamma proteobacteria	Fifteen gamma proteobacteria	Twenty six proteobacteria	Sixty six prokaryotes
E. coli	1834	1231	306	141
H. influenzae	1133	254	241	107
P. aeruginosa	2147	2668	367	163
V. cholerae	1532	910	333	161

5.5.2. Computing the core genome of gammaproteobacteria and reconstructing its history

Among the homologous SOHs, only a limited set of genes that are present in a genome have been found to have orthologues in all other prokaryotes. Table 4 shows that the number of these ubiquitous genes is decreasing rapidly when comparing more and more genomes which are increasingly diverse. Therefore, such universal genes would define a core of genes that are most probably essential and would be representative of the basic set of genes present in the common ancestor to all compared species. We further showed that more than 90% of these universal genes play essential roles in basic processes such as DNA replication, transcription and translation, cell wall biosynthesis and cell division, metabolism or active transport. This is strikingly reminiscent of the known universal genes which have been described by different approaches as being the core of the bacterial life (see, for instance, Ref. [59]). Thus, our approach may help to define the whole set of genes encoding essential functions including those which deserve to be further ascertained by experimental studies.

To better understand how this pool of ubiquitous genes was set up and how it has evolved we have taken as a test case the history of gammaproteobacteria. Accordingly, we have computed the relative distribution of conserved genes at the different nodes of the tree of life of this class of proteobacteria. Fig. 5 shows a maximum likelihood tree that we reconstructed using the 16S RNA sequences of the 15 compared species and for each of its nodes the number of families of SOHs that are ubiquitary in a set of species descending from a common node. For example, we have 2083 ancestral genes present in the ancestor of Enterobacteriales and 1190 ones present in the ancestor of Pseudomonadales. If we add Vibrionales and Pasteurellales, the number of ancestral genes present in the ancestor of all these families of ubiquitary genes goes down to only 493. Going deeper in the past, we found that only 281 ancestral genes are at the origin of the core genome (encoding 6327 present-day proteins) to the 15 studied gammaproteobacteria.

5.5.3. Numbering the different events of gene duplication and gene fusion

Fig. 5 further shows the distribution of the two orthologous classes of ubiquitary genes at the different nodes of the gammaproteobacteria tree. It can be seen that the proportion of para-ortho is becoming preponderant as we are going deeper

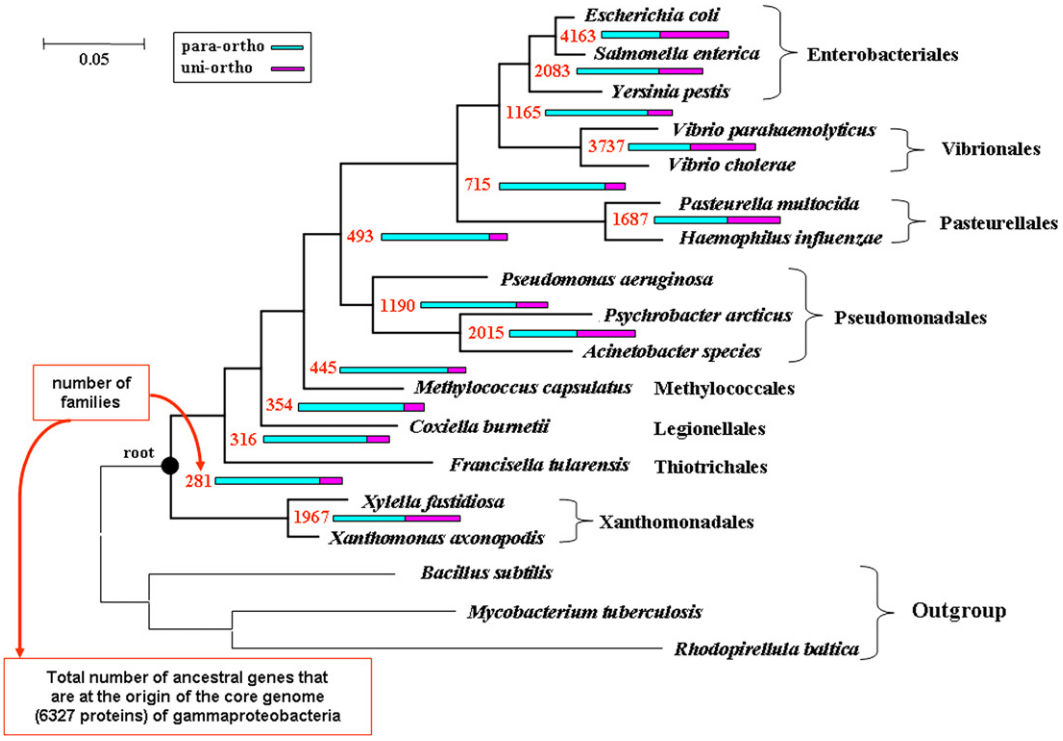


Fig. 5. Recent and ancient ancestral genes in the gammaproteobacteria tree. The 16S RNA sequences from the 15 compared gammaproteobacteria and three outgroup species were used to reconstruct a PhyML [67] tree. The number of families of modules and the distribution (para-ortho, blue – uni-ortho, pink) of homology classes are computed for each node of this tree, using the homology data obtained through the intergenomic comparison of the 15 gammaproteobacteria.

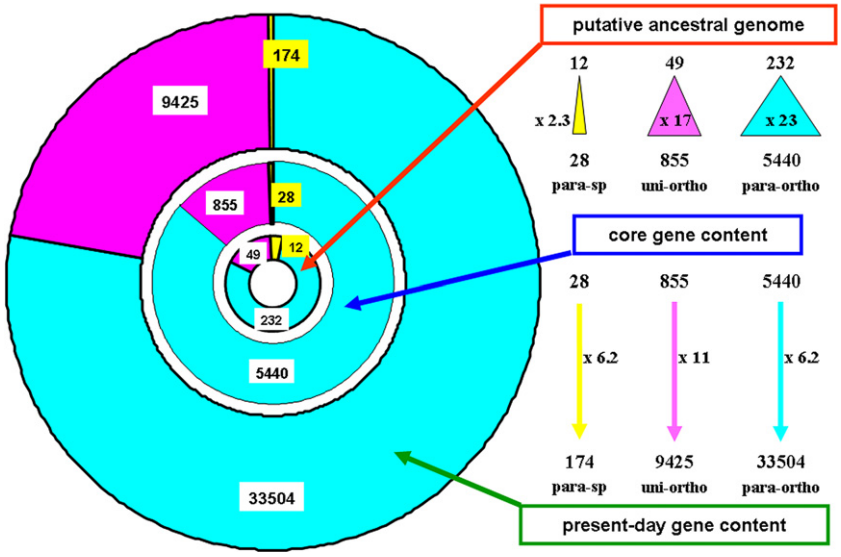


Fig. 6. From ancestral genes to present-day genomes of gammaproteobacteria. The distributions of the classes of genes are shown for the putative ancestral genome, the core genome and the present-day gene content of the 15 compared gammaproteobacteria (left part) with the same color code as in Figs. 3 and 5. On the right, the expansion rates are schematized by triangles and arrows for the two main steps of creation of multimodular proteins and their differentiation in present-day genomes respectively.

and deeper in the tree (232 para-ortho and 49 uni-ortho at the root of gammaproteobacteria, for instance). This confirms that duplication of ancestral genes has been a prevalent phenomenon in the history of prokaryotic genes.

Fig. 6 summarizes these data and shows the relative expansion rates of the three classes of para-sp, para-ortho and uni-ortho from the putative ancestral genome to the core one and then from this core genome to the full set of present-day genes found in the 15 species. It can be seen that many ancestral genes (detected as present-day modules) have fused to create more complex proteins that are found already encoded by the core genome. This first step corresponds to a rapid and large expansion rate, especially for the para-ortho class. In a second step with a lower expansion rate, the successive speciation events allow the differentiation from this ancestral core genome of the present-day multimodular genes. This suggests that many present-day proteins are the result of successive events of ancestral gene duplication and fusion of evolutionary unrelated genes which occurred at different periods and were at the origin of the present-day modules. In our eyes, the module (SOH) is a new unit of evolutionary descent. Identifying a module is operationally equivalent to determining the ancestor to this gene segment.

6. Inferring protein function from sequence homology

6.1. A large distribution of complex cases

The relationships between primary sequence, tertiary structure and function of a protein appear to be very complex when considering the large set of families we have progressively assembled after exhaustive comparison of completely sequenced genomes. We have previously described three main cases [41,56,57] that can be summarized as follows. (i) Homologous sequences code closely related functions as shown on Fig. 3. (ii) Sequences are homologous but have different functions (see a few examples in Ref. [41]). (iii) Sequences are not homologous but have related functions as exemplified on Fig. 7 that shows the genealogical tree of an eleven member family. These SOHs correspond to two isoenzymes (aconitate hydratases 1 and 2, EC 4.2.1.3), the large subunit (LeuC) of a functionally similar enzyme, the 3-isopropylmalate dehydratase (EC 4.2.1.33) and one unknown open reading frame (YbhJ), respectively. Surprisingly, the two isoenzymes are very distant and their similarities of sequence are too low for our criterions of homology. They are found belonging to the same group only because both aconitate hydratases display significant sequence similarities to LeuC and YbhJ. The common SOH of this family is about the size of the entire LeuC protein. The corresponding segment is located on the N-terminal side in the aconitate hydratase 1 and the same thing is true for YbhJ, but it is located on the C-terminal side in the aconitate hydratase 2. This striking difference in location suggests, at least in this case, that SOH location has no effect on protein activity. It also confirms that the evolutionary histories of both aconitate hydratases are significantly different. The long segment present in the N-terminal side of aconitate hydratase 2 has no homologue known in another protein beside the orthologues present in species closely related. On the other hand, we have detected on the C-terminal side of aconitate hydratase 1 a short module homologue to the corresponding segment present in the second subunit (LeuD) of the 3-isopropylmalate dehydratase. Thus, it seems that during their evolutionary history the two similar enzymes aconitate hydratase 1 and 3-isopropylmalate dehydratase have followed two different paths: in one case two independent modules remain as separate subunits (encoded by two

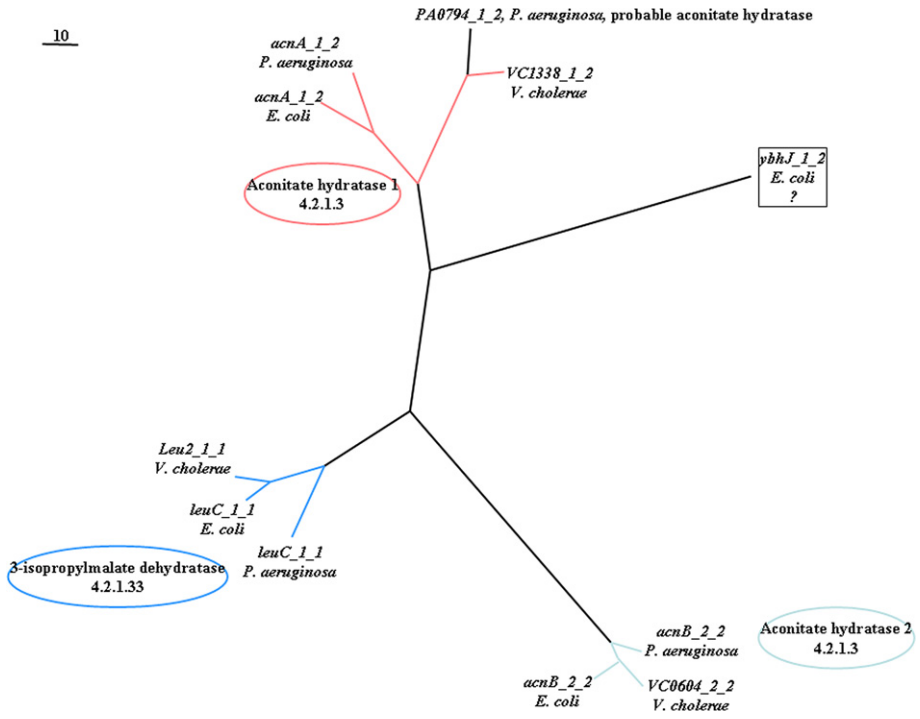


Fig. 7. Grouping SOHs which have no sequence similarity but display identical function. The tree was reconstructed as described in legend to Fig. 5.

adjacent genes inside the *leu* operon) which form a multimeric active complex. In the other case, a similar function is made by a monomer which corresponds to the fusion of the two ancestral modules. Note also that YbhJ is apparently too short to contain this entire 64 amino acid module, and this may predict absence of functionality.

6.2. SOH approach is helpful for annotating or reannotating genes in genome projects

The phylogenetic approach described above may help to transfer a function from known proteins to unknown homologues, as already shown in the case of *E. coli* periplasmic

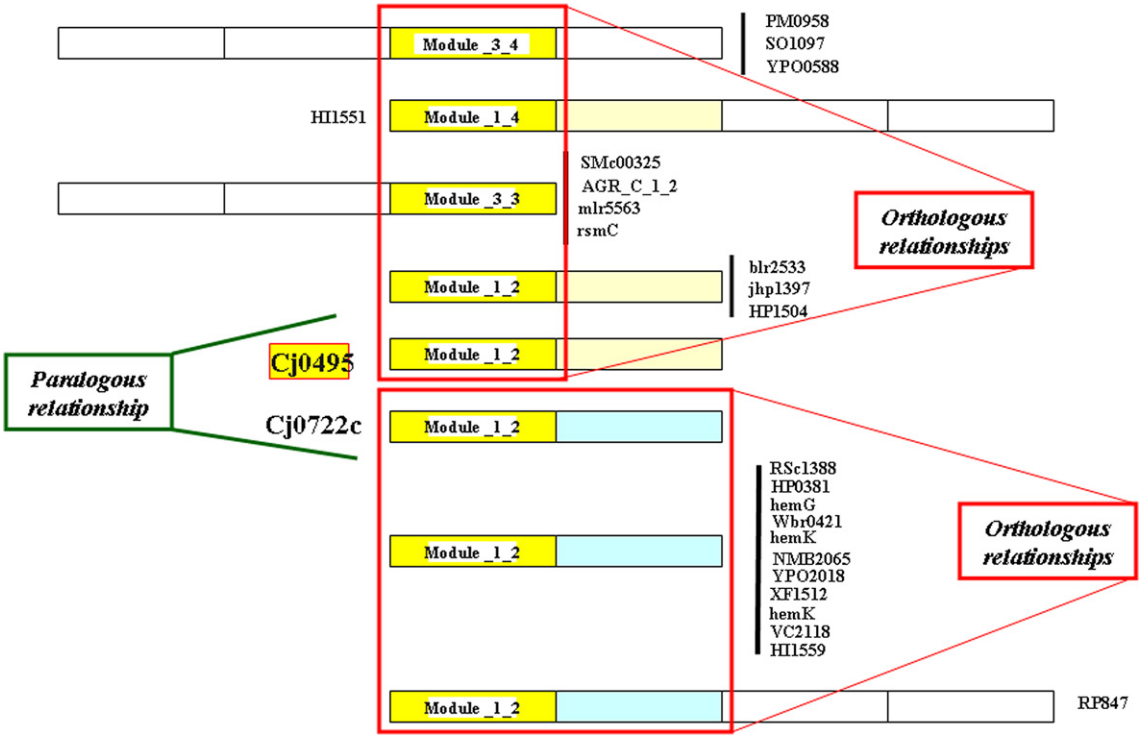


Fig. 8. Differentiating the homologous modules in a variety of multimodular proteins. The complexity of the paralogous and orthologous relationships is underlined.

binding proteins (see above and Fig. 3). To go a step further and to illustrate how helpful might be our approach in improving the annotation of essential proteins, we focused on genes that are universal to 26 proteobacteria but which remain annotated as hypothetical in *C. jejuni* [57]. Fig. 8 shows that the unknown protein Cj0495 affords complex homology relationships with various partners. We detected significant matches only between its segment 1_2 and either the segment 1_2 of its paralogue Cj0722c, or different segments of various proteobacterial proteins. Moreover, the Cj0722c protein is aligning along its full length with the putative HemK found in many proteobacteria. To better understand the complex relationships occurring between Cj0495 and its different orthologous and paralogous relatives, we aligned its segment 1_2 with the homologous segments (all the yellow segments in Fig. 8) in order to reconstruct an evolutionary tree. Fig. 9 shows that this tree is made of two, well separated, subtrees. In the subtree located in the upper part and made essentially of the HemK family, we found Cj0722c branching on a node common with the *H. pylori* HemK. These HemK proteins were annotated as DNA adenine methyltransferases. However, their function has been recently experimentally determined as N5 glutamine methyltransferase that acts on protein chain release factors PrfA and PrfB [60,61]. In the subtree located in the lower part, we find Cj0495 branching on a node common with its two unknown *H. pylori* orthologues. This lower subtree contains a set of proteins which were annotated as either hypothetical or predicted methyltransferases. Note that the three ribosomal RNA small subunit methyltransferases C (*rsmC*) present in *Xanca*, *Meslo*, *Agrtu* form a clade, strongly suggesting that the *S. meliloti* SMc00325 is also a rRNA small subunit methyltransferase C. Thus, Fig. 9 shows a clear evolutionary separation between the two paralogues Cj0722c that is most probably a N5 glutamine methyltransferase and Cj0495 which could be a nucleic acid (RNA?) methyltransferase.

7. Conclusion

Near 600 completely sequenced genomes of prokaryotes have been published in September 2007, and we are expecting very large figures in the next few years [62]. A pertinent treatment of this deluge of ongoing data requires — more than ever — a correct use of the homology approaches we have described in this review.

Our approach based on a very detailed analysis of homology relationships of modular proteins appears to be very useful as exemplified above and in our recent work on several genes involved in the biosynthesis of arginine [63,64].

A few crucial issues remain to be resolved. As already underlined, the main one is to set up *the* secure method allowing finding easily and indiscutably the genuine orthologues when comparing any set of species. Resolving this point will allow major development in understanding gene and genome evolutionary mechanisms and in genome annotation. For this last point, it would be advisable to combine contextual information (i.e. gene-neighbourhood) with protein sequence information to distinguish between the different putative co-orthologues detected in a genome as we described above.

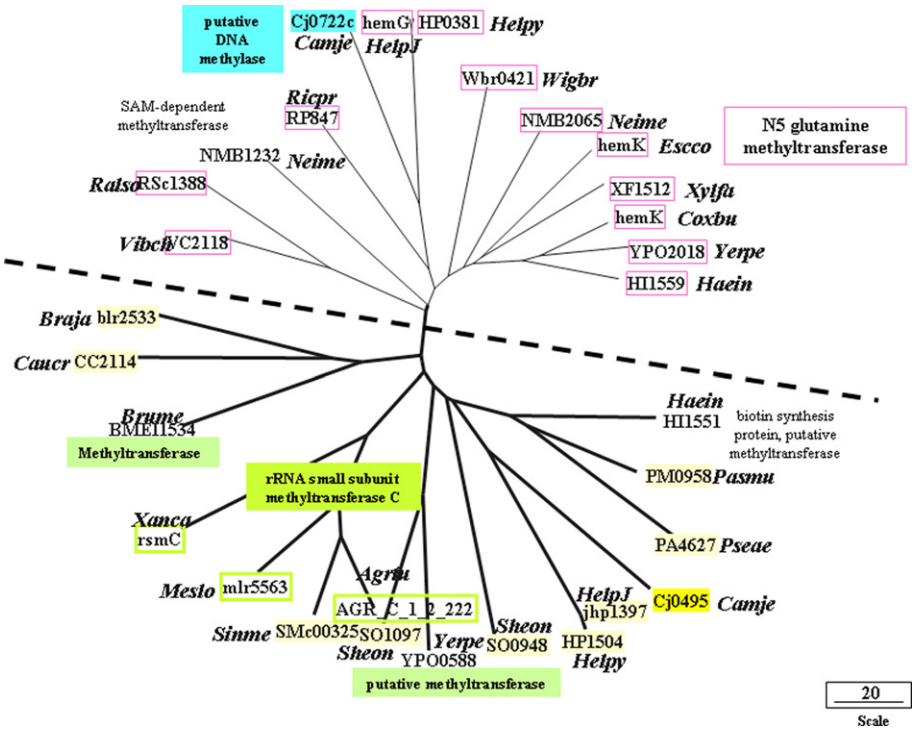


Fig. 9. Phylogenetic tree of the homologous modules identified in Fig. 8. The tree was reconstructed as described in legend to Fig. 3.

Such advances will be fundamental for the development of new fields such as systems biology [65] and synthetic biology [66]. This is equally important to progress in applied fields such as biotechnology, agronomy, medicine and pharmacology.

Acknowledgements

The undergraduate students Adrien BAZUREAU, Laurent FANT, Aurélie LEDUC, and Sylvain MAURER have been very helpful in participating to various aspects of the work described in this review during the last three years. We thank the four anonymous reviewers for their constructive and helpful comments on this manuscript.

Our lab is funded by the CNRS (UMR 8621), the PPF *Bioinformatique et Génomique* (Université Paris-Sud) and the Agence Nationale de la Recherche (ANR-05-MMSA-0009 MDMS_NV_10).

Supplementary data

Supplementary information associated with this article can be found in the online version, at doi:10.1016/j.biochi.2007.09.010.

References

[1] R. Owen, On the Archetype and Homologies of the Vertebrate Skeleton, J. van Voorst, London (1847).
[2] A.S. Wilkins, Homology, BioEssays 20 (1998) 1052–1053.
[3] W.M. Fitch, Homology: a personal view on some of the problems, Trends Genet. 16 (2001) 227–231.
[4] G.R. Reeck, C. de Haën, D.C. Teller, R.F. Doolittle, W.M. Fitch, R.E. Dickerson, P. Chambon, A.D. McLachlan, E. Margoliash, T.H. Jukes, et al., "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it, Cell 50 (1987) 667.
[5] J.W. Thornton, R. DeSalle, Gene family evolution and homology: genomics meets phylogenetics, Annu. Rev. Genomics Hum. Genet. 1 (2000) 41–73.
[6] M.C.C. De Pinna, Concepts and tests of homology in the cladistic paradigm, Cladistics 7 (1991) 367–394.
[7] S.F. Altschul, Amino acid substitution matrices from an information theoretic perspective, J. Mol. Biol. 219 (1991) 555–565.
[8] R.F. Doolittle, Similar amino acid sequences: chance or common ancestry? Science 214 (1981) 149–159.
[9] C. Patterson, Homology in classical and molecular biology, Mol. Biol. Evol. 5 (1988) 603–625.
[10] W.M. Fitch, Distinguishing homologous from analogous proteins, Syst. Zool. 19 (1970) 99–113.
[11] R.A. Jensen, Orthologs and paralogs – we need to get it right, Genome Biol. 2 (2001) 1002.
[12] W.F. Doolittle, Phylogenetic classification and the universal tree, Science 284 (1999) 2124–2129.
[13] M. Remm, C.E. Storm, E.L. Sonnhammer, Automatic clustering of orthologs and in-paralogs from pairwise species comparisons, J. Mol. Biol. 314 (2001) 1041–1052.
[14] R.L. Tatusov, E.V. Koonin, D.J. Lipman, A genomic perspective on protein families, Science 278 (1997) 631–637.
[15] B. Snel, P. Bork, M.A. Huynen, Genomes in flux: the evolution of archaeal and proteobacterial gene content, Genome Res. 1 (2002) 17–25.
[16] L.B. Koski, R.A. Morton, G.B. Golding, Codon bias and base composition are poor indicators of horizontally transferred genes, Mol. Biol. Evol. 18 (2001) 404–412.
[17] F. Swidan, E.P. Rocha, M. Shmoish, R.Y. Pinter, An integrative method for accurate comparative genome mapping, PLoS Comput. Biol. 2 (2006) e75.
[18] M. Huynen, B. Snel, W. Lathe, P. Bork, Predicting protein function by genomic context, quantitative evaluation and qualitative inferences, Genome Res. 10 (2000) 1204–1210.
[19] Y.I. Wolf, I.B. Rogozin, A.S. Kondrashov, E.V. Koonin, Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context, Genome Res. 11 (2001) 356–372.
[20] T. Dandekar, B. Snel, M. Huynen, P. Bork, Conservation of gene order: a fingerprint of proteins that physically interact, Trends Biochem. Sci. 23 (1998) 324–328.
[21] A. Enright, I. Ilipoulos, N. Kyrpides, C. Ouzounis, Protein interaction maps for complete genomes based on gene fusion events, Nature 402 (1999) 86–90.
[22] M.A. Huynen, P. Bork, Measuring genome evolution, Proc. Natl. Acad. Sci. U.S.A. 95 (1998) 5849–5856.
[23] E.M. Marcotte, M. Pellegrini, H. Ng, W.D. Rice, T.O. Yeates, D. Eisenberg, Detecting protein function and protein-protein interactions from genome sequences, Science 285 (1999) 751–753.
[24] R. Overbeek, M. Fonstein, M. D'Souza, G.D. Pusch, N. Maltsev, The use of gene clusters to infer functional coupling, Proc. Natl. Acad. Sci. U.S.A. 96 (1999) 2896–2901.
[25] M. Pellegrini, E.M.J. Marcotte, M. Thompson, D. Eisenberg, T.O. Yeats, Assigning protein functions by comparative genome analysis: protein phylogenetic profiles, Proc. Natl. Acad. Sci. U.S.A. 96 (1999) 4285–4288.
[26] M.Y. Galperin, E.V. Koonin, Who's your neighbor? New computational approaches for functional genomics, Nat. Biotechnol. 18 (2000) 609–613.
[27] A.R. Mushegian, E.V. Koonin, A minimal gene set for cellular life derived by comparison of complete bacterial genomes, Proc. Natl. Acad. Sci. U.S.A. 93 (1996) 10268–10273.
[28] L.B. Koski, G.B. Golding, The closest BLAST hit is often not the nearest neighbor, J. Mol. Evol. 52 (2001) 540–542.
[29] D.P. Wall, H.B. Fraser, A.E. Hirsh, Detecting putative orthologs, Bioinformatics 19 (2003) 1710–1711.
[30] F. Mao, Z. Su, V. Olman, P. Dam, Z. Liu, Y. Xu, Mapping of orthologous genes in the context of biological pathways, An application of integer programming, Proc. Natl. Acad. Sci. U.S.A. 103 (2006) 129–134.
[31] D.L. Fulton, Y.Y. Li, M.R. Laird, B.G. Horsman, F.M. Roche, F.S. Brinkman, Improving the specificity of high-throughput ortholog prediction, BMC Bioinformatics 7 (2006) 270.
[32] T.F. Deluca, I.H. Wu, J. Pu, T. Monaghan, L. Peshkin, S. Singh, D.P. Wall, Roundup: a multi-genome repository of orthologs and evolutionary distances, Bioinformatics 22 (2006) 2044–2046.
[33] G.H. Gonnet, M.A. Cohen, S.A. Benner, Exhaustive matching of the entire protein sequence database, Science 256 (1992) 1443–1445.
[34] G.H. Gonnet, M.T. Hallett, C. Korostensky, L. Bernardin, Darwin v. 2.0, an interpreted computer language for the biosciences, Bioinformatics 16 (2000) 101–103.
[35] C.R. Woese, Bacterial evolution, Microbiol. Rev. 51 (1987) 221–271.
[36] C.R. Woese, O. Kandler, M.L. Wheelis, Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya, Proc. Natl. Acad. Sci. U.S.A. 87 (1990) 4576–4579.
[37] Y. Wolf, I.B. Rogozin, N.V. Grishin, E.V. Koonin, Genome trees and the tree of life, Trends Genet. 18 (2002) 472–479.
[38] M.L. Saitou, M. Nei, The neighbour-joining method: a new method for reconstructing phylogenetic trees, Mol. Biol. Evol. 4 (1987) 406–425.
[39] M. Wolf, T. Müller, T. Dandekar, J.D. Pollack, Phylogeny of Firmicutes with special reference to Mycoplasma (Mollicutes) as inferred from phosphoglycerate kinase amino acid sequence data, Int. J. Syst. Evol. Microbiol. 54 (2004) 871–875.
[40] Y. Xu, N. Glansdorff, Lessons from extremophiles: early evolution and border conditions of life, in: C. Gerday, N. Glansdorff (Eds.), Physiology and Biochemistry of Extremophiles, ASM Press, Washington, 2007, pp. 409–421 DC 20036-2904.

[41] M. Riley, B. Labedan, Protein evolution viewed through *Escherichia coli* protein sequences: introducing the notion of a structural segment of homology, the module, J. Mol. Biol. 268 (1997) 857−868.
[42] S.J. Wheelan, A. Marchler-Bauer, S.H. Bryant, Domain size distributions can predict domain boundaries, Bioinformatics 16 (2000) 613−618.
[43] M. Gerstein, How representative are the known structures of the proteins in a complete genome? A comprehensive structural census, Fold. Des. 3 (1998) 497−512.
[44] R.D. Finn, J. Mistry, B. Schuster-Böckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S.R. Eddy, E.L.L. Sonnhammer, A. Bateman, Pfam, clans, web tools and services, Nucleic Acids Res. 34 (2006) D247−D251.
[45] N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, E. De Castro, P.S. Langendijk-Genevaux, M. Pagni, C.J.A. Sigrist, The PROSITE database, Nucleic Acids Res. 34 (2006) D227−D230.
[46] I. Letunic, R.R. Copley, S. Schmidt, F.D. Ciccarelli, T. Doerks, J. Schultz, C.P. Ponting, P. Bork, SMART 4.0: towards genomic data integration, Nucleic Acids Res. 32 (2004) D142−D144.
[47] C. Bru, E. Courcelle, S. Carrère, Y. Beausse, S. Dalmar, D. Kahn, The ProDom database of protein domain families: more emphasis on 3D, Nucleic Acids Res. 33 (2005) D212−D215.
[48] G.E. Schulz, Binding of nucleotides by proteins, Curr. Opin. Struct. Biol. 2 (1992) 61−67.
[49] A. Bouyoub, G. Barbier, P. Forterre, B. Labedan, The adenylosuccinate synthetase from the hyperthermophilic archaeon *Pyrococcus species* display unusual structural features, J. Mol. Biol. 261 (1996) 144−154.
[50] B.W. Poland, M.M. Silva, M.A. Serra, Y. Cho, K.H. Kim, E.M. Harris, R.B. Honzatko, Crystal structure of adenylosuccinate synthetase from *Escherichia coli*. Evidence for convergent evolution of GTP-binding domains, J. Biol. Chem. 268 (1993) 25334−25342.
[51] B.W. Poland, C. Bruns, H.J. Fromm, R.B. Honzatko, Entrapment of 6-thiophosphoryl-IMP in the active site of crystalline adenylosuccinate synthetase from *Escherichia coli*, J. Biol. Chem. 272 (1997) 15200−15205.
[52] S. Le Bouder-Langevin, I. Capron-Montaland, R. De Rosa, B. Labedan, A strategy to retrieve the whole set of protein modules in microbial proteomes, Genome Res. 12 (2002) 1961−1973.
[53] P. Liang, B. Labedan, M. Riley, Physiological genomics of *Escherichia coli* protein families, Physiol. Genomics. 9 (2002) 15−26.
[54] S. van Dongen, Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht, May 2000.
[55] A.J. Enright, S. Van Dongen, C.A. Ouzounis, An efficient algorithm for large-scale detection of protein families, Nucleic Acids Res. 30 (2002) 1575−1584.
[56] R. De Rosa, B. Labedan, The evolutionary relationships between the two bacteria *Escherichia coli* and *Haemophilus influenzae* and their putative last common ancestor, Mol. Biol. Evol. 15 (1998) 17−27.
[57] Q. Sculo, O. Lespinet, B. Labedan, Retrieving the whole set of protein modules of *Campylobacter jejeuni* and *Helicobacter pylori*, Genome Lett. 2 (2003) 2−9.
[58] G.A. Wilson, N. Bertrand, Y. Patel, J.B. Hughes, E.J. Feil, D. Field, Orphans as taxonomically restricted and ecologically important genes, Microbiology 151 (2005) 2499−2501.
[59] K. Kobayashi, S.D. Ehrlich, A. Albertini, G. Amati, K.K. Andersen, M. Arnaud, K. Asai, S. Ashikaga, S. Aymerich, P. Bessieres, F. Boland, S.C. Brignell, S. Bron, K. Bunai, J. Chapuis, L.C. Christiansen, A. Danchin, M. Débarbouillé, E. Dervyn, E. Deuerling, et al., Essential *Bacillus subtilis* genes, Proc. Natl. Acad. Sci. U.S.A. 100 (2003) 4678−4683.
[60] V. Heurgue-Hamard, S. Champ, A. Engstrom, M. Ehrenberg, R.H. Buckingham, The *hemK* gene in *Escherichia coli* encodes the N(5)-glutamine methyltransferase that modifies peptide release factors, EMBO J 21 (2002) 769−778.
[61] K. Nakahigashi, N. Kubo, S. Narita, T. Shimaoka, S. Goto, T. Oshima, H. Mori, M. Maeda, C. Wada, H. Inokuchi, HemK, a class of protein methyl transferase with similarity to DNA methyl transferases, methylates polypeptide chain release factors, and *hemK* knockout induces defects in translational termination, Proc. Natl. Acad. Sci. U.S.A. 99 (2002) 1473−1478.
[62] R. Overbeek, T. Begley, R.M. Butler, J.V. Choudhuri, H.Y. Chuang, M. Cohoon, V. de Crecy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E.D. Frank, S. Gerdes, E.M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A.C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G.D. Pusch, D.A. Rodionov, C. Ruckert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko, V. Vonstein, The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes, Nucleic Acids Res. 33 (2005) 5691−5702.
[63] D.G. Naumoff, Y. Xu, N. Glansdorff, B. Labedan, Retrieving sequences of enzymes experimentally characterized but erroneously annotated: the case of the putrescine carbamoyltransferase, BMC Genomics 5 (2004) 52.
[64] Y. Xu, B. Labedan, N. Glansdorff, Surprising arginine biosynthesis: a reappraisal of the enzymology and evolution of the pathway in microorganisms, Microbiol. Mol. Biol. Rev. 71 (2007) 36−47.
[65] B. Palsson, Systems Biology − Properties of Reconstructed Networks, Cambridge University Press, 2006.
[66] E. Andrianantoandro, S. Basu, D.K. Karig, R. Weiss, Synthetic biology: new engineering rules for an emerging discipline, Mol. Syst. Biol. 2 (2006) 2006.0028.
[67] S. Guindon, O. Gascuel, A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, Syst. Biol. 52 (2003) 696−704.

BMC Bioinformatics

BioMed Central

Software

Open Access

SynteBase/SynteView: a tool to visualize gene order conservation in prokaryotic genomes

Frédéric Lemoine[1,2], Bernard Labedan*[1] and Olivier Lespinet[1]

Address: [1]Institut de Génétique et Microbiologie, Université Paris Sud XI, CNRS UMR 8621, Bât. 400, 91405 Orsay Cedex, France and [2]Laboratoire de Recherche en Informatique, Université Paris Sud XI, CNRS UMR 8623, Bât. 490, 91405 Orsay Cedex, France

E-mail: Frédéric Lemoine - frederic.lemoine@igmors.u-psud.fr; Bernard Labedan* - bernard.labedan@igmors.u-psud.fr; Olivier Lespinet - olivier.lespinet@igmors.u-psud.fr;
*Corresponding author

Published: 16 December 2008

BMC Bioinformatics 2008, **9**:536 doi: 10.1186/1471-2105-9-536

Received: 4 September 2008

Accepted: 16 December 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/536

Abstract

Background: It has been repeatedly observed that gene order is rapidly lost in prokaryotic genomes. However, persistent synteny blocks are found when comparing more or less distant species. These genes that remain consistently adjacent are appealing candidates for the study of genome evolution and a more accurate definition of their functional role. Such studies require visualizing conserved synteny blocks in a large number of genomes at all taxonomic distances.

Results: After comparing nearly 600 completely sequenced genomes encompassing the whole prokaryotic tree of life, the computed synteny data were assembled in a relational database, SynteBase. SynteView was designed to visualize conserved synteny blocks in a large number of genomes after choosing one of them as a reference. SynteView functions with data stored either in SynteBase or in a home-made relational database of personal data. In addition, this software can compute *on-the-fly* and display the distribution of synteny blocks which are conserved in pairs of genomes. This tool has been designed to provide a wealth of information on each positional orthologous gene, to be user-friendly and customizable. It is also possible to download sequences of genes belonging to these synteny blocks for further studies. SynteView is accessible through Java Webstart at http://www.synteview.u-psud.fr.

Conclusion: SynteBase answers queries about gene order conservation and SynteView visualizes the obtained results in a flexible and powerful way which provides a comparative overview of the conserved synteny in a large number of genomes, whatever their taxonomic distances.

Background

As prokaryotic species diverge, their gene order is increasingly fading away, except in rare locations where a few genes retain their neighborhood. Such observations gave rise to the concept of genomic context [1-9]. Accordingly, it is assumed that a small number of genes remain adjacent either because their expressions occur at the same time, or because they encode proteins that are constituents of the same molecular machine (e.g. membrane ATPase) or involved in the same cellular function [10]. These genes that remain persistently adjacent in constantly moving genomes form synteny blocks. In a recent work [11], we have identified such synteny blocks in a large and diverse set of nearly 600 microbial genomes using a three-step process. In step one, we compared each protein encoded by a completely sequenced genome with all other available microbial proteomes in order to identify the full set of homologous proteins they share. In step two, we outlined

an approach allowing the identification of *bona fide* orthologues among all recognized homologues when comparing many pairs of genomes. This second step is based on an adaptation of the method designed by Wall et al. [12] to compute the reciprocal smallest distance (RSD) that separates the homologues present in a pair of genomes. Step three allowed further research among the correctly identified orthologues to pinpoint those that belong to a minimal unit that is conserved in each pair of genomes, i.e., a pair of positional orthologous genes (POGs) that remain adjacent in each genome. Then, after extending these minimal units as far as possible, it becomes feasible to assess the relative amount and size of synteny blocks in close and distant species. Such synteny blocks are appealing candidates in the study of the mechanisms of genome evolution and in the verification of the functional annotation of neighboring genes. Accordingly, visualizing these blocks in a large number of genomes at various taxonomic distances help to study their features. In this paper, we describe how to assemble all these synteny data in a relational database (SynteBase) and we develop a tool (SynteView) to visualize all conserved synteny blocks in a large number of completely sequenced prokaryotic genomes.

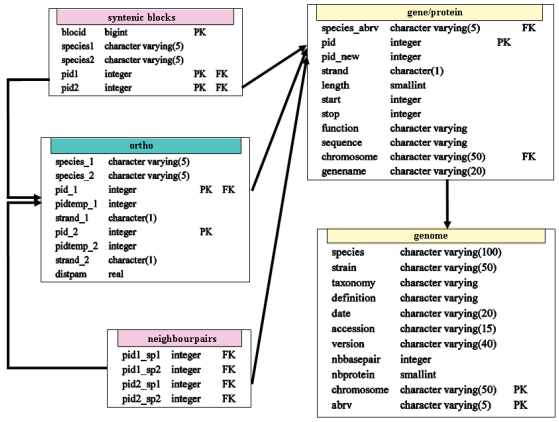


Figure 1
Relational schema of SynteBase. The SynteBase database is made up of five tables, which store information about genes/proteins, genomes, orthologous relationships, positional orthologous genes, and synteny blocks respectively. Relationships between tables are made through primary (PK) and foreign (FK) keys.

Implementation

SynteView was designed to display homology and gene context data that are organized in a relational database, SynteBase, described in detail below.

Creating a relational database for synteny data and populating its tables with a dedicated suite of softwares and other tools

We installed PostgreSQL [13], one of the most advanced open source relational database management systems, on a Linux platform and used it to create SynteBase, which is made up of five tables (Fig. 1). The database can be further populated with home-made data using the different tools we developed (see the user guide [Additional file 1]). Alternatively, one can directly use the SynteBase version we built for our own usage (this paper and [11]).

Step one: searching for homologues

Raw data extracted from public genomic databanks (GenBank/EMBL/DDBJ) were organized into two tables. The *genome* table contains information for the 598 prokaryotic genomes that were compared. The *gene/protein* table contains many features of their 1,928,135 encoded proteins, such as amino acid sequence, length, species name, location of encoding gene, etc. An exhaustive comparison of all these proteins led to the identification of all homologues. A complete suite of programs (Table 1) was used to compare each pair of proteomes using the following criteria: a pair of aligned proteins was retained as a couple of homologues if their E-values were smaller than 10^{-5}, and if the alignment extended for at least 80% of the length of the shorter matching protein [11, 14].

Table 1: A suite of programs to detect and identify synteny blocks

Step	Step designation	Tool	Reference
1	identifying homologues	protein Blast	[22]
2a	identifying orthologues by RBH	Perl script *rsd ortho*	FL, this work
2b	clustering homologues	Perl script *famtrans*	FL, this work
2c	breaking bridges	graph algorithm (Perl library)	FL, this work
2d	extracting significant clusters	MCL algorithm	[23]
3a	identifying pairs of adjacent orthologous genes	SQL query on SynteBase	FL, this work
3b	discovering synteny blocks	Perl script synblock	FL, this work

Step two: identifying orthologues among the collected homologues

We further adapted the Reciprocal Best Blast Hit approach [12] to analyze the Blast results obtained in the first step. The best RSD orthologous pairs were determined in each comparison of two proteomes as follows. Protein *a* encoded by genome G_A and protein *b* encoded by genome G_B form the best pair of orthologues if the distance separating *a* from *b* is smaller than the distance separating both *a* from any other protein encoded by G_B and *b* from any other protein encoded by G_A. We automated this search (Table 1, step 2a). The data obtained were used to populate the *ortho* table (Fig. 1).

Step three: identifying positional orthologous genes among the collected orthologues

Once populated, the first three tables were used to identify the synteny blocks. We devised a specific SQL query (see [Additional file 2]) to discover the pairs of adjacent orthologous genes (Table 1, step 3a). Then, blocks of size greater than 2 were detected by progressive accretion of blocks of size 2 which shared a common pair of orthologues (Table 1, step 3b). These computed data were entered in the *neighborpair*s and *synteny blocks* tables, respectively (Fig. 1).

Architecture of SynteView

To implement SynteView, we applied an object oriented programming paradigm using the Java programming language [15]. In this way, SynteView may be run either as a Java Webstart application or as a local application (Fig. 2). In both cases, SynteView can be used to query SynteBase through a web service (*web service mode*), or used to query a local synteny database (*loca access mode*). The web service mode allows the user to visualize the precomputed data that are present in our version of SynteBase. To do so, SynteView connects to the SynteView web service to retrieve synteny data present in SynteBase. The local access mode will be useful for those who wish to work online, with home-made computed data. This mode requires the local installation of the Data Base Management System PostgreSQL [13], and the creation of a committed SynteBase-like database that must be populated with home made synteny data after applying the following mandatory requirements to visualize these data. SynteView requires information on proteins (identifier, coding strand, sequence, function, and length), genomes (species name, species name abbreviation, strain name, taxonomy), and synteny blocks (identifier of the blocks, and pairs of identifiers of orthologous proteins belonging to this block). Note that it does not matter how the data are organized in the underlying local database. SynteView parameters can be set to retrieve the data it needs. However, while SynteView is independent of the name of the selected fields, their order is of importance for correct functioning. Components required to set up a local database are described in detail in the Additional file 1. Once the custom-made database has been built, SynteView can connect to it, after the settings, including connection information (server, login, etc) and all the mandatory queries have been filled out.

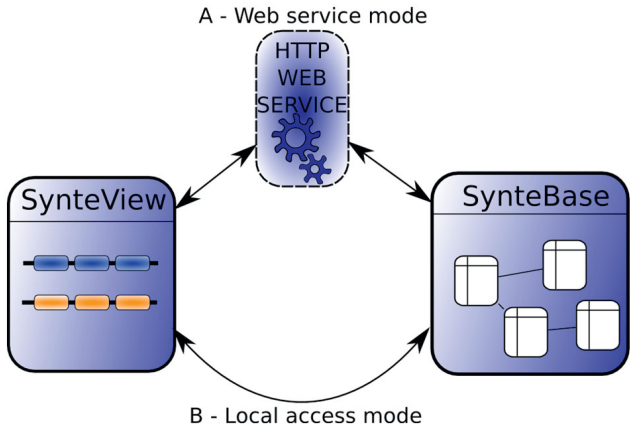


Figure 2
The two ways of using SynteView. In part A, the Web access mode connects the user to a Web service to retrieve synteny data stored in the SynteBase database http://www.synteview.u-psud.fr. In part B, the local access mode connects the user to a local database containing the home-computed synteny data to be visualized.

Results

Visualizing synteny data with SynteView

The whole set of synteny data that was stored in SynteBase was further examined using SynteView. This tool was designed to provide a wealth of information on each positional orthologous gene, to be user-friendly and customizable. For example, the user can choose the set of genomes to be studied by defining either an array of species names or a taxonomic sampling. The procedure used to visualize synteny between a reference species *s1* and a set of species (*s2*, *s3*, *s4*, *s5*) is straightforward. The user first chooses a reference species, in the "select reference genome panel" by selecting nodes in the species tree (Fig. 3). Clicking on a node produces a list of all the species that are its leaves (right panel). Then, the reference species is chosen by clicking on the species name in this list. Next, the set of compared species is determined by means of the "select compared species" tab (Fig. 3). As previously noted, the user browses the taxonomic tree of prokaryotes. When the user clicks on one node of the tree (e.g. Enterobacteriales), all the descendants of this node appear in the

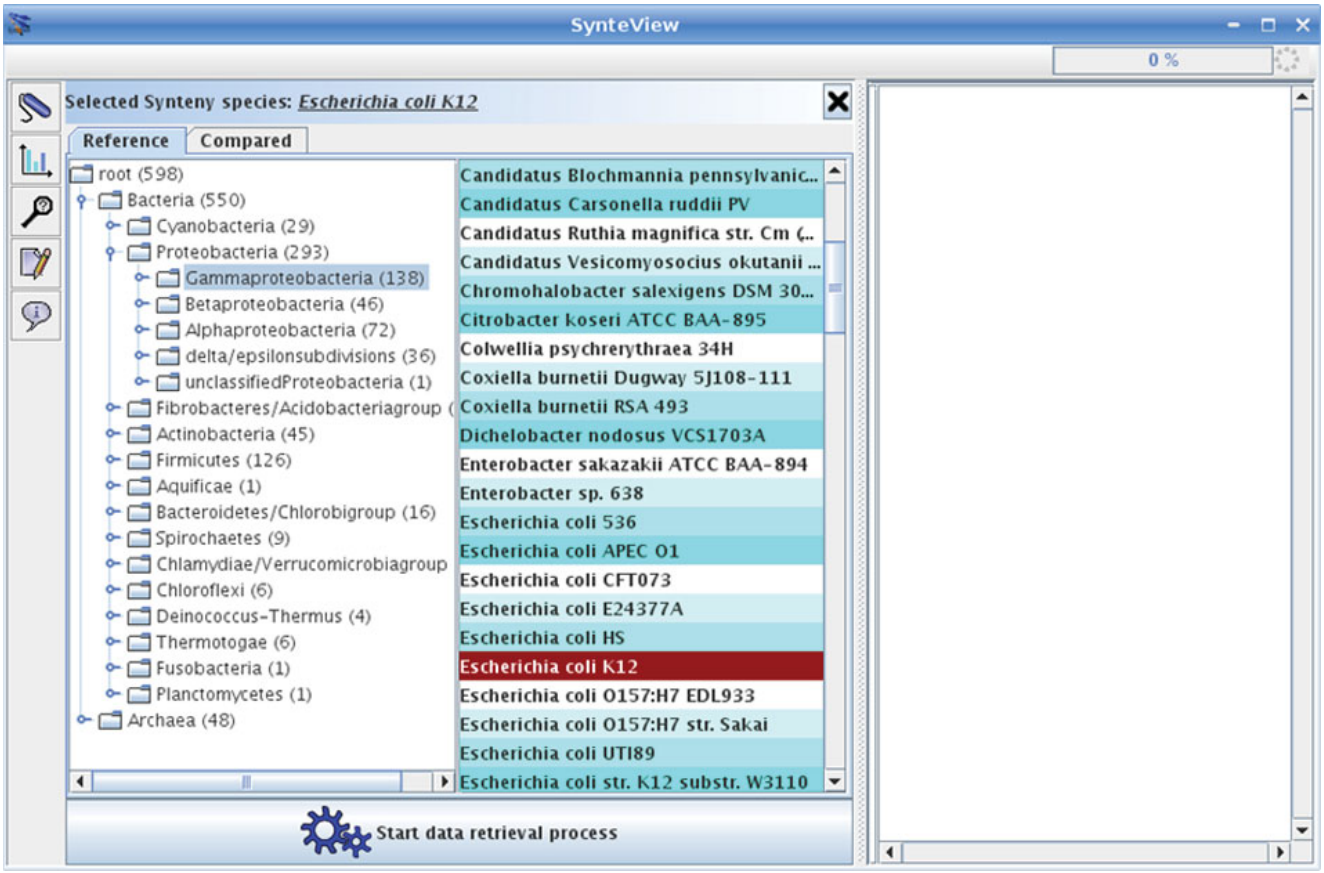


Figure 3

Selecting species to be compared. Species selection is driven by the species taxonomy. By selecting a node in the species taxonomy, the user of SynteView visualizes all the leaves descending from the selected node. It is then possible to select a species set, and finally drag and drop it into the right hand side list. The "Start data retrieval process" button starts the querying process and the visualization.

bottom panel. To choose one or several species, a drag and drop of the selected names will move the corresponding species into the right panel. This can be repeated several times, until the required set of species is selected. When this step is accomplished, clicking on the "Start data retrieval process" button on the bottom of the panel will launch the visualization step. The speed of this process depends on the number and nature of the chosen species. Once the retrieval process is completed, all regions of each compared genome become accessible for visualization in a scrollable window using the following features as shown in Fig. 4. Each line corresponds to a genome. The first line from the top (light blue background) shows gene adjacency in the chromosome of the reference species. Dark blue (positive DNA strand) and yellow (negative strand) rectangles stand for genes belonging to a synteny block that is conserved in at least one other species. Gray rectangles are genes of this reference genome that do not have any POGs in the set of compared genomes. Respective gene names are labeled on each rectangle. The following lines contain the different species that are compared to the reference genome. SynteView automatically sorts the chosen species by their taxonomic proximity to the reference genome. For each gene of the reference genome, columns contain the orthologous genes belonging to a synteny block found to be conserved in the different analyzed genomes with their respective names. The same color code (blue or yellow) helps to discriminate the strand of their respective location on each genome. The number of genes present in a block is displayed when the cursor is run over this block. Note that synteny blocks in compared genomes are defined exclusively with respect to the gene order in the reference genome. Thus, in a SynteView window of synteny blocks, the apparent proximity in compared genomes does not imply that they are as physically close in these genomes as their POGs are in the reference genome. By opening

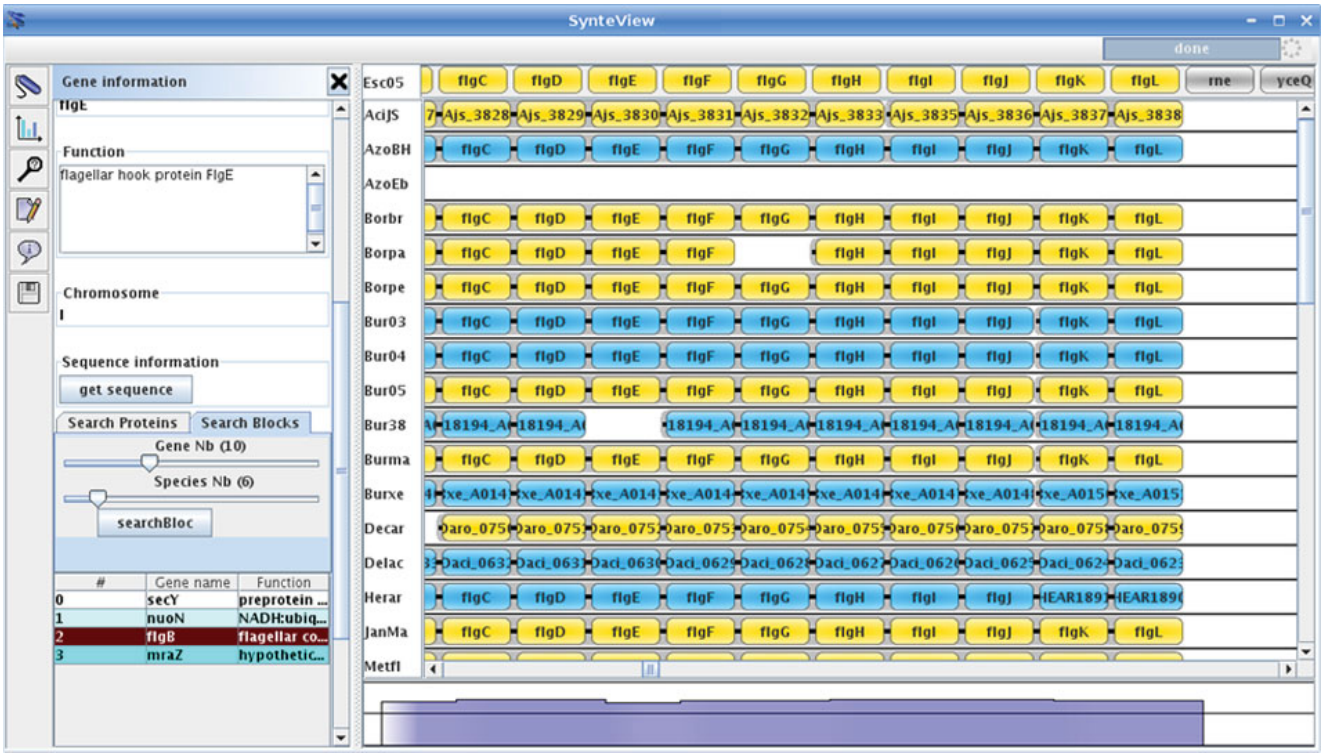


Figure 4

SynteView main window. The SynteView main window consists of a menu bar, a toolbar (on the left), a central panel which displays synteny relationships between a reference species and the compared ones, and a bottom panel, which shows the extent to which the reference species genes are conserved in blocks found in other species.

the *Setting*s panel (to do so, click on the "settings" button in the left toolbar menu) the user accesses a Dialog box where it is possible to modify various default parameters. For example, clicking on the "Database" tab allows the user to choose the retrieval mode (database or web service). Once these various parameters have been customized, it is possible to navigate along the reference genome to estimate the density of the synteny blocks present in the other genomes. For example, and as expected, comparing *E. coli* with the other gammaproteobacteria reveals a rather high density of gene conservation. The bottom blue background shape portrays this rate of conservation in the compared genomes as a histogram (Fig. 4).

Using SynteView for comparative analysis of gene context

Information about any annotated gene is immediately available by clicking on the corresponding rectangle. This opens, to the left of the window, the "gene information" panel (Fig. 4) in which, for the selected gene, its GenBank PID, its name, the species name and the replicon to which it belongs are given; the function of its product (if available), and its exact location on the chromosome are also mentioned. This information panel also contains a text field which permits simple queries such as a search for a protein function, a gene name or a PID in the analyzed genomes, as well as a search for synteny blocks containing at least x adjacent genes and having orthologous genes in at least y species. Moreover, clicking on a gene delivers complete information on its neighbors. For instance, it is possible to estimate the various levels of conservation of detected operons when comparing organisms separated by various taxonomic distances. While the operon histidine is rather well conserved in proteobacteria (Fig. 5, panel A), the neighboring clusters of genes involved in the O-specific lipopolysaccharide biosynthesis (*rfb* cluster) and the production of extracellular polysaccharide colanic acid (cluster *wca*), which are located at a short distance and on the other strand, are rapidly fragmented to a scarce number of 2–4 genes such as the *rml* genes in *Pseudomonas aeruginosa* (Fig. 5, panel B). In addition, clicking on the "get sequences" button in the information panel opens a dialog box. SynteView shows the sequence of the clicked gene in the first tab and that of its orthologues in the second tab in Fasta format. This further allows downloading of all these amino acid sequences for future work.

Figure 5

Displaying operon conservation. Part A shows that the order of genes belonging to the histidine operon is well conserved in this set of proteobacterial genomes. Part B shows the contrasting low conservation of the neighboring *rfb* cluster. Note that there is an interval of 12 *E. coli* genes (in the reference species) between gene *hisI* in panel A and gene *rfbC* in panel B.

Using SynteView for comparative analysis of multiple views

SynteView was also designed to allow complex studies by means of easy and simple operations. For example, looking at a peculiar set of species makes it possible to immediately visualize new assortments of synteny blocks. This is done simply by selecting a new reference species by clicking on a species name on the left of the display and/or by changing the list of compared genomes. Moreover, contrary to challenging tools (see Discussion below), SynteView allows global analyses of the synteny data using various points of view. Scrolling up and down the same window, one can assess the level of conservation of gene order at various taxonomic depths, the relative density of the synteny blocks along the whole genome, the relative size of the blocks, and the respective events of gene insertion/deletion in close and distant species.

Using SynteView to quantify synteny data

Besides being a visualization tool, SynteView can display various kinds of histograms which are computed *on-the-fly*. For example, the percentage of species displaying POGs in the same synteny block in the reference species is automatically computed and displayed as a histogram (blue background shape at the bottom of the main display). It is also easy to display the distribution of the size of synteny blocks which are conserved between genomes by selecting a pair of species and clicking on the *Histogra*m button in the left toolbar. This histogram may be saved for further use by selecting the "Save as" button located in the contextual menu in the window. Table 2 summarizes the data obtained when comparing the model organism *Bacillus subtilis* with various bacteria and archaea. It appears that the number of genes present in conserved synteny blocks depends on the

Table 2: Obtaining information on synteny blocks

species 1	species 2			synteny blocks	
B. subtilis[a] Taxonomy	Species name	Taxonomy	Proteome size	average size	longest size
Bacillaceae	*Oceanobacillus iheyensis*	Bacillaceae	3500	3.6	23
Firmicutes	*Shewanella oneidensis*	Proteobacteria	4471	2.4	8
Firmicutes	*Synechocystis species*	Cyanobacteria	3167	2.8	8
Firmicutes	*Mycobacterium tuberculosis*	Actinobacteria	4187	2.5	11
Bacteria	*Methanosarcina acetivorans*	Archaea	4540	2.3	7

The average size (number of genes) of synteny blocks is dependent on the taxonomic distance separating a pair of genomes that have been compared at the level of their genetic context.
[a] The size of the *B. subtilis* proteome is 4112.

phylogenetic (taxonomic) distance between species. Indeed, the mean size of synteny blocks is close to 3.3 genes when comparing two closely related bacteria such as the Bacillaceae *B. subtilis* and *Oceanobacillus iheyensis*, whereas it diminishes to nearly 2 when comparing a bacterium (*B. subtilis*) and an archaeon (*Methanosarcina acetivorans*), although these genomes encode a similar range of proteins (3000–4500). Likewise, the longest block ranges from 19 to 4 for the same species comparisons.

Discussion

SynteView was designed to allow fast and easy visualization of the conservation of gene adjacency in many genomes for which orthology and neighborhood data were computed and stocked in a dedicated relational database SynteBase. Our goal was to develop a flexible yet powerful tool to work directly with home-computed data obtained after comparing large and diverse sets of species. Indeed, our tool can be easily installed on any personal computer endowed with one of the main operating systems (Windows, Mac OS X or Linux). Moreover, SynteView can be customized in many aspects. In particular, it can be used with another, home-made, database in place of SynteBase. We observed that among the other tools to visualize synteny data [16-20] that have been designed to be locally installed, not one is adapted to the use of the abundant genomic data for prokaryotic species. Contrary to these previously published softwares [16-20], SynteView allows the user to compare the gene order in many different genomes in the same window. Finally, the strict relationship between SynteBase and SynteView allows their user to enlarge the study of gene order by means of specific queries on SynteBase. In addition to the visualization of synteny blocks, it is possible to obtain productive information through various requests such as "How many genes are involved in a neighbouring relationship, for each pair of genomes?"

Conclusion

We anticipate that we will be inundated by thousands of completely sequenced genomes in the next few years [21]. Our tool SynteBase/SynteView has been designed to support such large sets of prokaryotic data. This tool will serve to quickly evaluate the conservation of gene order in newly-published genomes as soon as they have been compared to those already analyzed.

Availability and requirements

- Project name: SynteView/SynteBase
- Project home page: http://www.synteview.u-psud.fr
- Operating System(s): Windows, Linux, MacOS X (Java web start)
- Programming Language: Java
- Other requirements: Java 1.5
- License: GNU GPL
- Any restrictions to use by non-academics: none
- Perl scripts: available on request

Abbreviations

POGs: Positional Orthologous Genes; RSD: Reciprocal Smallest Distance; SQL: Structured Query Language.

Authors' contributions

FL wrote the different programs necessary to collect all synteny data and to build up the relational database and the visualizing tool. He is responsible for the website. The three authors participated in the design of the experimental approach, the conception of the tools, and the data analysis. Together, the three authors wrote this manuscript.

Additional material

Additional file 1

SynteView user guide. *This guide helps the user in 1) Installing SynteView, 2) Using SynteView, 3) Adapting SynteBase/SynteView to his/her own purposes. It is available at http://www.synteview.u-psud.fr/documents.php*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-536-S1.pdf]

Additional file 2

SQL Query computing POGs. *This file contains the SQL query for computation of POGs using SynteBase. The main idea is to join the ortho table with itself, and to take only the tuples which form a gene quadruplet where each vertical pair is made up of orthologues and each horizontal pair consists of adjacent genes in their respective genomes.*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-536-S2.pdf]

Acknowledgements

FL is a PhD student supported by the French Ministry of Research. This work was funded by the CNRS (UMR 8621) and the Agence Nationale de la Recherche (ANR-05-MMSA-0009 MDMS NV 10). We gratefully acknowledge Stéphane Descorps-Declère for his help in designing the genome comparison pipeline and Mary Bouley (Université de Bourgogne) for her aid in improving the quality of our manuscript.

References

1. Dandekar T, Snel B, Huynen M and Bork P: **Conservation of gene order: a fingerprint of proteins that physically interact.** *Trends in Biochemical Sciences* 1998, **23:**324–328.
2. Huynen MA and Bork P: **Measuring genome evolution.** *Proc Natl Acad Sci USA* 1998, **95:**5849–5856.
3. Enright AJ, Iliopoulos I, Kyrpides NC and Ouzounis CA: **Protein interaction maps for complete genomes based on gene fusion events.** *Natur* 1999, **402:**86–90.
4. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO and Eisenberg D: **Detecting protein function and protein-protein interactions from genome sequences.** *Science* 1999, **285 (5428):**751–753.
5. Overbeek R, Fonstein M, D'Souza M, Pusch GD and Maltsev N: **The use of gene clusters to infer functional coupling.** *Proc Natl Acad Sci USA* 1999, **96:**2896–2901.
6. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D and Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc Natl Acad Sci USA* 1999, **96:**4285–4288.
7. Galperin MY and Koonin EV: **Who's your neighbor? New computational approaches for functional genomics.** *Nat Biotechnol* 2000, **18:**609–613.
8. Huynen M, Snel B, Lathe W and Bork P: **Exploitation of gene context.** *Curr Opin Struct Biol* 2000, **10:**366–370.
9. Wolf Y, Rogozin I, Kondrashov A and Koonin E: **Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context.** *Genome Research* 2001, **3:**356–372.
10. Huynen M, Snel B, Lathe W and Bork P: **Predicting Protein Function by Genomic Context: Quantitative Evaluation and Qualitative Inferences.** *Genome Research* 2000, **10:**1204–1210.
11. Lemoine F, Lespinet O and Labedan B: **Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data.** *BMC Evol Biol* 2007, **7:**237.
12. Wall D, Fraser H and Hirsh A: **Detecting putative orthologs.** *Bioinformatic* 2003, **19:**1710–1711.
13. PostgreSQL database management systems. http://www.postgresql.org/.
14. Le Bouder-Langevin S, Capron-Montaland I, De Rosa R and Labedan B: **A strategy to retrieve the whole set of protein modules in microbial proteomes.** *Genome Research* 2002, **12:**1961–1973.
15. Java Technology. http://java.sun.com/.
16. Sinha AU and Meller J: **Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms.** *BMC Bioinformatics* 2007, **8:**82.
17. Wang H, Su Y, Mackey AJ, Kraemer ET and Kissinger JC: **SynView: a GBrowse-compatible approach to visualizing comparative genome data.** *Bioinformatic* 2006, **22:**2308–2309.
18. Hunt E, Hanlon N, Leader DP, Bryce H and Dominiczak AF: **The visual language of synteny.** *OMIC* 2004, **8:**289–305.
19. Pan X, Stein L and Brendel V: **SynBrowse: a synteny browser for comparative sequence analysis.** *Bioinformatic* 2005, **21:**3461–3468.
20. Byrne KP and Wolfe KH: **The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species.** *Genome Research* 2005, **15:**1456–1461.
21. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Ruckert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O and Vonstein V: **The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes.** *Nucleic Acids Res* 2005, **33:**5691–5702.
22. protein BLAST. http://blast.ncbi.nlm.nih.gov/.
23. MCL – a cluster algorithm for graphs. http://micans.org/mcl/.

BMC Bioinformatics

BioMed Central

Database

ORENZA: a web resource for studying ORphan ENZyme activities

Olivier Lespinet and Bernard Labedan*

Address: Institut de Génétique et Microbiologie, CNRS UMR 8621, Université Paris-Sud, Bâtiment 400, 91405 Orsay Cedex, France

Email: Olivier Lespinet - olivier.lespinet@igmors.u-psud.fr; Bernard Labedan* - bernard.labedan@igmors.u-psud.fr

* Corresponding author

Published: 06 October 2006

BMC Bioinformatics 2006, **7**:436 doi:10.1186/1471-2105-7-436

Received: 25 July 2006

Accepted: 06 October 2006

This article is available from: http://www.biomedcentral.com/1471-2105/7/436

Abstract

Background: Despite the current availability of several hundreds of thousands of amino acid sequences, more than 36% of the enzyme activities (EC numbers) defined by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) are not associated with any amino acid sequence in major public databases. This wide gap separating knowledge of biochemical function and sequence information is found for nearly all classes of enzymes. Thus, there is an urgent need to explore these sequence-less EC numbers, in order to progressively close this gap.

Description: We designed ORENZA, a PostgreSQL database of ORphan ENZyme Activities, to collate information about the EC numbers defined by the NC-IUBMB with specific emphasis on orphan enzyme activities. Complete lists of all EC numbers and of orphan EC numbers are available and will be periodically updated. ORENZA allows one to browse the complete list of EC numbers or the subset associated with orphan enzymes or to query a specific EC number, an enzyme name or a species name for those interested in particular organisms. It is possible to search ORENZA for the different biochemical properties of the defined enzymes, the metabolic pathways in which they participate, the taxonomic data of the organisms whose genomes encode them, and many other features. The association of an enzyme activity with an amino acid sequence is clearly underlined, making it easy to identify at once the orphan enzyme activities. Interactive publishing of suggestions by the community would provide expert evidence for re-annotation of orphan EC numbers in public databases.

Conclusion: ORENZA is a Web resource designed to progressively bridge the unwanted gap between function (enzyme activities) and sequence (dataset present in public databases). ORENZA should increase interactions between communities of biochemists and of genomicists. This is expected to reduce the number of orphan enzyme activities by allocating gene sequences to the relevant enzymes.

Background

Since 1956, the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) has been classifying enzyme activities (EC numbers) in order to organize all contributions made by individual biochemists and to check their validity and consistency [1]. Such a standardization effort is based on the definition of the so-called EC numbers that comprise four digits. The first one (from 1 to 6) delineates the broad type of activity: Oxidoreductase, Transferase, Hydrolase,

Lyase, Isomerase, and Ligase respectively. The second and third digits detail the reaction that an enzyme catalyzes. For example (Table 1), among the 1065 items forming the class Hydrolases (EC 3), there are 163 Glycosylases forming the subclass EC 3.2, of which, 140 enzymes hydrolyse O- and S-glycosyl compounds (sub-subclass EC 3.2.1) and 23 hydrolyse N-Glycosyl compounds (sub-subclass EC 3.2.2). The last digit is a serial number that is used to identify a particular enzyme. For instance, EC 3.2.2.1 corresponds to the purine nucleosidase and EC 3.2.2.3 to the uridine nucleosidase, respectively. The EC categorization is constantly evolving as new enzyme activities are determined and new information comes to light on previously classified enzymes. Presently (June 2006), 3927 EC numbers correspond to a defined unambiguous activity encoded by a protein. Note that IntEnz, the integrated relational enzyme database [2], now provides easy access to updated and curated data of the NC-IUBMB [1].

Unexpectedly, Peter Karp [3] and us [4,5] independently observed that a significant part of these curated and approved EC numbers does not correspond to any amino acid sequence in public databases. Recent updates of our previous results confirm this very large gap between known enzyme function and recorded protein sequence. There are presently only 2483 EC numbers having at least one associated sequence in the release 8.1 (13-Jun-2006) of the UniProt Knowledgebase [6]. We have used the term orphan enzyme activities [4] for the 1444 EC numbers that do not have a sequence associated with them. Remarkably, these orphan enzyme activities currently represent 36.8% of the 3927 retained EC numbers.

We have already shown that orphans are present at about the same proportion in every class and subclass of enzyme activities [4]. Likewise, we found no correlation between orphan distribution and main functional categories. 25.3% of the enzyme activities involved in well-studied metabolic pathways are sequence-less while we found 49.5% orphans among non-metabolic enzyme activities [4].

Thus, it appears that there is an important gap between function and sequence, which implies that its progressive bridging would require a concerted effort as already underlined [3,4]. Accordingly, we have built ORENZA, a database of ORphan ENZyme Activities, to offer such a tool to the research community. Hereafter, we describe the content of this resource and we detail how to use it in order to reach the goals defined above.

Construction and content

Structure of the ORENZA database

In order to build an efficient relational database that will help to identify the encoding gene for the maximum number of sequence-less enzyme activities (the so-called orphan enzymes [4]) we have retrieved data from various public databases and we have organized them as described below.

Data collection

There are two primary sources of information about each enzyme, corresponding respectively to all data about its activity (EC number), namely the Enzyme Nomenclature [1,2] and amino acid sequences as recorded in UniProt Knowledgebase [6]. Fig. 1 shows the different fields that were collected from both these sources and how they are organized in one main table. Moreover, we added additional – but highly important – features about each enzyme such as its role in metabolism (data recovered from KEGG [7]), the names of the organism(s) where it has been studied (data extracted from BRENDA [8] and from UniProt [6]), and the taxonomy of these organisms (data extracted from the NCBI [9]), and various pieces of information extracted from ENZYME [10] such as cofactors, possible role in disease and motifs found in PROSITE [11]. These secondary characteristics are confined to small tables or added directly to the main one as in the case of the 3D structure (data recovered from PDB [12]). We wrote Perl scripts in order to extract and periodically update the relevant information from the following public resources: NC-IUBMB, ENZYME, KEGG, BRENDA, UNIPROT, and PDB. Note also the addition of a couple of other tables, one is listing ribozymes (only one, presently), the other one listing the individual contributions made by external experts on their sequence data (see below for more details).

Checking orphanity

A Perl script screened the occurrence of EC numbers in UniProt Knowledgebase [6]. Any EC number assigned by the NC-IUBMB [1] that is not referenced in UniProt is defined as an orphan enzyme activity. Note that we did not take into account partial or incomplete EC numbers

Table 1: Browsing the EC hierachy. For each level are indicated the total number of EC numbers and that of orphan EC numbers between brackets.

Class	**3** 1065 [336]												
Subclass	**3.1** 267 [113]	**3.2** 163 [56]		**3.3** 10 [4]	**3.4** 317 [49]	**3.5** 171 [70]	**3.6** 109 [36]	**3.7** 10 [4]	**3.8** 10 [1]	**3.9** 1 [1]	**3.10** 2 [1]	**3.11** 2 [0]	**3.12** 1 [1]	**3.13** 2 [0]
Sub-subclass		**3.2.1** 140 [45]	**3.2.2** 23 [11]											

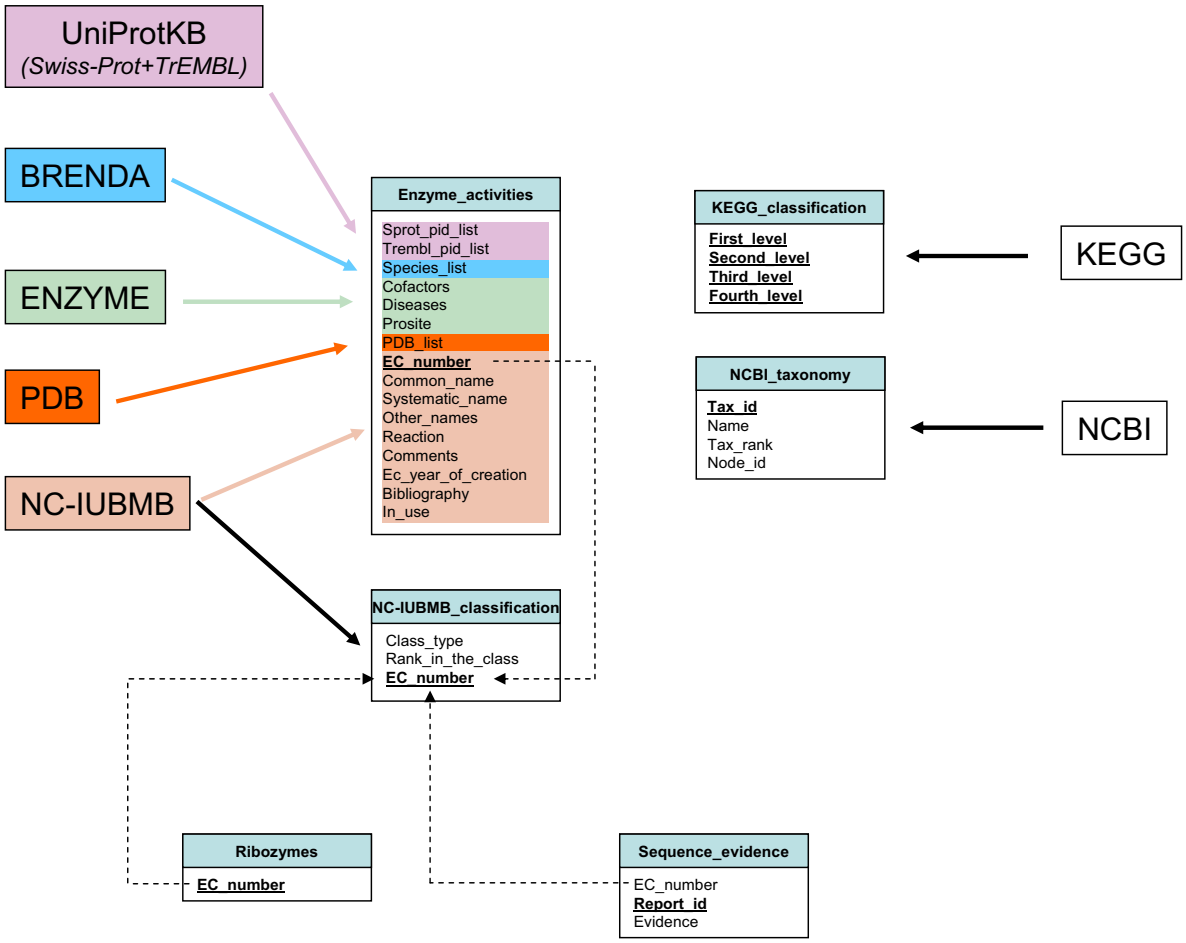


Figure 1
Schema of the ORENZA relational database. The primary key of each table is in bold underlined type. Dashed arrows indicate references to foreign keys. Plain arrows represent the origin of the data stored in each table. Moreover, for the table Enzyme_activities the origin of the data is indicated by the same color code used to identify each of the following major primary databases used in our analysis: UniProt (purple), BRENDA (blue), ENZYME (green), PDB (orange) and NC-IUBMB (beige).

(318 in the present version of UniProt) but too ambiguous [13] for sound use.

Structuring the relational database and implementing the web resource

We chose to use exclusively open source tools to build ORENZA database.

Accordingly, PostgreSQL 8.1 [14], one of the most advanced open source databases, was installed on a Linux platform. PHP language [15] was used to structure the Web service and to better exploit the queries from the relational database.

Utility

Browsing and searching ORENZA

One can browse and/or search ORENZA using three main avenues as described in detail below.

Browsing the whole set of EC numbers

The complete list of EC numbers is directly available by a simple click. It corresponds to the most recent version of NC-IUBMB [1]. The obtained view displays the list as a three-column table where each line corresponds to a specific EC number, the common name of the corresponding enzyme and a computed annotation about its possible orphanity, respectively (Fig. 2). Note also that the upper

3928 EC numbers are presently assigned by the NC-IUBMB. Among them 1444 are orphans.

EC NUMBER	COMMON NAME	ORPHAN
EC 1.1.1.1	alcohol dehydrogenase	-
EC 1.1.1.2	alcohol dehydrogenase (NADP+)	-
EC 1.1.1.3	homoserine dehydrogenase	-
EC 1.1.1.4	(R,R)-butanediol dehydrogenase	-
EC 1.1.1.5	acetoin dehydrogenase	-
EC 1.1.1.6	glycerol dehydrogenase	-
EC 1.1.1.7	propanediol-phosphate dehydrogenase	*Yes*
EC 1.1.1.8	glycerol-3-phosphate dehydrogenase (NAD+)	-
EC 1.1.1.9	D-xylulose reductase	-
EC 1.1.1.10	L-xylulose reductase	-
EC 1.1.1.11	D-arabinitol 4-dehydrogenase	-
EC 1.1.1.12	L-arabinitol 4-dehydrogenase	-
EC 1.1.1.13	L-arabinitol 2-dehydrogenase	-
EC 1.1.1.14	L-iditol 2-dehydrogenase	-
EC 1.1.1.15	D-iditol 2-dehydrogenase	-
EC 1.1.1.16	galactitol 2-dehydrogenase	*Yes*
EC 1.1.1.17	mannitol-1-phosphate 5-dehydrogenase	-
EC 1.1.1.18	inositol 2-dehydrogenase	-
EC 1.1.1.19	glucuronate reductase	-
EC 1.1.1.20	glucuronolactone reductase	-
EC 1.1.1.21	aldehyde reductase	-
EC 1.1.1.22	UDP-glucose 6-dehydrogenase	-
EC 1.1.1.23	histidinol dehydrogenase	-
EC 1.1.1.24	quinate dehydrogenase	-
EC 1.1.1.25	shikimate dehydrogenase	-

Figure 2
Extract from the full list of enzymes classified by the NC-IUBMB, along with their associated orphanity. For each line EC number, common name and orphanity are indicated. The total number of enzymes and the total number of orphan enzymes activities are indicated on top.

line of this view shows a summary indicating the total number of the EC numbers present in the selection (including the ribozyme) as well as that of the orphan EC numbers, respectively. The entire list, which can be easily downloaded as a text file, is completely dynamic. A click on a line opens a new view delivering a wealth of information about the selected EC number that is structured in three successive levels. Fig. 3A shows an example in the case of EC 1.1.1.125 with notification of many features.

The first level consists of characteristics of the enzymatic activity and its history. The description section contains information taken from the NC-IUBMB data such as the different names (common, systematic, and others) of the

A

EC 1.1.1.125

Common name :	2-deoxy-D-gluconate 3-dehydrogenase
Systematic name :	2-deoxy-D-gluconate:NAD+ 3-oxidoreductase
Other names :	2-deoxygluconate dehydrogenase
Reaction :	2-deoxy-D-gluconate + NAD+ = 3-dehydro-2-deoxy-D-gluconate + NADH + H+
References :	1. Eichhorn, M.M. and Cynkin, M.A. Microbial metabolism of 2-deoxyglucose; 2-deoxyglucose acid dehydrogenase. Biochemistry 4 (1965) 159-165.
Created :	EC 1.1.1.125 created 1972
KEGG MAP :	00040 Pentose and glucuronate interconversions
BRENDA organisms :	Aspergillus fumigatus Bacillus halodurans Bacillus subtilis --- Yersinia pseudotuberculosis
Prosite :	PDOC00060
Swiss-Prot	3 protein sequences in Swiss-Prot
	[P37769,KDUD_ECOLI_] [P50842,KDUD_BACSU_] [Q05528,KDUD_DICD3_]
TrEMBL	39 protein sequences in TrEMBL
	Q1J7L8 Q1JCS0 Q1JHT9 Q1JMP6 Q1NEJ6 Q1R7H0 Q1WRA1 Q2B7P5 Q2CAL6 Q2K140 Q2YIZ9 Q3BZC9 Q3JHK2 Q3JT32 Q4V1C0 Q5DYS9 Q5WJ75 Q5WJ78 Q5WJC3 Q62AB7 Q667B0 Q669X5 Q6D4I9 Q6MY53 Q6W2B4 Q746L6 Q82UC5 Q89VG3 Q8EMM8 Q8FE98 Q8KHI5 Q8YD61 Q8YIP8 Q8YIP9 Q8ZFH9 Q8ZHQ1 Q8ZM99 Q92V10 Q9KAW9

B

EC 1.1.1.126 is Orphan !

Common name :	2-dehydro-3-deoxy-D-gluconate 6-dehydrogenase
Systematic name :	2-dehydro-3-deoxy-D-gluconate:NADP+ 6-oxidoreductase
Other names :	2-keto-3-deoxy-D-gluconate dehydrogenase 2-keto-3-deoxygluconate dehydrogenase
Reaction :	2-dehydro-3-deoxy-D-gluconate + NADP+ = (4S,5S)-4,5-dihydroxy-2,6-dioxohexanoate + NADPH + H+
References :	1. Preiss, J. and Ashwell, G. Alginic acid metabolism in bacteria. II. The enzymatic reduction of 4-deoxy-L-erythro-5-hexoseulose uronic acid to 2-keto-3-deoxy-D-gluconic acid. J. Biol. Chem. 237 (1962) 317-321.
Created :	EC 1.1.1.126 created 1972
BRENDA organisms :	Pseudomonas sp.
Swiss-Prot	No protein sequences are associated with EC 1.1.1.126 in Swiss-Prot
TrEMBL	No protein sequences are associated with EC 1.1.1.126 in TrEMBL

Figure 3
Details of specific enzymes. 3A: example of an enzyme entry with associated amino acid sequences. 3B: example of an orphan EC number. The fact that the enzyme is an orphan enzyme is noted after the EC number and in the Swiss-Prot and TrEMBL fields.

enzyme, a scheme of the reaction(s) it catalyses and other data about the cofactors and NC-IUBMB comments about the reaction that are extracted from the ENZYME database [10]. In the history part, we list fundamental references, and the date of creation of the entry in the official NC-IUBMB nomenclature.

The second level presents information about the position of the enzyme in the cell metabolism with the corresponding number of a KEGG map [7], and its taxonomic ubiquity with a list of organisms where this enzymatic activity has been characterized as recorded in the BRENDA database [8].

The third level exhibits information about the peptidic molecule such as motifs (from PROSITE [11]), the lists of amino acid sequences found in SwissProt and TrEMBL, respectively [6]. If there is no sequence, as is the case for EC 1.1.1.126, which is labeled "orphan", this is clearly mentioned (Fig. 3B).

Browsing the orphan EC numbers

The second main avenue offered by ORENZA to explore the enzyme universe is the entire list, periodically updated, of the orphan enzyme activities. As described above, there are several ways to retrieve these orphans besides browsing the list in its entirety.

First, one can browse the different levels (class, subclass, etc.) of the EC hierarchy exactly as already described for the whole dataset of EC numbers.

A second approach is to explore the metabolism hierarchy proposed by KEGG. For instance, clicking on Lipid Metabolism (56 orphans out of 246) opens a view showing the distribution of these orphans inside the 12 corresponding pathways (Fig. 4A). Among these 12 pathways, glycerophospholipid metabolism appears to have the most orphans (19). Another click unveils the full list of these enzyme activities involved in glycerophospholipid metabolism for which no amino acid sequence is available (Fig. 4B). Again, one can explore each enzyme in detail and copy/paste the corresponding information to save it as a text file.

A third way to browse the orphan EC numbers is to sort them by their year of creation. This permits one to observe that the relative proportion of orphans is independent of the progress of genome sequencing. Fig. 5A shows that many orphans appeared during the period of gene sequencing and that the level remained unexpectedly high during the present era of heavy genome sequencing. Fig. 5B zooms in on the last seven years and confirms this trend with a high proportion of orphans in 2000, 2004 and 2005.

A fourth way to explore orphan enzyme activities is based on their occurrence in different organisms. Here, we access the entire list of orphan enzyme activities sorted by the number of organisms where these activities have been detected and experimentally studied. Beside the 39 EC numbers for which there is no information in the BRENDA resource, we find that a large majority (1286) of orphans is found in a limited number (1 to 10) of species (Fig. 6) but a few ones (132) have been found to have a large taxonomic distribution (Fig. 6, inset).

Searching ORENZA

It is possible to query ORENZA for a specific enzyme activity by entering either the EC number or the enzyme name. For example, entering the word "aspartate" recovers 41 EC numbers, 13 being presently not assigned to a sequence.

Another interesting feature is the possibility of searching by species. For instance, entering the phrase "*Homo sapiens*" retrieves 1560 EC numbers that are present in human cells. Looking at the obtained list shows again a significant number of 225 orphans. The same observation is true for four other model organisms as shown in Table 2.

Interestingly, the proportion of orphans that are common to these five model organisms is extremely low. Only three EC numbers are found as orphans in the five organisms: EC 3.6.1.18 (FAD diphosphatase), EC 3.6.4.4 (plus-end-directed kinesin ATPase), and EC 3.6.4.5 (minus-end-directed kinesin ATPase). Moreover, only three EC numbers are found as orphans in *E. coli*, fungi and animals but not in plants: EC 1.1.1.43 (phosphogluconate 2-dehydrogenase), EC 1.5.3.2 (N-methyl-L-amino-acid oxidase), and EC 3.6.4.1 (myosin ATPase).

On the other hand, we have a few species-specific orphans as shown further on Table 2. For instance, six orphan EC numbers are reported uniquely in human cells (listed in Table 3) but the corresponding figures are as high as 25 for *E. coli* and 14 for *S. cerevisiae*, two organisms that have been intensively studied at the biochemical level for 60 years by thousands of laboratories worldwide.

Building an ORENZA community

We clearly need the help of a large array of experts to identify the putative sequence(s) associated with orphan enzyme activities [3,4]. In order to encourage such a collective effort, we propose, as a part of this ORENZA resource, a friendly tool that will allow people having sound knowledge about specific enzyme activities to make helpful suggestions. Moreover, such a resource could help to establish fruitful and dynamic interactions between different experts interested in the same field. Indeed, each suggestion (with identification of its author) will appear on ORENZA resource as a new item on each

A

ORENZA 56 orphans are present in the KEGG pathway: *'01130 Lipid Metabolism'*

KEGG METABOLISM PATHWAY	ORPHANS
00061 Fatty acid biosynthesis	1
00062 Fatty acid elongation in mitochondria	1
00071 Fatty acid metabolism	6
00072 Synthesis and degradation of ketone bodies	-
00100 Biosynthesis of steroids	1
00120 Bile acid biosynthesis	6
00140 C21-Steroid hormone metabolism	2
00150 Androgen and estrogen metabolism	8
00561 Glycerolipid metabolism	8
00564 Glycerophospholipid metabolism	**19**
00590 Arachidonic acid metabolism	4
00591 Linoleic acid metabolism	1

B

ORENZA 19 orphans are present in the KEGG pathway: *'00564 Glycerophospholipid metabolism'*

EC NUMBER	COMMON NAME	ORPHAN
EC 1.14.99.19	plasmanylethanolamine desaturase	*Yes*
EC 2.3.1.23	1-acylglycerophosphocholine O-acyltransferase	*Yes*
EC 2.3.1.25	plasmalogen synthase	*Yes*
EC 2.3.1.52	2-acylglycerol-3-phosphate O-acyltransferase	*Yes*
EC 2.3.1.62	2-acylglycerophosphocholine O-acyltransferase	*Yes*
EC 2.3.1.63	1-alkylglycerophosphocholine O-acyltransferase	*Yes*
EC 2.3.1.67	1-alkylglycerophosphocholine O-acetyltransferase	*Yes*
EC 2.3.1.70	CDP-acylglycerol O-arachidonoyltransferase	*Yes*
EC 2.3.1.104	1-alkenylglycerophosphocholine O-acyltransferase	*Yes*
EC 2.3.1.105	alkylglycerophosphate 2-O-acetyltransferase	*Yes*
EC 2.3.1.121	1-alkenylglycerophosphoethanolamine O-acyltransferase	*Yes*
EC 2.3.1.149	platelet-activating factor acetyltransferase	*Yes*
EC 2.7.8.4	serine-phosphoethanolamine synthase	*Yes*
EC 2.7.8.22	1-alkenyl-2-acylglycerol choline phosphotransferase	*Yes*
EC 3.1.3.59	alkylacetylglycerophosphatase	*Yes*
EC 3.1.4.2	glycerophosphocholine phosphodiesterase	*Yes*
EC 3.3.2.2	alkenylglycerophosphocholine hydrolase	*Yes*
EC 3.3.2.5	alkenylglycerophosphoethanolamine hydrolase	*Yes*
EC 3.6.1.16	CDP-glycerol diphosphatase	*Yes*

Figure 4

List of orphan enzyme activities for various KEGG pathways. 4A: List for pathway '*01130 Lipid Metabolism*', sorted by sub-pathways. 4B: List of orphan EC numbers for pathway '*00564 Glycerophospholipid metabolism*.'

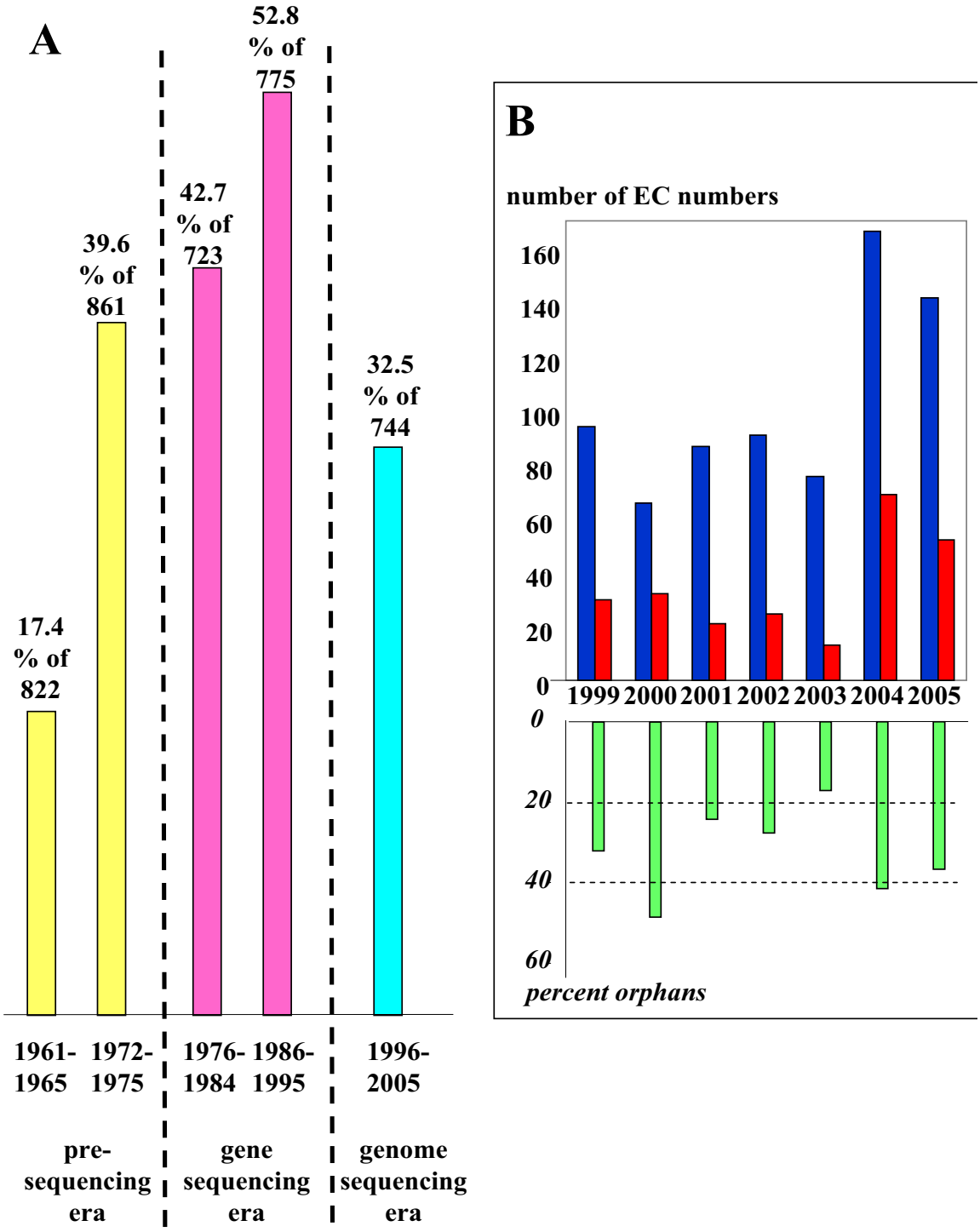


Figure 5
Distribution of the creation year of orphan enzyme activities. 5A. Distribution during the pre-sequencing era (yellow), the gene sequencing era (pink) and the genome sequencing era (cyan). 5B Number of enzymes created within the past seven years that have/lack sequence data. Total number of EC numbers is in blue, total number of orphan EC numbers in red and percentage of orphans in green.

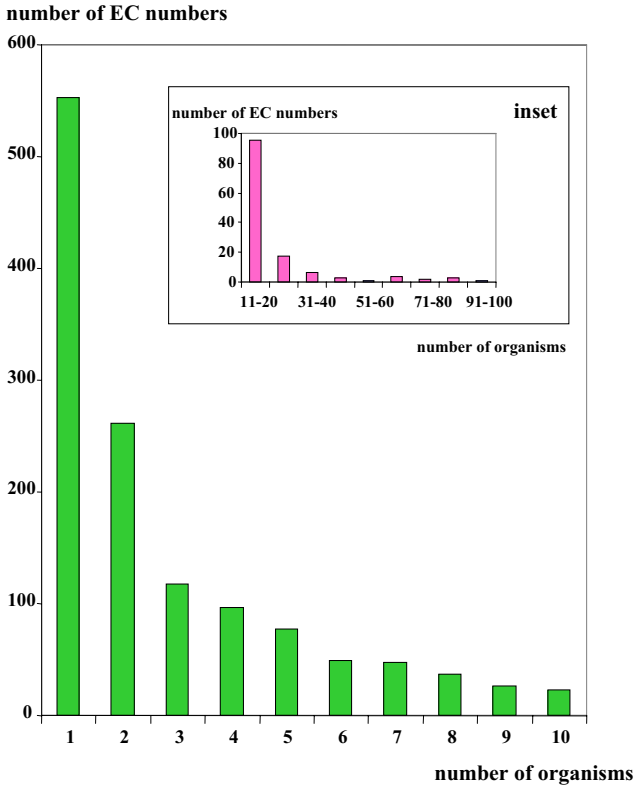


Figure 6

Taxonomic distribution of orphan enzyme activities. Green bars correspond to the distribution of the number of organisms (ranging from one to ten) where orphan EC numbers have been experimentally identified. In the inset, pink bars correspond to the number of orphan EC numbers identified in various ranges of number of organisms larger than ten organisms.

EC number's individual files. If several experts agree on the same suggestion, it would be transmitted to the curators of UniProt with a high degree of confidence. In cases where experts provide conflicting advice, all versions of the advice provided will be published as they have been set and validated. This would allow the community to decide, eventually.

Discussion

The presence of so many EC numbers that do not have an associated sequence appears rather extraordinary at a time where we are inundated by genomic data. Such a situation is encroaching Research at different levels. Alleviating this problem would be very helpful for the difficult task of annotating and/or reannotating genomes. Thus, there is an urgent need to bridge this unwanted gap between biochemical knowledge and massive identification of coding sequences and we and others (see Karp [3]) think that the whole community must contribute to this task. This is why we built this ORENZA resource.

We designed this database to be an interactive tool allowing each expert to exploit his/her knowledge about an (or a group of related) enzyme(s) that have been registered as being an orphan enzyme activity.

Different cases may exist and we already described three of them where personal expertise would eliminate many errors and/or neglected instances. (i) A trivial error takes place when the enzyme has been correctly described in a sequence database but its EC number is not indicated. This is the case for example of glyceraldehyde 3-phosphate dehydrogenases as already shown [5]. One of these sequences (GAPOR, EC 1.2.7.6) has been entered in UniProt without its EC number although the information was given in relevant published papers. Presently, we estimate that up to 20% of the so-called orphan EC numbers might correspond to such a trivial incomplete annotation in the sequence databases (OL & BL, unpublished results). (ii) A sequence or a partial sequence has been previously determined but has not been published. We recently described such an instance in the case of putrescine carbamoyltransferases [16]. (iii) We further observed that around 50% of the present orphan EC numbers are found in only one species or a few closely related organisms as shown on Fig. 6. This is due, in the large majority of the cases, to the fact that we miss genetic tools for such imperfectly studied organisms. Moreover, the availability of genomic sequences for closely related species is useless when the orphan EC numbers are specific for the studied organisms (see Tables 2 and 3).

Table 2: Distribution of orphan enzyme activities in a few model organisms

Model organisms	Total	EC numbers	
		Orphans (/total)	Species specific orphans (/total orphans)
Escherichia coli	1792	189 (0.11)	25 (0.13)
Arabidopsis thaliana	651	22 (0.03)	0 (0)
Saccharomyces cerevisiae	1254	129 (0.10)	14 (0.11)
Drosophila melanogaster	417	16 (0.04)	4 (0.25)
Homo sapiens	1560	225 (0.14)	6 (0.02)

Table 3: The six orphan enzyme activities that are specific to *Homo sapiens*.

EC number	Enzyme name	role in human physiology
EC 2.3.1.125	1-alkyl-2-acetylglycerol O-acyltransferase	platelet activation
EC 3.1.6.15	N-sulfoglucosamine-3-sulfatase	urinary infection by *Flavobacterium heparinum*
EC 1.1.1.160	dihydrobunolol dehydrogenase	liver physiology
EC 2.4.1.153	dolichyl-phosphate α-N-acetylglucosaminyltransferase	liver physiology
EC 3.1.2.13	S-succinylglutathione hydrolase	liver physiology
EC 5.1.3.19	chondroitin-glucuronate 5-epimerase	blood coagulation, cardiovascular disease, carcinogenesis

Conclusion

We consider ORENZA to be a useful resource for all categories of biologists. Let us take for instance the data summarized in Table 2 and more precisely the observation that human cells harbour six enzyme activities that are not found elsewhere and that are not associated with any amino acid sequence (Table 3).

Any biologist would attempt to better understand the origin of such metabolic specificities. Any progress in this field could have positive consequences in terms of medical advances (see Table 3).

The genomicist would wonder if the occurrence of these six orphans is not an indicator of a big annotation problem in the current analysis of the human genome. The expert for either a specific enzyme or a physiological aspect related with these orphan enzyme activities would feel personally concerned and we hope that he/she will promptly answer such a challenge.

Availability and requirements

ORENZA resource is freely available via the Internet at http://www.orenza.u-psud.fr. The web accessibility has been tested to work with the Mozilla 1.7.12, Mozilla Firefox 1.5, and Internet Explorer 6.0 web browsers.

Complete lists of all EC numbers and of orphan EC numbers are available and will be periodically updated. All data can be easily downloaded as text files.

Authors' contributions

OL wrote the different programs necessary to collect all data from public sources and to build the relational database and the web server. Both authors participated in the data analysis and wrote the paper.

Acknowledgements

We thank the two anonymous reviewers for their constructive comments and Claudio Scazzocchio for critical reading of the manuscript and help with the English language. The Agence Nationale de Recherche (programme Masse de Données) and the CNRS have funded this project, including the processing charge for publishing this paper.

References

1. **Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB)** *Eur J Biochem* 1999, **264:**610-650 [http://www.chem.qmul.ac.uk/iubmb/enzyme/index.html]. Enzyme Nomenclature
2. Fleischmann A, Darsow M, Degtyarenko K, Fleischmann W, Boyce S, Axelsen KB, Bairoch A, Schomburg D, Tipton KF, Apweiler R: **IntEnz, the integrated relational enzyme database.** *Nucleic Acids Res* 2004, **32:**D434-437 [http://www.ebi.ac.uk/intenz/index.html].
3. Karp PD: **Call for an enzyme genomics initiative.** *Genome Biol* 2004, **5:**401.
4. Lespinet O, Labedan B: **Orphan enzymes?** *Science* 2005, **307:**42.
5. Lespinet O, Labedan B: **Puzzling over orphan enzymes.** *Cell Mol Life Sci* 2006, **63:**517-523.
6. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS: **The Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2005, **33:**D154-159 [http://www.expasy.uniprot.org/index.shtml].
7. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: **The KEGG resources for deciphering the genome.** *Nucleic Acids Res* 2004, **32:**D277-D280 [http://www.genome.ad.jp/kegg].
8. Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D: **BRENDA, the enzyme database: updates and major new developments.** *Nucleic Acids Res* 2004, **32:**D431-D433 [http://www.brenda.uni-koeln.de/].
9. Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA, Rapp BA: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2000, **28:**10-14 [http://www.ncbi.nlm.nih.gov/Taxonomy/].
10. Bairoch A: **The ENZYME database in 2000.** *Nucleic Acids Res* 2000, **28:**304-305 [http://www.expasy.org/enzyme/].
11. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ: **The PROSITE database.** *Nucleic Acids Res* 2006, **34:**D227-D230 [http://www.expasy.org/prosite/].
12. Berman HM, Henrick K, Nakamura H: **Announcing the worldwide Protein Data Bank.** *Nature Structural Biology* 2003, **10:**980 [http://www.wwpdb.org/].
13. Green ML, Karp PD: **Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers.** *Nucleic Acids Res* 2005, **33:**4035-4039.

14. **PostgreSQL** [http://www.postgresql.org/]
15. **PHP** [http://www.php.net/]
16. Naumoff DG, Xu Y, Glansdorff N, Labedan B: **Retrieving sequences of enzymes experimentally characterized but erroneously annotated: the case of the putrescine carbamoyltransferase.** *BMC Genomics* 2004, **5**:52.

Research

The genome sequence of the model ascomycete fungus *Podospora anserina*

Eric Espagne[¤*†], Olivier Lespinet[¤*†], Fabienne Malagnac[¤*†‡], Corinne Da Silva[§], Olivier Jaillon[§], Betina M Porcel[§], Arnaud Couloux[§], Jean-Marc Aury[§], Béatrice Ségurens[§], Julie Poulain[§], Véronique Anthouard[§], Sandrine Grossetete[*†], Hamid Khalili[*†], Evelyne Coppin[*†], Michelle Déquard-Chablat[*†], Marguerite Picard[*†], Véronique Contamine[*†], Sylvie Arnaise[*†], Anne Bourdais[*†], Véronique Berteaux-Lecellier[*†], Daniel Gautheret[*†], Ronald P de Vries[¶], Evy Battaglia[¶], Pedro M Coutinho[¥], Etienne GJ Danchin[¥], Bernard Henrissat[¥], Riyad EL Khoury[#], Annie Sainsard-Chanet[#**], Antoine Boivin[#**], Bérangère Pinan-Lucarré[††], Carole H Sellem[#], Robert Debuchy[*†], Patrick Wincker[§], Jean Weissenbach[§] and Philippe Silar[*†‡]

Addresses: [*]Univ Paris-Sud, Institut de Génétique et Microbiologie, UMR8621, 91405 Orsay cedex, France. [†]CNRS, Institut de Génétique et Microbiologie, UMR8621, 91405 Orsay cedex, France. [‡]UFR de Biochimie, Université de Paris 7 - Denis Diderot, case 7006, place Jussieu, 75005, Paris, France. [§]Genoscope (CEA) and UMR 8030 CNRS-Genoscope-Université d'Evry, rue Gaston Crémieux CP5706, 91057 Evry, France. [¶]Microbiology, Department of Biology, Utrecht University, Padulaan, 3584 CH Utrecht, The Netherlands. [¥]UMR 6098, Architecture et Fonction des Macromolecules Biologiques, CNRS/univ. Aix-Marseille I et II, Marseille, France. [#]CNRS, Centre de Génétique Moléculaire, UPR 2167, 91198 Gif-sur-Yvette, France. [**]Université Paris-Sud, Orsay, 91405, France. [††]Institut de Biochimie et de Génétique Cellulaires, UMR 5095 CNRS/Université de Bordeaux 2, rue Camille St. Saëns, 33077 Bordeaux Cedex, France.

¤ These authors contributed equally to this work.

Correspondence: Philippe Silar. Email: philippe.silar@igmors.u-psud.fr

Published: 6 May 2008

Genome ***Biology*** 2008, **9:**R77 (doi:10.1186/gb-2008-9-5-r77)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2008/9/5/R77

Received: 26 November 2007
Revised: 12 February 2008
Accepted:

Abstract

Background: The dung-inhabiting ascomycete fungus *Podospora anserina* is a model used to study various aspects of eukaryotic and fungal biology, such as ageing, prions and sexual development.

Results: We present a 10X draft sequence of *P. anserina* genome, linked to the sequences of a large expressed sequence tag collection. Similar to higher eukaryotes, the *P. anserina* transcription/splicing machinery generates numerous non-conventional transcripts. Comparison of the *P. anserina* genome and orthologous gene set with the one of its close relatives, *Neurospora crassa*, shows that synteny is poorly conserved, the main result of evolution being gene shuffling in the same chromosome. The *P. anserina* genome contains fewer repeated sequences and has evolved

new genes by duplication since its separation from *N. crassa*, despite the presence of the repeat induced point mutation mechanism that mutates duplicated sequences. We also provide evidence that frequent gene loss took place in the lineages leading to *P. anserina* and *N. crassa*. *P. anserina* contains a large and highly specialized set of genes involved in utilization of natural carbon sources commonly found in its natural biotope. It includes genes potentially involved in lignin degradation and efficient cellulose breakdown.

Conclusion: The features of the *P. anserina* genome indicate a highly dynamic evolution since the divergence of *P. anserina* and *N. crassa*, leading to the ability of the former to use specific complex carbon sources that match its needs in its natural biotope.

Background

With one billion years of evolution [1], probably more than one million species [2] and a biomass that may exceed that of animals [3,4], eumycete fungi form one of the most successful groups of eukaryotes. Not surprisingly, they have developed numerous adaptations allowing them to cope with highly diverse environmental conditions. Presently, virtually all biotopes, with the exception of extreme biotopes (that is, hyperthermophilic areas), contain some representative eumycetes. They feed by osmotrophy and import through very efficient transporters the nutrients they take up from the environment, often by degrading complex material, such as plant cell walls, that few other organisms can use.

Eumycete fungi have a huge impact on the global carbon cycle in terrestrial biotopes. Some species associate with plant and algae, helping them to scavenge mineral nutrients and to cope with various stresses, such as poor soils, desiccation, parasites and herbivore damage. These mutualistic relationships lead to better carbon dioxide fixation. In contrast, many species parasitize plants and algae, resulting in reduced carbon fixation [5], as well as causing serious economic losses to human agriculture. The majority, however, are saprobic and live on dead plant material, such as fallen plant debris, plants ingested by herbivores or the remains of plants in feces of herbivores. It is estimated that saprobes release 85 billion tons of carbon dioxide annually [6,7], much higher than the 7 billion tons emitted by humans [8]. Finally, some fungi can infect and kill animals, especially invertebrates, which results in diminished carbon fluxes within the food chain. A few are opportunists able to infect humans. Impact on human health is increasing because of the higher prevalence of immunodeficiency, a condition favoring fungal infection.

In addition to these global effects, eumycetes impact their biotope and humans in many ways. Indeed, humans have been using them for thousands of years as food, to process other plant or animal materials and to produce compounds of medicinal interest. A few species degrade human artifacts, causing permanent damage to irreplaceable items. Furthermore, due to their ease of handling, some species, such as *Saccharomyces cerevisiae* or *Neurospora crassa*, have been exploited as research tools to make fundamental biological discoveries. In recent years, a number of genome initiatives have been launched to further knowledge of the biology and evolution of these organisms. Presently, a large effort is dedicated to saccharomycotina yeasts (formerly hemiascomycetes) [9]. Other efforts are concentrated towards human parasites and plant mutualists or pathogens. The genomes of *Magnaporthe grisea*, a rice pathogen, *Fusarium graminearum*, a wheat pathogen, *Ustilago maydis*, a maize pathogen, *Cryptococcus neorformans* and *Aspergillus fumigatus*, two human pathogens, have been published [10-14]. In addition, saprobic fungi are also considered, since the genome sequences of the basidiomycete *Phanerochaete chrysosporium* [15], of the ascomycetes *N. crassa* [16] and *Schizosaccharomyces pombe* [17], and three strictly saprobic Aspergilli, *A. nidulans*, *A. oryzae* and *A. niger* [18-20], are available.

Because of its ease of culture and the speed of its sexual cycle, which is completed within a week, the saprobic filamentous ascomycete *Podospora anserina* (Figure 1) has long been used as a model fungus in several laboratories [21,22] to study general biological problems, such as ageing, meiosis, prion and related protein-based inheritance, and some topics more restricted to fungi, such as sexual reproduction, heterokaryon formation and hyphal interference (Table 1). *P. anserina* and *N. crassa* both belong to the sordariomycete clade of the pezizomycotina (formerly euascomycete). Based on the sequence divergence between the *P. anserina* and *N. crassa* 18S rRNA, the split between the two species has been estimated to have occurred at least 75 million years ago [23]. However, the average amino acid identity between orthologous proteins of the two species is 60-70% [24], the same percentage observed between human and teleost fishes [25], which diverged about 450 million years ago [26,27]. It is not surprising, therefore, that despite similar life cycles and saprobic lifestyles, each species has adopted a particular biotope and displays many specific features (Table 2). To better comprehend the gene repertoire enabling *P. anserina* to adapt to its biotope and permit this fungus to efficiently complete its life cycle, we have undertaken to determine the genome sequence of *P. anserina* and have compared it to that of *N. crassa*, its closest relative for which the genome sequence is already known. We started with a pilot project of about 500 kb (about 1.5% of the

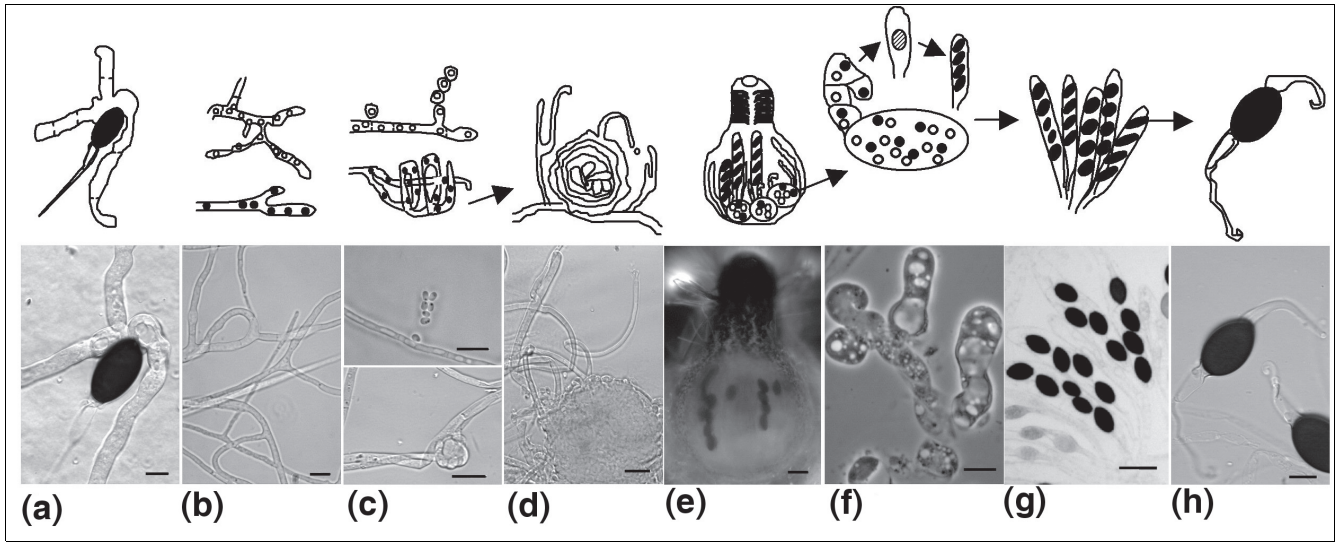


Figure 1
The major stages of the life cycle of *P. anserina* as illustrated by light microphotography, with a corresponding schematic representation shown above. **(a)** The cycle starts with the germination of an ascospore, after the transit in the digestive tract of an herbivore in the wild. **(b)** Then, a mycelium, which usually carries two different and sexually compatible nuclei (pseudo-homothallism), called *mat+* and *mat-*, develops and invades the substratum. **(c)** On this mycelium, male (top; microconidia) and female (bottom; ascogonium) gametes of both mating types differentiate after three days. In the absence of fertilization, ascogonium can develop into protoperithecium by recruiting hyphae proliferating from nearby cells. **(d)** This structure, in which an envelope protects the ascogonial cell, awaits fertilization. **(e,f)** This occurs only between *mat+* and *mat-* sexually compatible gametes (heterothallism) and triggers the development completed in four days of a complex fructification (e) or perithecium, in which the dikaryotic mat+/mat- fertilized ascogonium gives rise to dikaryotic ascogenous hyphae (f). **(g)** These eventually undergo meiosis and differentiate into ascii, mostly with four binucleate mat+/mat- ascospores (pseudo-homothallism), but sometime with three large binucleate ascospores and two smaller uninucleate ones (bottom asci is five-spored). Unlike those issued from large binucleate ascospores, mycelia issued from these smaller ascospores are self-sterile because their nuclei carry only one mating type. **(h)** When ripe, ascospores are expelled from perithecia and land on nearby vegetation awaiting ingestion by an herbivore. Scale bar: 10 μm in (a-d,f,h); 200 μm in (e,g).

genome) [24] and in this paper we present the establishment of a 10X draft sequence.

Results and discussion

Acquisition, assembly and main features of the sequence

The genome of the laboratory reference S mat+ strain was sequenced using a whole-genome shotgun approach (see Materials and methods for a detailed explanation of the sequencing and assembly strategies). Ten-fold coverage permitted complete assembly of the mitochondrial genome as a single circular contig of about 95 kb and most of the nuclear genome (Table 3). The latter was assembled in 1,196 contigs clustered into 33 large scaffolds, comprising nearly all unique sequences, and 45 small scaffolds composed almost exclusively of transposon sequences, collectively totaling 35 Mb. Based on the frequency of sequence runs corresponding to the rDNA compared to that of unique sequences, we estimated that 75 rDNA units are present in the genome. With this assumption, the total sequence length of the genome is 35.5-36 Mb, a value somewhat superior to pulse field estimates [28,29]. Presently, all large scaffolds are assigned to a chromosome as defined by the genome map that now includes over 300 markers (see Materials and methods; Additional data file 1).

The annotation strategy, described in the Materials and methods section, identified 10,545 putative coding sequences (CDSs), including two inteins [30]. 5S rRNA, tRNA, as well as several small nuclear RNAs (snRNAs) and small nucleolar RNAs (snoRNAs) were also identified. Statistics concerning the protein coding capacity of the *P. anserina* genome and the main features of the CDSs are indicated in Table 3. The present estimates of the coding capacity of *N. crassa* are 9,826 CDSs at the Broad Institute [31] and 9,356 CDSs at the Munich Information Center for Protein Sequences (MIPS) [32]. It remains to be established whether the higher coding capacity of *P. anserina* is real or due to the differences in strategies used to annotate the genomes of these fungi. We have searched for orthologous genes between *P. anserina*, *N. crassa*, *M. grisea* and *A. nidulans* by the best reciprocal hit method and found that these four fungi share a common core of 2,876 genes (Figure 2a). Comparison of the *P. anserina* CDSs with *N. crassa* orthologues (Figure 2b) indicates that they are, on average, 60.5 ± 16.0 percent identical, a value similar to the one calculated previously on a small sample [24]. The *P. anserina* CDSs were 54.7 ± 15.8% identical to *M. grisea* and 47.9 ± 15.1% to *A. nidulans* orthologues. The

Table 1

Areas of research that should benefit from the *P. anserina* complete genome sequence

	Original report	Recent works that have benefited from the genome sequence
Ageing and cell degeneration	[40,103]	[104-106]
Cell death	[79]	[104,107]
Self/non-self recognition (vegetative incompatibility and hyphal interference)	[76,79]	[65]
Mating type and inter-nuclear recognition	[108]	[109]
Cell differentiation and cell signaling in filamentous fungi	[110]	[111]
Sexual reproduction in fungi	[21]	[64,111]
Mechanism of meiosis	[22,112]	
Meiotic drive	[113]	
Translation accuracy determinants and role	[114]	[115]; this paper
Mitochondrial physiology	[116,117]	[105]
Peroxisomal physiology and function	[118]	[119]
Prions and other protein-based inheritance	[120,121]	[106]
Biomass conversion	This paper	
Secondary metabolism		[122]

identities reflect the known phylogenetic relationship between these four pezizomycotina and are comparable to those found between species of saccharomycotina [9].

The expressed sequence tag database analysis

In addition to genomic DNA sequencing, a collection of 51,759 cDNAs was sequenced. These originate from libraries constructed at different stages of the *P. anserina* life cycle (Table 4). The resulting expressed sequence tags (ESTs) were mapped on the genomic sequence to help with the annotation but also to gain insight into the transcriptional ability of *P. anserina*. As seen in Table 4, these cDNAs confirmed 5,848 genes. However, we detected alternative splicing events in 3.8% of the clusters. This suggests that the *P. anserina* proteome might be more complex than concluded from the present annotation. Of interest is the presence of 668 transcribed regions without obvious protein-coding capacity (designated here as 'non-coding transcripts'). Some of these produce ESTs that are spliced, poly-adenylated or present in multiple copies, suggesting that they originate from true transcription units. Although some genes may have been miss-called during annotation, these transcription units may correspond to transcriptional noise, code for catalytic/regulatory RNA or reflect polycistronic units coding for small peptides as described recently [33,34]. Finally, we detected 45 antisense transcripts corresponding to 36 different CDSs. These transcripts might potentially be involved in proper gene regulation, as described for the *S. cerevisiae PHO5* gene [35]. In large scale analyses of *Fusarium verticilloides* [36] and *S. cerevisiae* [37] ESTs, similar arrays of alternatively spliced, 'non-coding' and antisense transcripts were detected, suggesting that the production of these 'unusual' transcripts is, in fact, a normal situation in ascomycete fungi, as described for other eukaryotes [38].

Genes putatively expressed through frame-shift or read-through

During the manual annotation of the genome, we detected 14 genes possibly requiring a frame-shift or a read-through to be properly expressed (Additional data file 2). In all cases, sequencing errors were discounted. In addition, ESTs covering putative read-through or frame-shift sites confirm six of them. Some of the putative frame-shifts and read-throughs detected could correspond to first mutations that will lead to pseudogene formation. However, four sites seem conserved during evolution, arguing for a physiological role. One of the putative -1 frame-shift sequences is located in the Yeti retrotransposon, a classic feature of this type of element. The 13 others affect genes coding for cellular proteins. Factors involved in the control of translation fidelity and affecting rates of frame-shift and read-through have been studied in *P. anserina* and shown to strongly impact physiology [39-42]. To date, the reasons for these effects are not known. None of the components responsible for insertion of selenocysteine are found in the *P. anserina* genome, excluding a role in the observed phenotypes of the non-conventional translation insertion of this amino acid, which takes place at specific UGA stop codons [43]. Similarly, no obvious suppressor tRNA was discovered in the genome.

Synteny with *N. crassa*

We have explored in more detail the synteny between orthologous genes in the *P. anserina* and *N. crassa* genomes (Figures 3 and 4). Synteny was defined as orthologous genes that have the same order and are on the same DNA strand. As observed for other fungal genomes [18,44], extensive rearrangements have occurred since the separation of the two fungi. However, most of them seem to happen within chromosomes since a good correlation exists between the gene

Table 2

Comparison between *P. anserina* and *N. crassa* biology

	P. anserina [80]	*N. crassa* [123]
Ecology		
Habitat	Restricted on dung of herbivores	Prefers plants killed by fire
	Always small biotopes and high competition	Often large biotopes and low competition
Distribution	Worldwide	Prefers hot climate
Vegetative growth		
Growth rate	Average (7 mm/d)	High (9 cm/d)
Ageing syndrome	Senescence in all investigated strains	Mostly immortal with some ageing strains
Hyphal interference	Present	Not yet described
Major pigments	Melanins (green)	Carotenoids (orange)
Reproduction		
Asexual reproduction	None	Efficient with germinating conidia
Sexual generation time	One week	Three weeks
Mating physiology	Pseudohomothallic	Strict heterothallic
Ascospore dormancy	No	Yes
Ascospore germination trigger	Passage through digestive track of herbivores in nature (on low nutrient media containing ammonium acetate in the laboratory)	60°c heat shock or chemicals (for example, furfural)
Gene inactivation processes		
RIP	Not efficient	Very efficient
MSUD	Not yet described	Efficient
Quelling	Not yet described	Efficient

Features and references pertaining to the biology of both fungi can be found at the corresponding reference.

contents of many chromosomes, even though a few translocations are detected (Figure 3). For example, most of *P. anserina* chromosome 1 corresponds to *N. crassa* chromosome I except for a small part, which is translocated to the *N. crassa* chromosome IV. Within the chromosomes, numerous rearrangements have occurred, compatible with the prevalence of small inversions in fungal genome evolution as observed previously between genes of saccharomycotina (hemiascomycetous) yeasts [45]. The size of the synteny blocks loosely follows an exponential decrease (Figure 4), compatible, therefore, with the random breakage model [46], suggesting that most breaks occur randomly, as observed for genome evolution in Aspergilli [18]. However, in both Aspergilli and saccharomycotina yeasts, blocks of synteny have been dispersed among the various chromosomes [18,47], unlike what is observed between *P. anserina* and *N. crassa*. This discrepancy of genome evolution between the three groups of fungi might stem from the fact that *P. anserina* and *N. crassa* have likely had a long history of heterothallism, whereas Aspergilli and saccharomycotina yeasts are either homothallic, undergo a parasexual cycle or switch mating types. In heterothallics, the presence of interchromosomic translocation results in chromosome breakage during meiosis and, hence, reduced fertility. On the contrary, homothallism, parasexualilty or mating-type switching may allow translocation to be present in both partners during sexual reproduction and, therefore, have fewer consequences on fertility. Additionally, meiotic silencing of unpaired DNA (MSUD), an epigenetic gene silencing mechanism operating in *N. crassa* [48], abolishes fertility in crosses involving rearranged chromosomes in one of the partners.

Interestingly, the largest synteny block between *P. anserina* and *N. crassa*, with 37 orthologous genes, encompasses the mating type, a region involved in sexual incompatibility. A similar trend in conserved synteny in the mating-type region has been observed in the genus *Aspergillus* [18]. This suggests that recombination may be inhibited in this region on an evolutionary scale. In both *P. anserina* and *N. crassa*, the mating-type regions are known to display peculiar properties. In *P. anserina*, meiotic recombination is severely repressed around the mating-type locus [49], as also described in *Neurospora tetrasperma* [50]. In *N. crassa*, MSUD is inhibited in the *mat* region [48]. However, recombination is not completely abolished around this locus. Indeed, between pairs of orthologous genes, a few species-specific CDSs were detected. These genes may come from *de novo* insertion or, alternatively, these species-specific genes have been lost in the other

Table 3

Main features of the *P. anserina* genome

Genome features	Value
Nuclear genome	
Size	35.5-36 Mb
Chromosomes	7
GC percentage (total genome)	52.02
GC percentage in coding sequences	55.87
GC percentage in non-coding regions	48.82
tRNA genes	361
rDNA repeat number	75
Consensus rDNA repeat size	8192 pb
5S rRNAs	87
snRNA genes	14
snoRNA genes	13
Protein coding genes (CDSs)	10545
Percent coding	44.75
Average CDS size (min; max)	496.4 codons (10; 8,070)
Average intron number/CDS (max)	1.27 (14)
Average intron size (max)	79.32 nucleotides (2,503 nucleotides)
Mitochondrial genome	
Size	94,197 bp
Chromosome	1 (circular)
GC percentage	30%

species. This lends credit to the hypothesis put forward to explain the mating-type region of *Cryptococcus neoformans* [51], in which the genetic incompatibility is driven by two genetically different sequences of 100 kb. In these regions, not only the mating-type regulatory genes are different, but also housekeeping genes. Inhibition of recombination at this locus may have driven the differential acquisition of genes by the two haplotypes within the same species. Note that on a longer evolutionary scale, inhibition of recombination cannot be detected because the synteny of the mating-type region of *P. anserina* with that of *M. grisea* or *A. nidulans* is absent or limited to very few genes.

Repeated sequences in the *P. anserina* genome

The pilot project that sequenced about 500 kb around the centromere of chromosome 5 revealed an apparent paucity in repeated sequences in *P. anserina* [24]. The draft sequence reported here confirms a paucity of repeats but not as much as suggested by the pilot project. In fact, repeats cover about 5% of the *P. anserina* genome (omitting the rDNA cluster). They can be divided into four categories: RNA genes (Table 3; see Materials and methods), true transposons (Additional data file 3), repetitive elements of unknown origin (Additional data file 3) and segmental duplications (Additional data file 4). Collectively, the transposons occupy about 3.5% of the genome. However, as many transposons border the sequence gaps present in the draft assembly, the actual percentage in the complete genome may be higher. This is about three times less than in the genomes of *M. grisea* [11] and *N. crassa* [16], close relatives of *P. anserina*. Most segmental amplifications are small (Additional data file 4), although one is 20 kb large. They occupy about 1.5% of the genome. An interesting feature of all these repeated sequences (except for the 5S RNA and tRNA genes) is that they are nested together (Figure 5), as previously described for *Fusarium oxysporum* transposons [52]. In particular, large parts of many chromosomes are almost devoid of these repeated sequences whereas chromosome 5 is enriched in repeats. Ironically, the pilot project sequenced a region of this chromosome 5 almost devoid of repeated sequences.

Nearly all copies of these repeated elements differ by polymorphisms, many of which appear to be caused by repeat induced point mutation (RIP). RIP is a transcriptional gene silencing and mutagenic process that occurs during the sexual dikaryotic stage of many pezizomycotina [53]. *P. anserina* displays a very weak RIP process [54,55]. It results, as in *N. crassa*, in the accumulation of C●G to T●A transitions in duplicated sequences present in one nucleus, and, therefore, 'ripped' sequences present a higher than average T/A content. However, although the RIP process acts in the *P. anserina* genome, it does not account for all the mutations found in

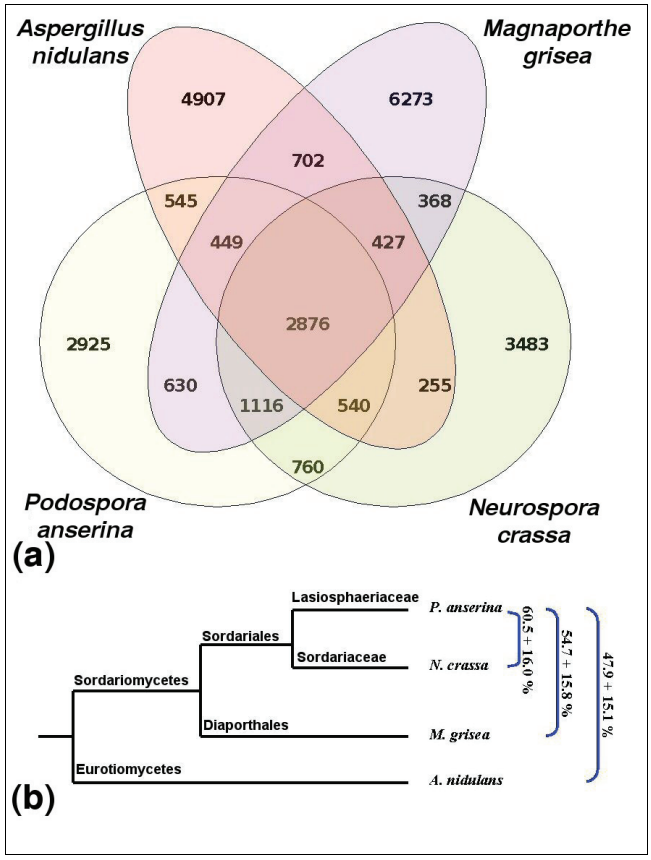


Figure 2
Orthologue conservation in some Pezizomycotina. **(a)** Venn diagram of orthologous gene conservation in four ascomycete fungi. The diagram was constructed with orthologous genes identified by the best reciprocal hit method with a cut-off e-value lower than 10^{-3} and a BLAST alignment length greater than 60% of the query CDS. **(b)** Phylogenetic tree of the four fungal species. The average percentage of identity ± standard deviation between orthologous proteins of *P. anserina* and the three other fungi are indicated on the right.

these inactivated paralogues. For example, the copies of 'rainette', the last transposon to have invaded the *P. anserina* genome (Additional data file 3), differ by 30 polymorphic sites. Twenty-five of them (83%) were C●G versus T●A polymorphisms and may, therefore, be accounted for by RIP, while the five others (17%) cannot. A reciprocal ratio was observed in other instances as seen for the largest segmental triplication with two copies present on chromosome 5 and one on chromosome 1. The three members share a common region of about 9 kb. In this region they differ by numerous indels and in about 20% of their nucleotides. More precisely, in the 4,000 nucleotide-long core region where the three sequences can unambiguously be aligned, there are 1,341 polymorphic sites in which at least one sequence differs from the others. For 418 of them (31%), two members have a C●G polymorphism whereas the other has a T●A polymorphism, strongly suggesting that these polymorphisms may originate from RIP, whereas for the remaining 923 (69%), the variations are small indels or single nucleotide variations not accounted for by RIP. Therefore, in the case of rainette, RIP polymorphisms are foremost, whereas for the triplication, non-RIP polymorphisms are more frequent. This is compatible with a model in which RIP occurs first and is then followed by accumulation of other types of mutations.

Overall, these data suggest that *P. anserina* has experienced a fairly complex history of transposition and duplications, although it has not accumulated as many repeats as *N. crassa*. *P. anserina* possesses all the orthologues of *N. crassa* factors necessary for gene silencing (Additional data file 5), including RIP, meiotic MSUD [48] and also vegetative quelling, a post transcriptional gene silencing mechanism akin to RNA interference [56]. However, to date, no MSUD or quelling has been described in *P. anserina*, despite the construction of numerous transgenic strains since transformation was first performed [57]. Surprisingly, the DIM-2 DNA methyltransferase [58], the RID DNA methyltransferase-related protein [59] and the HP1 homolog necessary for DNA methylation [60] described in *N. crassa* are present in the genome of *P. anserina*. Although the *P. anserina* orthologues of these two proteins seem functional based on the analysis of the conserved catalytic motifs, no cytosine methylation has been reported to occur in this fungus [54]. A possibility would be that methylation is restricted to a specific developmental stage or genomic region that has not yet been investigated. Overall, the apparent absence (quelling and MSUD) or lack of efficiency (RIP) of these genome protection mechanisms in *P. anserina* questions their true impact on genome evolution, especially since this fungus contains less repeated sequences than *N. crassa*. Maybe the life strategy of *P. anserina* makes it less exposed to incoming selfish DNA elements, therefore diminishing the requirement of highly efficient gene silencing mechanisms. Supporting this assumption is the fact that, although heterothallic, formation of ascospores makes *P. anserina* pseudo-homothallic (Figure 1), with seemingly very little out-crossing [61], whereas *N. crassa* is strictly heterothallic and presents a low fertility in crosses between closely related strains [62].

Gene evolution by duplication and loss in fungi

The detection of segmental duplications raised the question of whether new genes evolved through duplication in the lineage that gave rise to *P. anserina*. It is known that creating new genes through duplication in *N. crassa*, in which RIP is very efficient, is almost impossible [16]. On the contrary, RIP is much less efficient in *P. anserina*; in particular, RIP is absent in progeny produced early during the maturation of the fructifications [55]. In addition, the mutagenic effect of RIP is very slight since it has been estimated that at most 2% of cytosines are mutated when RIP affects duplicated sequences present on two different chromosomes [63]. We previously reported that some thioredoxin isoforms were encoded by a triplicated gene set in *P. anserina* as compared to *N. crassa* [64], showing that gene duplications can indeed generate new genes in *P. anserina*. However, thioredoxins are

Table 4

EST analysis

	Number of sequenced cDNA clones	Number of clusters	Confirmed genes*	Alternatively spliced transcripts			Non-coding transcripts not covering a predicted CDS	Antisense transcripts
				Exon cassette	Alternative splice site	Retained intron		
Bank								
Mycelium grown for 48 h	27,291	6,054	5,780	1	155	137	322	19
Young perithecia of less than 48 h	7,695	2,392	2,236	2	46	55	258	12
Perithecia older than 48 h	7,814	2,373	2,088	2	26	51	440	4
Ascospores 20 h after germination trigger	5,570	1,589	1,502	0	29	28	125	3
Senescent mycelium	1,136	718	665	0	10	9	59	4
Incompatible mycelium	1,133	514	474	1	7	6	54	1
Rapamycin induced mycelium	1,120	593	543	1	3	11	68	2
All databanks	51,759	6,618	5,848	5	80	167	668	36

*Cluster covering a CDS.

small proteins encoded by small genes. To test if large genes were duplicated, we performed a three-way comparison between the *P. anserina*, *N. crassa* and *M. grisea* putative CDSs and screened for *P. anserina* CDSs that show a best hit with another *P. anserina* CDS to the exclusion of proteins from *N. crassa* and *M. grisea*. Such CDSs may originate from duplication that occurred in the *P. anserina* lineage after its divergence from *N. crassa*. In this analysis, small genes were excluded because the putative candidates were selected on the basis of an e-value of less than 10^{-190} in Blast comparison against the database containing the three predicted proteomes (as a consequence, the thioredoxin genes were not included in the set).

To confirm that the candidates recovered indeed originated from recent duplications, phylogenetic trees were constructed with the CDSs from *P. anserina*, *N. crassa*, *M. grisea* and additional fungal CDSs. In some instances, the trees confirmed a recent duplication event in the *P. anserina* lineage after the split between *P. anserina* and *N. crassa*, because the phylogenetic analysis clustered the *P. anserina* paralogues with high statistical confidence. Figure 6 shows the trees obtained for three such couples of paralogues, for example, genes coding for putative alkaline phosphatase D precursors (Pa_4_1520 and Pa_6_8120; Figure 6a), putative HC-toxin efflux carrier proteins related to ToXA from *Cochliobolus carbonum* (Pa_2_7900 and Pa_6_8600; Figure 6b) and putative chitinases related to the killer toxin of *Kluyveromyces lactis* (Pa_4_5560 and Pa_5_1570; Figure 6c). Overall, our analysis detected an initial set of 33 putative duplicated gene families, including the het-D/E gene family, whose evolutionary history has been reported elsewhere [65]. Among these, at least nine (including the het-D/E genes) have duplicated recently. However, some additional recent duplication events may have occurred but are not supported with sufficient statistical confidence to differentiate between recent duplications followed by rapid divergence, and ancient duplications (see Figure 6c for an example of such duplications with putative chitinases). The fact that large genes may duplicate in *P. anserina* is not contradictory to the presence of RIP, since if RIP may inactivate genes when efficient, it can accelerate gene divergence when moderately efficient, as described for the het-D/E family [65].

The phylogenetic analyses of the multigene families suggest that gene loss may also have occurred during fungal evolution. The putative chitinases related to the killer toxin of *K. lactis* provide a clear example of this situation. *N. crassa* and *M. grisea* have two paralogues, whereas *P. anserina* has eight. The phylogenetic tree including the ten paralogues present in *A. nidulans* (Figure 6c) suggests that these proteins can be grouped into two families. Surprisingly, the *P. anserina* proteins cluster in one subfamily, whereas the *M. grisea* proteins cluster in the other, indicating differential gene losses. In *P. anserina*, even if Pa_4_5560 and Pa_5_1570 seem to have duplicated recently, this is not as clear for the other members since they are not very similar. They may result from ancient gene duplications or from recent duplications followed by rapid evolution, possibly thanks to RIP. Evolution of this family seems thus to proceed by a complex set of gain and loss at various times. The same holds true for polyketide synthase (PKS) genes. Seven PKSs were

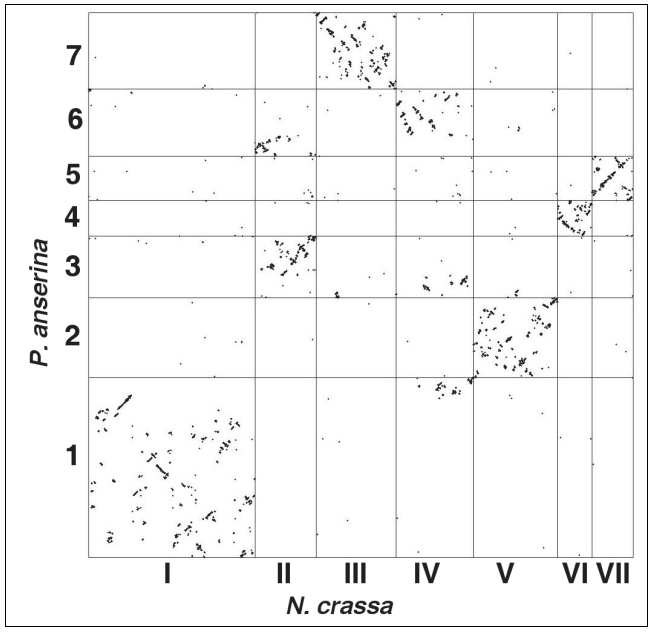


Figure 3
Genome-wide comparison of orthogolous genes of *N. crassa* (x-axis) and *P. anserina* (y-axis). Each dot corresponds to a couple of orthologous genes. The lines delimit the chromosomes. The scale is based on the number of orthologous genes per chromosome.

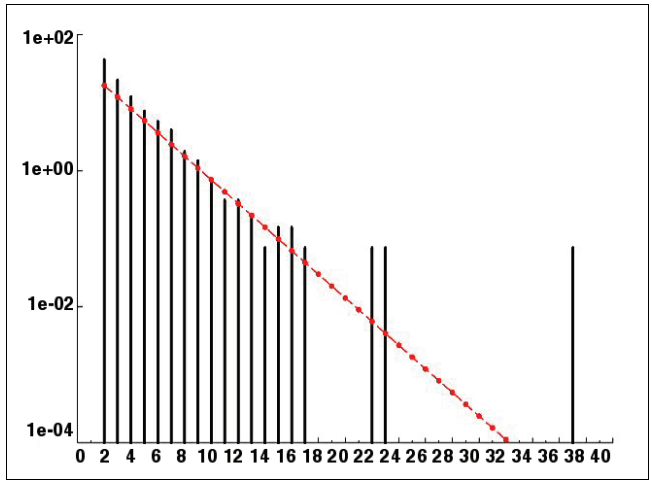


Figure 4
Size distribution of synteny block between *P. anserina* and *N. crassa*. Block size is given on the x-axis and frequency on the y-axis. Black bars indicate the actual value, and the red line shows the theoretical curve expected in the case of the random break model. The two distribution functions are not statistically different (Kolmogorov-Smirnov test, $p >> 5\%$).

reported for *N. crassa* [16], while *M. grisea* has 23 [11], and we identified 20 PKS genes for *P. anserina*. A comparison of all these PKSs (data not shown) indicates a complex evolution process in which *N. crassa* has probably lost most of its PKSs and the two other fungi present several duplications yielding very different copies. Again, this does not permit us to establish whether the duplications are ancient or recent but followed by intense divergence. See also below for additional examples of losses and amplifications of genes involved in carbon source degradation.

Such gene losses may be frequent events in filamentous ascomycete. As seen in Figure 2a, *P. anserina*, *M. grisea* and *A. nidulans* share 1,624 genes that seem to be lacking in *N. crassa* (among these, 449 are present in the three fungi, 630 in both *P. anserina* and *M. grisea*, and 545 in both *P. anserina* and *A. nidulans*), even though *M. grisea* and *A. nidulans* are more distantly related to *P. anserina* than is *N. crassa* (Figure 2b). Although some genes may have evolved beyond recognition specifically in *N. crassa*, the most parsimonious explanation is that *P. anserina* has retained many genes that *N. crassa* has lost. Similarly, *N. crassa*, *M. grisea* and *A. nidulans* share 1,050 genes that are absent in *P. anserina*. Therefore, we tentatively suggest that genomes from sordariomycetes may be shaped by more gene loss and gene duplications than anticipated by the presence of RIP. Similar rates of gene loss in filamentous ascomycetes have recently been demonstrated [66].

Carbon catabolism

In nature, *P. anserina* lives exclusively on dung of herbivores. In this biotope, a precise succession of fungi fructifies [67]. An explanation put forward to account for this succession is nutritional. The first fungi to appear feed preferably on simple sugars, which are easy to use, followed by species able to digest more complex polymers that are not easily degraded. Indeed, the mucormycotina zygomycetes, which are usually the first ones to fructify on dung, prefer glucose and other simple sugars as carbon sources. They are followed by ascomycetes that use more complex carbohydrates such as (hemi)cellulose but rarely degrade lignin. The succession ends with basidiomycetes, some of which can degrade lignin to reach the cellulose fiber not available to other fungi [68-70].

Usually, *P. anserina* fructifies in the late stage of dung decomposition [67]. This late appearance of the *P. anserina* fruiting body is hard to correlate with slow growth of the mycelium and delay in fructification since in laboratory conditions ascospore germination occurs overnight and fruit body formation takes less than a week. However, *P. anserina* harbors unexpected enzymatic equipment, suggesting that it may be capable of at least partly degrading lignin, which concurs with the nutritional hypothesis (Table 5). It includes a large array of glucose/methanol/choline oxidoreductases [71], many of which are predicted to be secreted, two cellobiose dehydrogenases, a pyranose oxidase, a galactose oxidase, a copper radical oxidase, a quinone reductase, several laccases and one putative Lip/Mn/Versatile peroxidase. Enzymes homologous to these CDSs are known to produce or use reactive oxygen species during lignin degradation [68-70]. This ascomycete fungus may thus be able to access car-

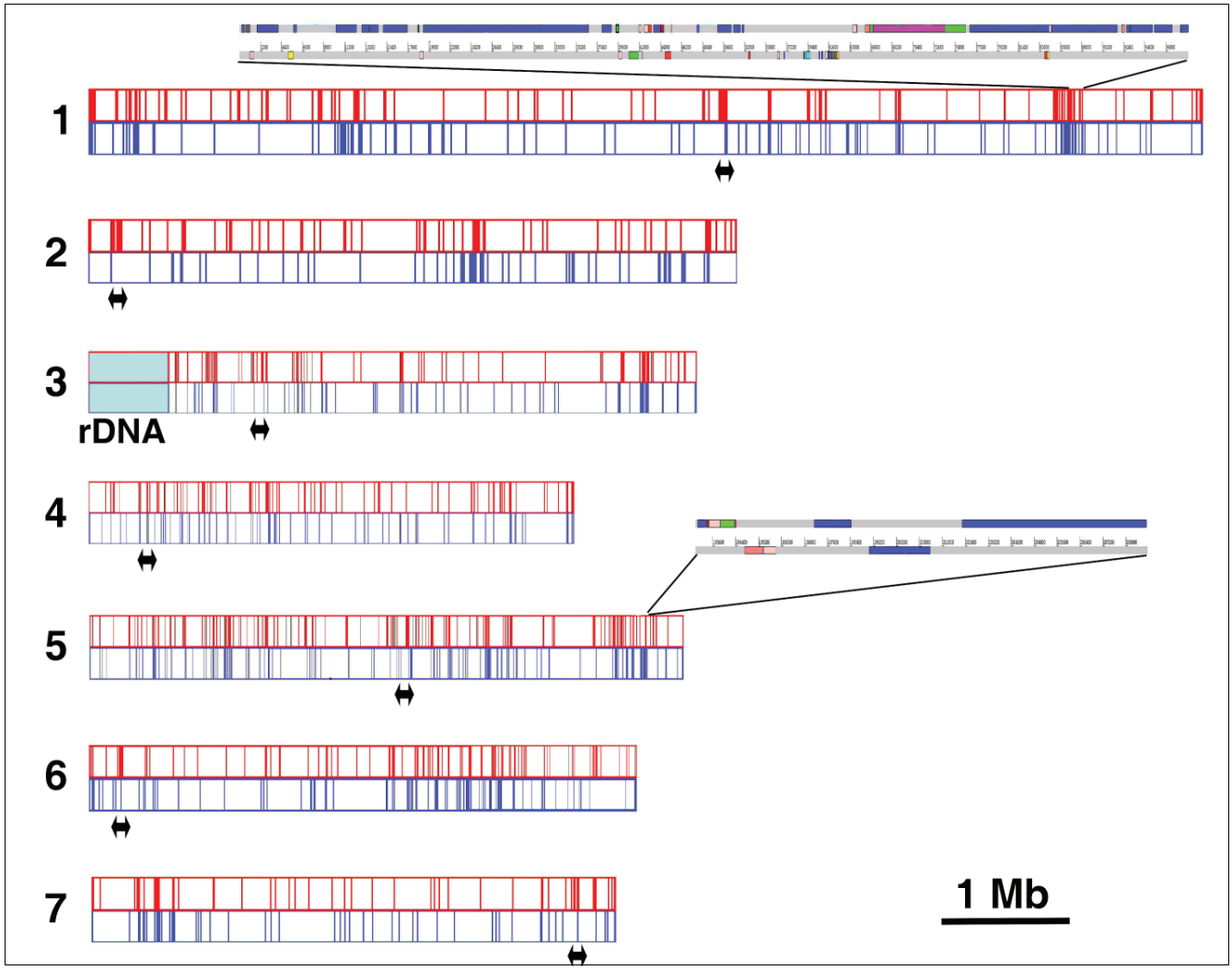


Figure 5
Repartition of transposons (top in red) and segmental duplications (bottom in blue) in the *P. anserina* genome. Chromosome numbering and orientation is that of the genetic map [85]. The double arrows indicate the putative centromere positions. Two regions have been expanded to show the interspacing of segmental duplications (in blue) with transposons (other colors); numbering refers to the nucleotide position with respect to the beginning of the scaffolds.

bon sources normally available mainly to basidiomycetes. Interestingly, *P. anserina* is closely related to xylariales, a group of ascomycete fungi that seems to contain true white rot fungi capable of degrading lignin [72]; also, *P. anserina* has the most complete enzymatic toolkit involved in lignin degradation when compared to the three other ascomycetes included in Table 5. The comparison with *N. crassa* is particularly striking. This is in line with the fact that *N. crassa* in its less competitive biotope may have access to more easily digestible carbon sources.

As mentioned above, *P. anserina* is considered a late growing ascomycete on herbivorous dung. This suggests that the fungus is likely to target lignocellulose as a carbon source, since most hemicellulose and pectin would probably be consumed by zygomycetes and early ascomycetes. A close examination of the genome sequence of *P. anserina* for the presence of carbohydrate active functions (Additional data file 6) and a comparison with the genome sequence of other fungi confirmed the adaptation capacity of *P. anserina* to growth on lignocellulose. The total number of putative glycoside hydrolases (GHs), glycoside transferases, polysaccharide lyases (PLs) and carbohydrate esterases (CEs) are similar to those of other ascomycetes, such as *A. niger* [20] and *M. grisea* [73], but *P. anserina* has the highest number of carbohydrate-binding modules (CBMs) of all the fungal genomes sequenced to date. Despite possessing similar numbers of putative enzymes, the distribution of the possible enzyme functions related to plant cell wall degradation (Table 6) is significantly different in *P. anserina* from that of other fungi. *P. anserina* has the largest fungal set of candidate enzymes for cellulose degradation described to date. This is particularly remarkable in GH family 61 (GH61) with 33 members, two-fold higher than the phytopathogen ascomycete *M. grisea* and the white rot basid-

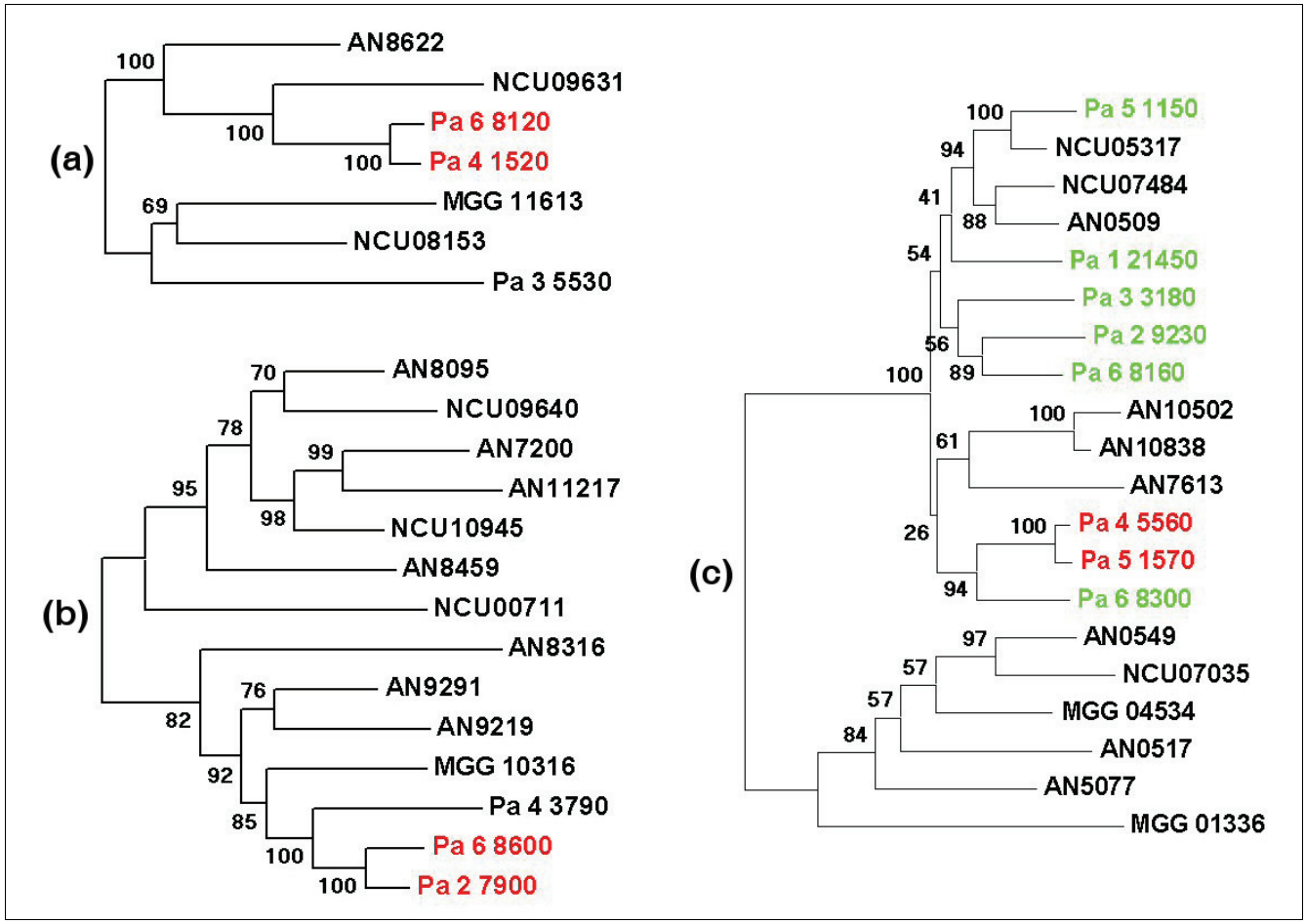


Figure 6
Gene gain and loss in fungal genomes. **(a-c)** Unrooted phylogenetic trees of putative alkaline phosphatase D precusors (a), putative HC-toxin efflux carrier proteins related to ToXA from *Cochliobolus carbonum* (b), and putative chitinases related to the killer toxin of *Kluyveromyces lactis* (c). The putative CDSs were aligned with ProbCons 1.10 [101] and manually edited to eliminate poorly conserved regions, resulting in alignment over 565, 544, 505 amino acids, respectively. Phylogenetic trees were constructed with Phyml 2.4.4 [102] under the WAG model of amino acid substitution. The proportion of variable sites and the gamma distribution parameters of four categories of substitution rate were estimated by phyml. For each tree, we performed 100 boostrap replicates. The recently duplicated *P. anserina* paralogues are highlighted in red and the divergent duplication of chitinases in green. Trees with similar topologies and statistical support (1,000 boostrap replicates) were recovered with the neighbor joining method. Especially, recent duplication of Pa_4_1520/Pa_6_8120, Pa_2_7900/Pa_6_8600 and Pa_4_5560/Pa_5_1570 as well as the distinction of the two subfamilies of chitinases were recovered with 100% confidence. AN, *A. nidulans*; MGG, *M. grisea*; NC, *N. crassa*; Pa, *P. anserina*.

iomycete *P. chrysosporium*. Similar patterns are visible for other cellulose-degrading families (for example, GH6, GH7, GH45) and in the high number of CBM1 (possibly cellulose-binding) modules found, which are only equivalent to the sets of *P. chrysosporium* and *M. grisea*.

Strikingly, *P. anserina* also has an increased potential for xylan degradation, with abundant enzyme sets in families GH10 and GH11, together with a relative abundance of exo-acting enzymes in families GH3 and GH43. Interestingly, no α-fucosidases of families GH29 and GH95 are found, suggesting a depletion of xyloglucan prior to growth of *P. anserina*. During the stage at which *P. anserina* grows in dung, significant amounts of cellulose, but also xylan, are still available. Xylan can be cross-linked to lignin through ferulic acid [74] or 4-O-methyl-glucuronic acid [75]. In light of the potential of *P. anserina* for lignin degradation, it is conceivable that this fungus particularly consumes lignin-linked xylan that could not be degraded by 'earlier' growing organisms that lack a lignin-degradation system. The relatively high number of putative CE1 acetyl xylan and feruloyl esterases found in *P. anserina* by comparison with other fungi correlates with this hypothesis.

In contrast to the increased potential for cellulose and xylan degradation, a significantly weak potential for pectin degradation was observed for *P. anserina*. No members of GH28 (containing pectin hydrolases) were detected in the genome and only a single α-rhamnosidase (GH78). In comparison, *A. niger* contains 21 GH28 members and 8 GH78 members [20].

Table 5

CDSs putatively involved in lignin degradation

		P. anserina				
	Reference		Secretion*	*N. crassa*	*M. grisea*	*A. nidulans*
GMC oxidoreductases	[124]	Pa_0_190	+	NCU09798.3	MGG_07580.5	AN2175.3
		Pa_5_1280	+?	NCU04938.3	MGG_07941.5	AN7998.3
		Pa_1_15920	+	NCU01853.3	MGG_08438.5	AN4006.3
		Pa_5_4870	-	NCU07113.3	MGG_10479.5	AN3229.3
		Pa_4_5130	+	NCU09024.3	MGG_05055.5	AN4212.3
		Pa_5_5180	+?	NCU08977.3	MGG_10933.5	AN9011.3
		Pa_6_6430	-?		MGG_11204.5	AN7267.3
		Pa_1_23060	+		MGG_12623.5	AN9348.3
		Pa_6_7550	+		MGG_12626.5	AN6445.3
		Pa_2_7270	+		MGG_14477.5	AN7812.3
		Pa_1_470	+		MGG_02127.5	AN1093.3
		Pa_5_9820	-		MGG_09072.5	AN2704.3
		Pa_6_1080	+		MGG_06596.5	AN3531.3
		Pa_1_24480	-		MGG_00779.5	AN7408.3
		Pa_0_340	-		MGG_02371.5	AN1429.3
		Pa_4_880	+		MGG_10948.5	AN8329.3
		Pa_7_4250	+		MGG_11676.5	AN8547.3
		Pa_5_12190	+		MGG_13253.5	AN3206.3
		Pa_4_4320	+?		MGG_11317.5	AN0567.3
		Pa_3_11130	+?		MGG_13583.5	AN3960.3
		Pa_1_21970	-		MGG_08487.5	AN7890.3
		Pa_3_1060	+		MGG_09189.5	AN7832.3
		Pa_7_4780	+		MGG_07569.5	AN7056.3
		Pa_0_440	+			

Table 5 *(Continued)*

CDSs putatively involved in lignin degradation						
		Pa_5_4150	+			
		Pa_6_11490	+			
		Pa_5_12200	+			
		Pa_5_13040	-?			
		Pa_6_11360	-			
Cellobiose dehydrogenases	[125]	Pa_7_2650	+	NCU00206.3	MGG_11036.5	AN7230.3
		Pa_0_280	+	NCU05923.3	MGG_13809.5	
Pyranose oxidases	[126]	Pa_6_8060	?	-	-	AN5281.3
Galactose oxidases	[127]	Pa_1_18310	+	NCU09209.3	MGG_10878.5	-
					MGG_12681.5	
Copper radical oxidases	[128]	Pa_1_7300	+	NCU09267.3	MGG_01655.5	-
					MGG_05865.5	
Quinone reductase	[129]	Pa_1_6390	-?	NCU02948.3	MGG_01569.5	AN0297.3
Laccases	[130]	Pa_5_1200	+?	NCU04528.3	MGG_09102.5	AN0901.3
		Pa_5_4660	+	NCU05113.3	MGG_08523.5	AN6635.3
		Pa_7_4200	+	NCU05604.3	MGG_07771.5	AN0878.3
		Pa_5_9860	+	NCU09279.3	MGG_02876.5	AN6830.3
		Pa_7_3560	+?	NCU02201.3	MGG_09139.5	AN5397.3
		Pa_6_10630	+	NCU00526.3	MGG_05790.5	AN9170.3
		Pa_1_15470	+	NCU07920.3	MGG_11608.5	
		Pa_6_7880	-	NCU09023.3	MGG_08127.5	
		Pa_1_16470	+		MGG_13464.5	
		Pa_5_4140	?			
lip/Mn/versatile peroxidases	[70,131]	Pa_1_5970	?	-	MG_07790.5	-
					MGG_03873.5	

*Orthologues were identified by the best reciprocal hit method. Putative secretion was evaluated by searching for the presence of a secretion signal peptide with Interproscan or by evaluating the most probable localization with WolfPSORT. In most instances, both methods yielded the same result. '+', protein likely secreted; '-', protein likely not secreted; '?', no firm conclusion could be reached as to the actual localisation. GMC, glucose/methanol/choline.

Table 6

Comparison of relevant CAZy family content related to plant cell wall polysaccharide degradation

CAZy family	Main substrate	*P. anserina*	*N. crassa*	*M. grisea*	*A. nidulans*	*A. niger*	*P. chrysosporium*
Plant cell wall degradation							
GH1	Cellulose/hemicellulose	1	1	2	3	3	2
GH2	Hemicellulose	8	5	6	10	6	2
GH3	Cellulose/hemicellulose/xylan	11	9	19	21	17	11
GH5	Cellulose	15	7	13	16	10	20
GH6	Cellulose	4	3	3	2	2	1
GH7	Cellulose	6	5	6	3	2	9
GH10	Xylan	9	4	5	3	1	6
GH11	Xylan	6	2	5	2	4	1
GH12	Cellulose/xylan	2	1	3	1	3	2
GH28	Pectin	0	2	3	10	21	4
GH29	Hemicellulose	0	0	4	0	1	0
GH35	Hemicellulose	1	2	0	4	5	3
GH36	Hemicellulose	1	1	2	4	3	0
GH43	Hemicellulose	13	7	19	18	10	4
GH45	Cellulose	2	1	1	1	0	0
GH51	Hemicellulose	1	1	3	3	3	2
GH53	Hemicellulose	1	1	1	1	2	1
GH54	Hemicellulose	0	1	1	1	1	0
GH61	Cellulose	33	14	17	9	7	15
GH62	Hemicellulose	2	0	3	2	1	0
GH67	Xylan	1	1	1	1	1	0
GH74	Hemicellulose	1	1	1	2	1	4
GH78	Pectin	1	0	1	9	8	1
GH88	Pectin	0	0	1	3	1	1
GH93	Hemicellulose	3	2	1	2	0	0
GH94	Cellulose	1	1	1	0	0	0
GH95	Hemicellulose	0	0	1	3	2	1
GH105	Pectin	0	1	3	4	2	0
PL1	Pectin	4	1	2	9	6	0
PL3	Pectin	2	1	1	5	0	0
PL4	Pectin	1	1	1	4	2	0
PL9	Pectin	0	0	0	1	0	0
PL11	Pectin	0	0	0	1	0	0
CE1	Xylan	14	7	10	4	3	5
CE8	Xylan	1	1	1	3	3	2
CE12	Xylan	1	1	2	2	2	0
CBM1	Cellulose	28	20	22	7	8	30
Other relevant families							
GH18	Chitin	20	12	14	20	14	11
GH32	Sucrose/inulin	0	1	5	2	6	0
CBM18	Chitin	30	3	29	19	13	1

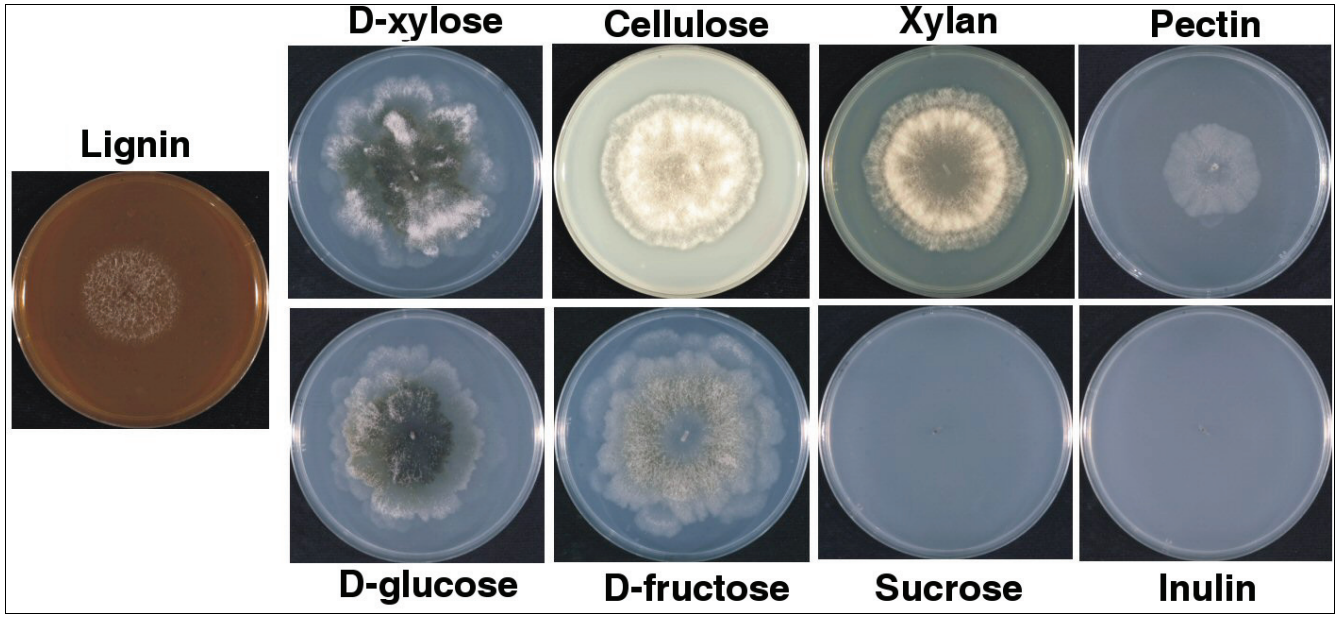


Figure 7
Carbohydrate utilization in *P. anserina*. Cultures were incubated for one week with 1% of the indicated compounds as carbon source.

The number of putative pectin lyases is also much smaller than that observed for *A. niger*. The auxiliary activities of GH88 and GH105, likely to act on pectin lyase degradation products, are equally absent from *P. anserina* while present in all pectin-degrading organisms (Table 6). The absence of the potential to degrade sucrose and inulin is concluded from the lack of enzymes in the GH32 family. This also correlates with the low capacity of *P. anserina* to grow on rapidly degradable carbohydrates that are most likely depleted by 'earlier' organisms. Furthermore, the large number of GH18 and CBM18 modules, 20 and 30 respectively, could indicate that *P. anserina* has the ability to degrade exogenous chitin and possibly to depend on available fungal cell material (derived from the set of fungi that grow earlier on dung of herbivores and that *P. anserina* may kill by hyphal interference [76]).

To evaluate whether the enzymatic potential reflects the ability of *P. anserina* to degrade plant polymeric substrates, growth was monitored on minimal medium plates containing lignin, cellulose, beech wood xylan, apple pectin, inulin and 25 mM sucrose, D-glucose, D-fructose or D-xylose (Figure 7). *P. anserina* did grow on lignin, indicating that it is able to degrade lignin. However, it is suspected that in nature lignin degradation, an energy consuming process, may not be to obtain a carbon source, but mainly to gain access to the (hemi-)cellulose. Growth on cellulose, xylan and D-xylose was significantly faster than on pectin, which agrees with the enzymatic potential based on the genome sequence as described above. No growth was observed on inulin or sucrose, while efficient growth was observed on D-fructose and D-glucose. This is in agreement with the absence of genes required to degrade sucrose and inulin from the genome of *P. anserina*. Overall, these data suggest that *P. anserina* has all the enzymatic complement necessary to efficiently scavenge the carbohydrates it encounters in its natural biotope. Selection has in fact evolved its genome to deal efficiently with these carbon sources, first by duplicating genes involved in cellulose degradation, as shown by the high number of GH61 CDSs, and second by deleting genes required to use carbon sources not commonly encountered (for example, pectin, inulin, and sucrose). This demonstrates the high environmental pressure on evolution as well as the high level of specialization that occurs in the fungal kingdom.

Conclusion

Our analysis of the genome sequence of *P. anserina*, a saprophytic model ascomycete, provides new insights into the genomic evolution of fungi. EST analysis indicates that similar to other eukaryotes, the transcription machinery generates a large array of RNAs with potential regulatory roles. Functional characterization of these RNAs might be one of the most interesting perspectives of this study. Strikingly, in addition to abundant inversions of chromosome segments and gene losses, substantial gene duplications were uncovered. Since this fungus displays a mild RIP, these findings allow us to ask whether the RIP process, when relatively inefficient, might be more of a genome evolution tool rather than a genome defense mechanism.

Moreover, availability of the genome sequence has also already permitted the development of new tools that will bolster research in *P. anserina*. The polymorphic markers designed to plot scaffolds onto the genetic map are now suc-

cessfully used for positional cloning. Gene deletion is facilitated thanks to the availability of the *PaKu70* mutant strain, which greatly enhanced homologous recombination [77], similarly to the deletion of the homologous gene in *N. crassa*, *mus-51* [78]. The identification of the *PaPKS1* gene by a candidate gene approach permits us to envision the design of new genetic tools based on mycelium or ascospore color [63]. The design of microarrays for transcriptome analyses is under way.

As for other saprophytic fungi, the *P. anserina* genome sequence has opened new avenues in the comprehensive study of a variety of biological processes. Of importance is the novel discovery of a large array of *P. anserina* genes potentially involved in lignin and cellulose degradation, some of which may be used for biotechnology applications. It also demonstrates how *P. anserina* is well adapted at the genome level to its natural environment, which was confirmed by the analysis of growth profiles. This result emphasizes the necessity to study several less well-tracked organisms in addition to those well known in the scientific community, as these may yield unexpected new insights into biological phenomena of general interest.

Materials and methods

Strains and culture conditions

The sequenced strain is the S mat+ homokaryotic strain [79]. Culture conditions for this organism were described [61], and currently used methods and culture media can be accessed at the *Podospora anserina* Genome Project web site [80].

Genomic DNA library construction

Nuclear genomic DNA was extracted and separated from mitochondrial DNA as described [81]. Residual mitochondrial DNA present in the preparation was sufficient to allow sequencing of the full mitochondrial DNA circular chromosome. Construction of plasmid DNA libraries was made at Genoscope. The construction of the bacterial artificial chromosome (BAC) library is described in [24].

Construction of cDNA library

Two strategies were used to construct the cDNA libraries. First, a mycelium library was constructed in the yeast expression vector pFL61 [82]. Total RNA was extracted from the s wild-type strain (mat-) and polyA$^+$ RNA was purified twice on oligo (dT)-cellulose columns (mRNA purification kit, Amersham Pharmacia Biotech, GE Healthcare Bio-Sciences AB, Uppsala, Sweden). Anchored dT25 primers were used to obtain double-stranded DNA (cDNA kit, Amersham Pharmacia Biotech, GE Healthcare Bio-Sciences AB, Uppsala, Sweden). Three cDNA libraries, corresponding to three ranges of molecular weight cDNA (0.2-1 kb, 1-2.5 kb, > 2.5 kb) were cloned using BstX1 adaptators in the pFL61 vector between the 5' (promoter) and 3' (terminator) sequences of the *S. cerevisiae pgk1* gene as described previously [82].

Second, total RNA obtained under various physiological conditions (Table 4) was extracted as described [83], using the 'RNeasy Maxi Kit' (Qiagen, Germantown, MD, USA). PolyA$^+$ mRNAs were extracted with the 'Oligotex mRNA Maxi Kit' (Qiagen), reverse transcribed and cloned with the 'cloneMiner cDNA library construction Kit' into plasmid pDONR222 (Invitrogen, Carlsbad, CA, USA).

Sequencing and assembly strategy

The genome of *P. anserina* was sequenced using a 'whole genome shotgun and assembly' strategy. We generated 510,886 individual sequences from two plasmid libraries of 3.3 and 12 kb insert sizes, and from one BAC library of about 90 kb insert size. This corresponds to genome coverage of 9.7-fold. The reads were automatically assembled using Arachne [84], and the initial assembly was improved by eliminating small redundant scaffolds. Additionally, in cases when the genetic map indicated the proximity of two scaffolds (see below), we joined them if there was some additional read pair information between them that was not used by Arachne. Some inter-contig gaps were also filled by placing a contig between two other contigs when matches and read pair information existed and were coherent. The final automatic assembly consisted of 2,784 contigs of N50 size 43 kb, grouped in 728 scaffolds of N50 size 638 kb, for a total genome size (without gaps) of 35.7 Mb. Manual sequence gap filling and removal of contigs corresponding to rDNA genes permitted the decrease of scaffolds and contig numbers to 1,196 contigs clustered into 78 scaffolds.

To connect the genome sequence with the genetic map [85], two approaches were followed. First, sequenced genes, whose positions on the genetic map were known, were mapped by searching the corresponding sequence in the scaffolds, enabling the attribution of some scaffolds to known chromosomes. Second, potential molecular polymorphic markers (microsatellites, minisatellites and indels) were searched and their polymorphisms were assessed in geographic isolates D, E M, T and U. It rapidly appeared that strain T was the genetically most distant strain from strain S, since about three-quarters of tested markers were actually polymorphic between the two strains. A cross between the T and S strains was set up and 51 homokaryotic progenies from this cross were assayed for 120 polymorphic sites scattered onto the 36 largest scaffolds that represented all the coding parts of the genome (except for one putative CDS). Linkage analysis made it possible to define seven linkage groups that were matched with the chromosomes thanks to the already known genes mapped on the sequence by the first approach. Additional polymorphic markers were then used to confirm local assembly, resulting in the new genome map, which contain 325 markers (Additional data file 1). No discrepancy was observed between the established genetic map, the newly defined linkage groups and the sequence assembly. Presently, all but one CDS-containing scaffold are attributed to a chromosome position, although in a few cases orientation of some scaffolds

within the chromosome could not be accurately defined because of their small size. One 33 kb scaffold containing one predicted CDS as well as small scaffolds exclusively made up of repeated sequences are presently not mapped. Collectively, they represent about 1% of the genome.

EMBL accession numbers

Chromosome 1: CU633438; CU633901; CU633867; CU633899; CU633445; CU633897. Chromosome 2: CU633446; CU640366; CU633447. Chromosome 3: CU633448; CU633447; CU633453. Chromosome 4: CU633454; CU633455; CU633456; CU633895. Chromosome 5: CU633457; CU633458; CU633459; CU633866; CU633871; CU607053; CU633461, CU633870, CU633865, CU633876. Chromosome 6: CU633898; CU638744; CU633463, CU633872. Chromosome 7: CU633900; CU633464; CU633873.

Annotation and analysis of genomic sequences

CDSs were annotated by a combination of semi-automatic procedures. First, *P. anserina* open reading frames longer than 20 codons that are evolutionary conserved in *N. crassa* were retrieved by TBLASTN analysis. Candidates with an e-value lower than 10^{-18} were conserved as hypothetical exons. Exons separated by less than 200 nucleotides were merged into putative CDSs and putative introns were predicted thanks to the *P. anserina* consensus sequences defined in the pilot project [24]. Then, 5' and 3' smaller exons were searched by the same procedure except that open reading frames longer than five codons surrounding putative CDSs were analyzed by BLAST with the homologous *N. crassa* region. Candidates with an e-value lower than 10^{-5} were conserved and added to the putative CDSs. CDS and intron predictions were edited with Artemis [86] and manually corrected after comparison with available ESTs. Finally, *ab initio* prediction with GeneID [87] using the *N. crassa* and *Chaetomium globosum* parameter files were performed on regions devoid of annotated features. Manual verification was then applied to improve prediction. This resulted in the definition of 10,545 putative CDSs.

A canonic rDNA unit was assembled. A junction sequence between the left arm of chromosome 3 and an rDNA unit was observed, confirming the position of the cluster on this chromosome based on pulse field electrophoresis data [28]. On the other end of the cluster a junction between an incomplete rDNA repeat and CCCTAA telomeric repeats [88] was detected showing that the cluster is in a subtelomeric position. Similar to the previously investigated filamentous fungi [89], 5S rRNAs were detected by comparison with the *N. crassa* 5S genes. They are encoded by a set of 87 genes, including 72 full-length copies dispersed in the genome. tRNAs were identified with tRNAscan [90]. A total of 361 genes encode the cytosolic tRNA set, which is composed of 48 different acceptor families containing up to 22 members. This set enabled us to decode the 61 sense codons with the classical wobble rule. Other non-coding RNAs were detected with a combination of the Erpin [91], Blast [92] and Yass [93] programs. Homology search included all RNAs contained in the RFAM V.8 [94] and ncRNAdb [95] databases. Any hit from either program with an e-value below 10^{-4} was retained, producing a list of 28 annotated non-coding RNA genes or elements, including 12 spliceosomal RNAs, 15 snoRNAs (mostly of the C/D box class) and one thiamine pyrophosphateriboswitch.

Alignment of EST sequences on the *P. anserina* genome

A two-step strategy was used to align the EST sequences on the *P. anserina* genome. As a first step, BLAST [92] served to generate the alignments between the microsatellite repeat-masked EST sequences and the genomic sequence using the following settings: W = 20, X = 8, match score = 5, mismatch score = -4. The sum of scores of the high-scoring pairs was then calculated for each possible location, then the location with the highest score was retained if the sum of scores was more than 1,000. Once the location of the transcript sequence was determined, the corresponding genomic region was extended by 5 kb on either side. Transcript sequences were then realigned on the extended region using EST_GENOME [96] (mismatch 2, gap penalty 3) to define transcript exons [97]. These transcript models were fused by a single linkage clustering approach, in which transcripts from the same genomic region sharing at least 100 bp are merged [98]. These clusters were used to detect alternative splicing events [99].

Detection, functional annotation and comparative analysis of carbohydrate-active enzymes

Catalytic modules specific to carbohydrate-active enzymes (CAZymes: GHs, glycoside transferases, PLs and CEs) and their ancillary CBMs in fungi were searched by comparison with a library of modules derived from all entries of the Carbohydrate-Active enZymes (CAZy) database [73]. Each protein model was compared with a library of over 100,000 constitutive modules (catalytic modules, CBMs and other non-catalytic modules or domains of unknown function) using BLASTP. Models that returned an e-value passing the 0.1 threshold were automatically sorted and manually analyzed. The presence of the catalytic machinery was verified for distant relatives whenever known in the family. The models that displayed significant similarities were retained for functional annotation and classified in the appropriate classes and families.

Many of the sequence similarity-based families present in CAZy do not coincide with a single substrate or product specificity and, therefore, they are susceptible to grouping proteins with different Enzyme Commission (EC) numbers. Similarly to what has been provided for other genome annotation efforts, we aimed at producing annotations for each protein model that will survive experimental validation, avoiding over-interpretation. A strong similarity to an

enzyme with a characterized activity allows annotation as 'candidate activity', but often for a safe prediction of substrate specificity, annotation such as 'candidate α- or β-glycosidase' may be provided, as the stereochemistry of the α- or β-glycosidic bond is more conserved than the nature of the sugar itself. Each protein model was compared to the manually curated CAZy database, and a functional annotation was assigned according to the relevance. All uncharacterized protein models were thus annotated as 'candidates' or 'related to' or 'distantly related to' their characterized match as a function of their similarity. The overall results of the annotation of the set of CAZymes from *P. anserina* were compared to the content and distribution of CAZymes in several fungal species (Danchin *et al.*, in preparation) in order to identify singularities in the families' distributions and sizes per genome (data not shown). This allowed the identification of significant expansions and reductions of specific CAZyme families in *P. anserina*.

Growth tests

M2 minimal medium contained per liter: 0.25 g KH_2PO_4, 0.3 g K_2HPO_4, 0.25 g $MgSO_4 \cdot 7H_2O$, 0.5 g urea, 0.05 mg thiamine, 0.25 μg biotine and trace elements [100], 12.5 g agar; it was adjusted to pH 7 with KH_2PO_4. Standard M2 contains also 5.5 g/l dextrine, which was replaced by the other tested carbon sources. Sucrose, D-glucose, D-fructose, D-xylose, inulin, Apple pectin, carboxymethyl cellulose and Birchwood xylan were from Sigma-Aldrich (Gillingham, UK) and were added before autoclaving. *P. anserina* was grown for 7 days at 25°C.

Abbreviations

BAC, bacterial artificial chromosome; CAZymes, carbohydrate-active enzymes; CBM, carbohydrate binding module; CDS, coding sequence; CE, carbohydrate esterase; EST, expressed sequence tag; GH, glycoside hydrolase; MSUD, meiotic silencing of unpaired DNA; PKS, polyketide synthase; PL, polysaccharide lyase; RIP, repeat induced point mutation.

Authors' contributions

RD and PS initiated the project. Funding was secured thanks to Genoscope and CNRS. PS coordinated the project. AC, JMA, BS, JP, VA, PW and JW acquired and assembled the sequence. FM, EE, PS, ASC, AB, HK, EC, MDC, MP, VC, SA, AB and CHS contributed to the assembly and juxtaposition of the sequence with the genetic map. BPL, RD and CHS constructed the cDNA libraries. SG and OL developed the bioinformatic tools. MP and CDS analyzed the EST database. OJ and RD performed the synteny analysis. DG identified the non-coding RNA. EE, OL and PS analyzed the repeated sequences and the gain/loss of genes. EE, OL, FM, VBL, RPdV, EB, PMC, EGJD, BH, REK and PS analyzed the genome content. EE, OL, FM, CDS, PW, RPdV, PC, VB, AS, RD and PS contributed to writing the manuscript. All authors read and approved the final manuscript.

Additional data files

The following additional data are available with the online version of this paper. Additional data file 1 is a figure of the *P. anserina* genome map as defined by classic genetic markers and molecular markers, mainly microsatellites, that are polymorphic between strains S and T. Additional data file 2 is a table listing CDSs potentially expressed through frame-shift and read-through. Additional data file 3 is a table listing transposons and transposon-like elements of the *P. anserina* genome. Additional data file 4 is a table listing segmental duplications in the *P. anserina* genome. Additional data file 5 is a table listing CDSs putatively involved in genome protection mechanisms. Additional data file 6 is a list of putative CDSs involved in (hemi-)cellulose and pectin degradation.

Acknowledgements

We thank Anne-Lise Haenni for reading the manuscript and Gaël Lecellier for performing statistical analysis. Sequencing of the genome was funded by Consortium National de Recherche en Génomique, 'séquencage à grande échelle 2002' by CNRS and IFR 115 'Génome: structure, fonction, évolution'. RPdeV and EB were supported by The Netherlands Technology Foundation (STW) VIDI project 07063. DG was supported in part by the ACI-IMPBIO program of the French Research Ministry. The Annie-Sainsard-Chanet laboratory was supported by the 'Centre National de la Recherche Scientifique' and grants from 'Association Française contre les Myopathies'. BPL was a recipient of a fellowship from the Ministere de la Recherche.

References

1. Hedges SB, Blair JE, Venturi ML, Shoe JL: **A molecular timescale of eukaryote evolution and the rise of complex multicellular life.** *BMC Evol Biol* 2004, **4:**2.
2. Hawskworth DL: **The magnitude of fungal diversity: the 1.5 million species revisited.** *Mycol Res* 2001, **105:**1422-1432.
3. Bills GF, Christensen M, Powell M, Thorn G: **Saprobic soil fungi.** In *Biodiversity of the Fungi, Biodiversity and Monitoring Methods* Edited by: Mueller GM, Bills GF, Foster MS. Amsterdam: Elsevier; 2004:271-302.
4. Durrieu G: *Ecologie des Champignons* Paris: Masson; 1993.
5. Money MP: *The Triumph of Fungi: a Rotten History* Oxford: Oxford University press; 2007.
6. Gilbertson RI: **Wood-rotting fungi of north america.** *Mycologia* 1980, **72:**1-49.
7. Spooner B, Roberts P: *Fungi* London: HarperCollins Publishers; 2005.
8. **Carbon Dioxide Information Analysis Center** [http://cdiac.ornl.gov/]
9. Dujon B: **Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution.** *Trends Genet* 2006, **22:**375-387.
10. Cuomo CA, Guldener U, Xu JR, Trail F, Turgeon BG, Di Pietro A, Walton JD, Ma LJ, Baker SE, Rep M, Adam G, Antoniw J, Baldwin T, Calvo S, Chang YL, Decaprio D, Gale LR, Gnerre S, Goswami RS, Hammond-Kosack K, Harris LJ, Hilburn K, Kennell JC, Kroken S, Magnuson JK, Mannhaupt G, Mauceli E, Mewes HW, Mitterbauer R, Muehlbauer G, *et al.*: **The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization.** *Science* 2007, **317:**1400-1402.
11. Dean RA, Talbot NJ, Ebbole DJ, Farman ML, Mitchell TK, Orbach MJ, Thon M, Kulkarni R, Xu JR, Pan H, Read ND, Lee YH, Carbone I, Brown D, Oh YY, Donofrio N, Jeong JS, Soanes DM, Djonovic S, Kolomiets E, Rehmeyer C, Li W, Harding M, Kim S, Lebrun MH, Bohnert H, Coughlan S, Butler J, Calvo S, Ma LJ, *et al.*: **The genome sequence of the rice blast fungus *Magnaporthe grisea*.** *Nature* 2005, **434:**980-986.
12. Kamper J, Kahmann R, Bolker M, Ma LJ, Brefort T, Saville BJ, Banuett

F, Kronstad JW, Gold SE, Muller O, Perlin MH, Wosten HA, de Vries R, Ruiz-Herrera J, Reynaga-Pena CG, Snetselaar K, McCann M, Perez-Martin J, Feldbrugge M, Basse CW, Steinberg G, Ibeas JI, Holloman W, Guzman P, Farman M, Stajich JE, Sentandreu R, Gonzalez-Prieto JM, Kennell JC, Molina L, *et al.*: **Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*.** *Nature* 2006, **444:**97-101.

13. Loftus BJ, Fung E, Roncaglia P, Rowley D, Amedeo P, Bruno D, Vamathevan J, Miranda M, Anderson IJ, Fraser JA, Allen JE, Bosdet IE, Brent MR, Chiu R, Doering TL, Donlin MJ, D'Souza CA, Fox DS, Grinberg V, Fu J, Fukushima M, Haas BJ, Huang JC, Janbon G, Jones SJ, Koo HL, Krzywinski MI, Kwon-Chung JK, Lengeler KB, Maiti R, *et al.*: **The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*.** *Science* 2005, **307:**1321-1324.
14. Nierman WC, Pain A, Anderson MJ, Wortman JR, Kim HS, Arroyo J, Berriman M, Abe K, Archer DB, Bermejo C, Bennett J, Bowyer P, Chen D, Collins M, Coulsen R, Davies R, Dyer PS, Farman M, Fedorova N, Fedorova N, Feldblyum TV, Fischer R, Fosker N, Fraser A, Garcia JL, Garcia MJ, Goble A, Goldman GH, Gomi K, Griffith-Jones S, *et al.*: **Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*.** *Nature* 2005, **438:**1151-1156.
15. Martinez D, Larrondo LF, Putnam N, Gelpke MD, Huang K, Chapman J, Helfenbein KG, Ramaiya P, Detter JC, Larimer F, Coutinho PM, Henrissat B, Berka R, Cullen D, Rokhsar D: **Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78.** *Nat Biotechnol* 2004, **22:**695-700.
16. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S, Rehman B, Elkins T, Engels R, Wang S, Nielsen CB, Butler J, Endrizzi M, Qui D, Ianakiev P, Bell-Pedersen D, Nelson MA, Werner-Washburne M, Selitrennikoff CP, Kinsey JA, Braun EL, Zelter A, Schulte U, Kothe GO, Jedd G, Mewes W, *et al.*: **The genome sequence of the filamentous fungus *Neurospora crassa*.** *Nature* 2003, **422:**859-868.
17. Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J, Baker S, Basham D, Bowman S, Brooks K, Brown D, Brown S, Chillingworth T, Churcher C, Collins M, Connor R, Cronin A, Davis P, Feltwell T, Fraser A, Gentles S, Goble A, Hamlin N, Harris D, Hidalgo J, Hodgson G, Holroyd S, *et al.*: **The genome sequence of *Schizosaccharomyces pombe*.** *Nature* 2002, **415:**871-880.
18. Galagan JE, Calvo SE, Cuomo C, Ma LJ, Wortman JR, Batzoglou S, Lee SI, Basturkmen M, Spevak CC, Clutterbuck J, Kapitonov V, Jurka J, Scazzocchio C, Farman M, Butler J, Purcell S, Harris S, Braus GH, Draht O, Busch S, D'Enfert C, Bouchier C, Goldman GH, Bell-Pedersen D, Griffiths-Jones S, Doonan JH, Yu J, Vienken K, Pain A, Freitag M, *et al.*: **Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*.** *Nature* 2005, **438:**1105-1115.
19. Machida M, Asai K, Sano M, Tanaka T, Kumagai T, Terai G, Kusumoto K, Arima T, Akita O, Kashiwagi Y, Abe K, Gomi K, Horiuchi H, Kitamoto K, Kobayashi T, Takeuchi M, Denning DW, Galagan JE, Nierman WC, Yu J, Archer DB, Bennett JW, Bhatnagar D, Cleveland TE, Fedorova ND, Gotoh O, Horikawa H, Hosoyama A, Ichinomiya M, Igarashi R, *et al.*: **Genome sequencing and analysis of *Aspergillus oryzae*.** *Nature* 2005, **438:**1157-1161.
20. Pel HJ, de Winde JH, Archer DB, Dyer PS, Hofmann G, Schaap PJ, Turner G, de Vries RP, Albang R, Albermann K, Andersen MR, Bendtsen JD, Benen JA, van den Berg M, Breestraat S, Caddick MX, Contreras R, Cornell M, Coutinho PM, Danchin EG, Debets AJ, Dekker P, van Dijck PW, van Dijk A, Dijkhuizen L, Driessen AJ, d'Enfert C, Geysens S, Goosen C, Groot GS, *et al.*: **Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88.** *Nat Biotechnol* 2007, **25:**221-231.
21. Dowding ES: **The sexuality of the normal, giant and dwarf spores of *Pleurage anserina*. (Ces) Kuntze.** *Ann Bot* 1931, **45:**1-14.
22. Rizet G: **Sur l'analyse génétique des asques du *Podospora anserina*.** *C R Acad Sci Paris* 1941, **212:**59-61.
23. Saupe SJ, Clave C, Sabourin M, Begueret J: **Characterization of hch, the *Podospora anserina* homolog of the het-c heterokaryon incompatibility gene of *Neurospora crassa*.** *Curr Genet* 2000, **38:**39-47.
24. Silar P, Barreau C, Debuchy R, Kicka S, Turcq B, Sainsard-Chanet A, Sellem CH, Billault A, Cattolico L, Duprat S, Weissenbach J: **Characterization of the genomic organization of the region bordering the centromere of chromosome V of *Podospora anserina* by direct sequencing.** *Fungal Genet Biol* 2003, **39:**250-263.
25. Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, Nicaud S, Jaffe D, Fisher S, Lutfalla G, Dossat C, Segurens B, Dasilva C, Salanoubat M, Levy M, Boudet N, Castellano S, Anthouard V, Jubin C, Castelli V, Katinka M, Vacherie B, Biemont C, Skalli Z, Cattolico L, Poulain J, *et al.*: **Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype.** *Nature* 2004, **431:**946-957.
26. Hedges SB: **The origin and evolution of model organisms.** *Nat Rev Genet* 2002, **3:**838-849.
27. Kumar S, Hedges B: **A molecular timescale for vertebrate evolution.** *Nature* 1998, **392:**917-920.
28. Javerzat JP, Jacquier C, Barreau C: **Assignment of linkage groups to the electrophoretically-separated chromosomes of the fungus *Podospora anserina*.** *Curr Genet* 1993, **24:**219-222.
29. Osiewacz HD, Clairmont A, Huth M: **Electrophoretic karyotype of the ascomycete *Podospora anserina*.** *Curr Genet* 1990, **18:**481-483.
30. Butler MI, Goodwin TJ, Poulter RT: **Two new fungal inteins.** *Yeast* 2005, **22:**493-501.
31. ***Neurospora crassa* Database** [http://www.broad.mit.edu/annotation/genome/neurospora/Home.html]
32. **The MIPS *Neurospora crassa* database (MNCDB)** [http://mips.gsf.de/projects/fungi/neurospora]
33. Galindo MI, Pueyo JI, Fouix S, Bishop SA, Couso JP: **Peptides encoded by short ORFs control development and define a new eukaryotic gene family.** *PLoS Biol* 2007, **5:**e106.
34. Kondo T, Hashimoto Y, Kato K, Inagaki S, Hayashi S, Kageyama Y: **Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA.** *Nat Cell Biol* 2007, **9:**660-665.
35. Uhler JP, Hertel C, Svejstrup JQ: **A role for noncoding transcription in activation of the yeast PHO5 gene.** *Proc Natl Acad Sci USA* 2007, **104:**8011-8016.
36. Brown DW, Cheung F, Proctor RH, Butchko RA, Zheng L, Lee Y, Utterback T, Smith S, Feldblyum T, Glenn AE, Plattner RD, Kendra DF, Town CD, Whitelaw CA: **Comparative analysis of 87,000 expressed sequence tags from the fumonisin-producing fungus *Fusarium verticillioides*.** *Fungal Genet Biol* 2005, **42:**848-861.
37. Miura F, Kawaguchi N, Sese J, Toyoda A, Hattori M, Morishita S, Ito T: **A large-scale full-length cDNA analysis to explore the budding yeast transcriptome.** *Proc Natl Acad Sci USA* 2006, **103:**17846-17851.
38. Kapranov P, Willingham AT, Gingeras TR: **Genome-wide transcription and the implications for genomic organization.** *Nat Rev Genet* 2007, **8:**413-423.
39. Coppin-Raynal E, Dequard-Chablat M, Picard M: **Genetics of ribosomes and translational accuracy in *Podospora anserina*.** In *Genetics of Translation: New Approaches* Edited by: Tuite M, Picard M, Bolotin-Fukuhara M. Berlin/Heidelberg Springer-Verlag; 1988:431-442.
40. Silar P, Haedens V, Rossignol M, Lalucque H: **Propagation of a novel cytoplasmic, infectious and deleterious determinant is controlled by translational accuracy in *Podospora anserina*.** *Genetics* 1999, **151:**87-95.
41. Silar P, Koll F, Rossignol M: **Cytosolic ribosomal mutations that abolish accumulation of circular intron in the mitochondria without preventing senescence of *Podospora anserina*.** *Genetics* 1997, **145:**697-705.
42. Silar P, Lalucque H, Haedens V, Zickler D, Picard M: ***eEF1A* Controls ascospore differentiation through elevated accuracy, but controls longevity and fruiting body formation through another mechanism in *Podospora anserina*.** *Genetics* 2001, **158:**1477-1489.
43. Allmang C, Krol A: **Selenoprotein synthesis: UGA does not end the story.** *Biochimie* 2006, **88:**1561-1571.
44. Fischer G, Rocha EP, Brunet F, Vergassola M, Dujon B: **Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages.** *PLoS Genet* 2006, **2:**e32.
45. Seoighe C, Federspiel N, Jones T, Hansen N, Bivolarovic V, Surzycki R, Tamse R, Komp C, Huizar L, Davis RW, Scherer S, Tait E, Shaw DJ, Harris D, Murphy L, Oliver K, Taylor K, Rajandream MA, Barrell BG, Wolfe KH: **Prevalence of small inversions in yeast gene order evolution.** *Proc Natl Acad Sci USA* 2000, **97:**14433-14437.
46. Nadeau J, Taylor B: **Lengths of chromosomal segments conserved since divergence of man and mouse.** *Proc Natl Acad Sci USA* 1984, **81:**814-818.
47. Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, De Montigny J, Marck C, Neuveglise C, Talla E, Goffard N, Frangeul

L, Aigle M, Anthouard V, Babour A, Barbe V, Barnay S, Blanchin S, Beckerich JM, Beyne E, Bleykasten C, Boisrame A, Boyer J, Cattolico L, Confanioleri F, De Daruvar A, Despons L, Fabre E, Fairhead C, Ferry-Dumazet H, *et al.*: **Genome evolution in yeasts.** *Nature* 2004, **430:**35-44.

48. Shiu PK, Raju NB, Zickler D, Metzenberg RL: **Meiotic silencing by unpaired DNA.** *Cell* 2001, **107:**905-916.
49. Marcou D, Masson A, Simonet JM, Piquepaille G: **Evidence for non-random spatial distribution of meiotic exchanges in *Podospora anserina* : comparison between linkage groups I and 6.** *Mol Gen Genet* 1979, **176:**67-79.
50. Gallegos A, Jacobson DJ, Raju NB, Skupski MP, Natvig DO: **Suppressed recombination and a pairing anomaly on the mating-type chromosome of *Neurospora tetrasperma*.** *Genetics* 2000, **154:**623-633.
51. Fraser JA, Diezmann S, Subaran RL, Allen A, Lengeler KB, Dietrich FS, Heitman J: **Convergent evolution of chromosomal sex-determining regions in the animal and fungal kingdoms.** *PLOS Biol* 2004, **2:**e384.
52. Hua-Van A, Daviere JM, Kaper F, Langin T, Daboussi MJ: **Genome organization in *Fusarium oxysporum* : clusters of class II transposons.** *Curr Genet* 2000, **37:**339-347.
53. Galagan JE, Selker EU: **RIP: the evolutionary cost of genome defense.** *Trends Genet* 2004, **20:**417-423.
54. Graia F, Lespinet O, Rimbault B, Dequard-Chablat M, Coppin E, Picard M: **Genome quality control: RIP (repeat-induced point mutation) comes to *Podospora*.** *Mol Microbiol* 2001, **40:**586-595.
55. Bouhouche K, Zickler D, Debuchy R, Arnaise S: **Altering a gene involved in nuclear distribution increases the repeat-induced point mutation process in the fungus *Podospora anserina*.** *Genetics* 2004, **167:**151-159.
56. Fulci V, Macino G: **Quelling: post-transcriptional gene silencing guided by small RNAs in *Neurospora crassa*.** *Curr Opin Microbiol* 2007, **10:**199-203.
57. Begueret J, Razanamparany V, Perrot M, Barreau C: **Cloning gene ura5 for the orotidylic acid pyrophosphorylase of the filamentous fungus *Podospora anserina* : transformation of protoplasts.** *Gene* 1984, **32:**487-492.
58. Kouzminova E, Selker EU: **dim-2 encodes a DNA methyltransferase responsible for all known cytosine methylation in *Neurospora*.** *EMBO J* 2001, **20:**4309-4323.
59. Freitag M, Williams RL, Kothe GO, Selker EU: **A cytosine methyltransferase homologue is essential for repeat-induced point mutation in *Neurospora crassa*.** *Proc Natl Acad Sci USA* 2002, **99:**8802-8807.
60. Freitag M, Hickey PC, Khlafallah TK, Read ND, Selker EU: **HP1 is essential for DNA methylation in *Neurospora*.** *Mol Cell* 2004, **13:**427-434.
61. Rizet G, Engelmann C: **Contribution à l'étude génétique d'un Ascomycète tétrasporé : *Podospora anserina* (Ces.) Rehm.** *Rev Cytol Biol Veg* 1949, **11:**201-304.
62. Raju NB, Perkins DD, Newmeyer D: **Genetically determined nonselective abortion of asci in *Neurospora crassa*.** *Can J Botany* 1987, **65:**1539-1549.
63. Coppin E, Silar P: **Identification of PaPKS1, a polyketide synthase involved in melanin formation and its utilization as a genetic tool in *Podospora anserina*.** *Mycol Res* 2007, **111:**901-908.
64. Malagnac F, Klapholz B, Silar P: **PaTrx1 and PaTrx3, two cytosolic thioredoxins of the filamentous ascomycete *Podospora anserina* involved in sexual development and cell degeneration.** *Eukaryot Cell.* 2007, **6:**2323-2331.
65. Paoletti M, Saupe SJ, Clave C: **Genesis of a fungal non-self recognition repertoire.** *PLoS ONE* 2007, **2:**e283.
66. Wapinski I, Pfeffer A, Friedman N, Regev A: **Natural history and evolutionary principles of gene duplication in fungi.** *Nature* 2007, **449:**54-61.
67. Webster J: **Coprophilous fungi.** *Trans Br Mycol Soc* 1970, **54:**161-180.
68. Martinez AT, Speranza M, Ruiz-Duenas FJ, Ferreira P, Camarero S, Guillen F, Martinez MJ, Gutierrez A, del Rio JC: **Biodegradation of lignocellulosics: microbial, chemical, and enzymatic aspects of the fungal attack of lignin.** *Int Microbiol* 2005, **8:**195-204.
69. ten Have R, Teunissen PJ: **Oxidative mechanisms involved in lignin degradation by white-rot fungi.** *Chem Rev* 2001, **101:**3397-3413.
70. Wesenberg D, Kyriakides I, Agathos SN: **White-rot fungi and their enzymes for the treatment of industrial dye effluents.** *Biotechnol Adv* 2003, **22:**161-187.
71. Cavener DR: **GMC oxidoreductases. A newly defined family of homologous proteins with diverse catalytic activities.** *J Mol Biol* 1992, **223:**811-814.
72. Pointing SB, Parungao MM, Hyde KD: **Production of wood-decay enzymes, mass loss and lignin solubilization in wood by tropical Xylariaceae.** *Mycol Res* 2003, **107:**231-235.
73. **CAZy~Carbohydrate-Active enZymes** [http://www.cazy.org/]
74. Ishii T: **Isolation and characterization of a diferuloyl arabinoxylan hexasaccharide from bamboo shoot cell-walls.** *Carbohydr Res* 1991, **219:**15-22.
75. Imamura T, Watanabe T, Kuwahara M, Koshijima. T: **Ester linkages between lignin and glucuronic acid in lignin-carbohydrate complexes from *Fagus crenata*.** *Phytochem* 1994, **37:**1165-1173.
76. Silar P: **Peroxide accumulation and cell death in filamentous fungi inudced by contact with a contestant.** *Mycol Res* 2005, **109:**137-149.
77. El-Khoury R, Sellem CH, Coppin E, Boivin A, Maas MFPM, Debuchy R, Sainsard-Chanet A: **Gene deletion and allelic replacement in the filamentous fungus *Podospora anserina*.** *Curr Genet* 2008, **53:**249-258.
78. Ninomiya Y, Suzuki K, Ishii C, Inoue H: **Highly efficient gene replacements in *Neurospora* strains deficient for nonhomologous end-joining.** *Proc Natl Acad Sci USA* 2004, **101:**12248-12253.
79. Rizet G: **Les phénomènes de barrage chez *Podospora anserina*. I. Analyse génétique des barrages entre souches S and s.** *Rev Cytol Biol Veg* 1952, **13:**51-92.
80. ***Podospora anserina* Genome Project** [http://podospora.igmors.u-psud.fr]
81. Cummings DJ, Belcour L, Grandchamp C: **Mitochondrial DNA from *Podospora anserina*. I. Isolation and characterization.** *Mol Gen Genet* 1979, **171:**229-238.
82. d'Enfert C, Minet M, Lacroute F: **Cloning plant genes by complementation of yeast mutants.** *Methods Cell Biol* 1995, **49:**417-430.
83. Chomczynski P, Sacchi N: **Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction.** *Anal Biochem* 1987, **162:**156-159.
84. Jaffe DB, Butler J, Gnerre S, Mauceli E, Lindblad-Toh K, Mesirov JP, Zody MC, Lander ES: **Whole-genome sequence assembly for mammalian genomes: Arachne 2.** *Genome Res* 2003, **13():**91-96.
85. Marcou D, Picard-Bennoun M, Simonet JM: **Genetic Map of *Podospora anserina*.** In *Genetic Maps* 6th edition. Edited by: O'Brien S. Cold Spring Harbor: Cold Spring Harbor laboratory Press; 1993:3.92-3.101.
86. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16:**944-945.
87. Parra G, Blanco E, Guigo R: **GeneID in *Drosophila*.** *Genome Res* 2000, **10:**511-515.
88. Javerzat JP, Bhattacherjee V, Barreau C: **Isolation of telomeric DNA from the filamentous fungus *Podospora anserina* and construction of a self-replicating linear plasmid showing high transformation frequency.** *Nucleic Acids Res* 1993, **21:**497-504.
89. Rooney AP, Ward TJ: **Evolution of a large ribosomal RNA multigene family in filamentous fungi: birth and death of a concerted evolution paradigm.** *Proc Natl Acad Sci USA* 2005, **102:**5084-5089.
90. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res* 1997, **25:**955-964.
91. Gautheret D, Lambert A: **Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles.** *J Mol Biol* 2001, **313:**1003-1011.
92. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
93. Noe L, Kucherov G: **YASS: enhancing the sensitivity of DNA similarity search.** *Nucleic Acids Res* 2005, **33:**W540-W543.
94. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A: **Rfam: annotating non-coding RNAs in complete genomes.** *Nucleic Acids Res* 2005, **33:**D121-D124.
95. Szymanski M, Erdmann VA, Barciszewski J: **Noncoding RNAs database (ncRNAdb).** *Nucleic Acids Res* 2007, **35:**D162-D164.
96. Mott R: **EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA.** *Comput Appl Biosci* 1997, **13:**477-478.
97. Castelli V, Aury J-M, Jaillon O, Wincker P, Clepet C, Menard M, Cruaud C, Quetier F, Scarpelli C, Schachter V, Temple G, Caboche M,

Weissenbach J, Salanoubat M: **Whole genome sequence comparisons and 'full-length' cDNA sequences: a combined approach to evaluate and improve *Arabidopsis* genome annotation.** *Genome Res* 2004, **14:**406-413.
98. Porcel BM, Delfour O, Castelli V, De Berardinis V, Friedlander L, Cruaud C, Ureta-Vidal A, Scarpelli C, Wincker P, Schachter V, Saurin W, Gyapay G, Salanoubat M, J. W: **Numerous novel annotations of the human genome sequence supported by a 5'-end-enriched cDNA collection.** *Genome Res* 2004, **14:**463-471.
99. Thill G, Castelli V, Pallud S, Salanoubat M, Wincker P, De la Grange P, Auboeuf D, Schachter V, Weissenbach J: **ASEtrap: a biological method for speeding up the exploration of spliceomes.** *Genome Res* 2006, **16:**776-786.
100. Vishniac W, Santer M: **The thiobacilli.** *Bacteriol Rev* 1957, **21:**195-213.
101. Do C, Mahabhashyam M, Brudno M, Batzoglou S: **ProbCons: probabilistic consistency-based multiple sequence alignment.** *Genome Res* 2005, **15:**330-340.
102. Guindon S, Gascuel O: **Simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol* 2003, **52:**696-704.
103. Rizet G: **Sur l'impossibilité d'obtenir la multiplication végétative ininterrompue et illimitée de l'Ascomycète *Podospora anserina*.** *C R Acad Sci* 1953, **237:**838-840.
104. Hamann A, Brust D, Osiewacz HD: **Deletion of putative apoptosis factors leads to lifespan extension in the fungal ageing model *Podospora anserina*.** *Mol Microbiol* 2007, **65:**948-958.
105. Sellem CH, Marsy S, Boivin A, Lemaire C, Sainsard-Chanet A: **A mutation in the gene encoding cytochrome c1 leads to a decreased ROS content and to a long-lived phenotype in the filamentous fungus *Podospora anserina*.** *Fungal Genet Biol* 2007, **44:**648-658.
106. Kicka S, Bonnet C, Sobering AK, Ganesan LP, Silar P: **A mitotically inheritable unit containing a MAP kinase module.** *Proc Natl Acad Sci USA* 2006, **103:**13445-13450.
107. Dementhon K, Saupe SJ: **DNA-binding specificity of the IDI-4 basic leucine zipper factor of *Podospora anserina* defined by systematic evolution of ligands by exponential enrichment (SELEX).** *Eukaryot Cell* 2005, **4:**476-483.
108. Picard M, Debuchy R, Coppin E: **Cloning the mating types of the heterothallic fungus *Podospora anserina* : developmental features of haploid transformants carrying both mating types.** *Genetics* 1991, **128:**539-547.
109. Coppin E, de Renty C, Debuchy R: **The function of the coding sequences for the putative pheromone precursors in *Podospora anserina* is restricted to fertilization.** *Eukaryot Cell* 2005, **4:**407-420.
110. Nguyen Hv: **Rôle des facteurs internes et externes dans la manifestation de rythmes de croissance chez l'ascomycète *Podospora anserina*.** *C R Acad Sci Paris* 1962, **254:**2646-2648.
111. Jamet-Vierny C, Debuchy R, Prigent M, Silar P: **IDC1, a Pezizomycotina-specific gene that belongs to the PaMpk1 MAP kinase transduction cascade of the filamentous fungus *Podospora anserina*.** *Fungal Genet Biol* 2007, **44:**1219-1230.
112. Mannot F: **Sur la localisation du gène S et sur quelques particularités du crossing-over chez *Podospora anserina*.** *C R Acad Sci Paris* 1953, **236:**2330-2332.
113. Padieu E, Bernet J: **Mode d'action des gènes responsables de l'avortement de certains produits de la méiose chez l'Ascomycète *Podospora anserina*.** *C R Acad Sci* 1967, **264:**2300-2303.
114. Picard M: **Genetic evidence for a polycistronic unit of transcription in the complex locus '14' in *Podospora anserina*. II. Genetic analysis of informational suppressors.** *Genet Res Camb* 1973, **21:**1-15.
115. Dequard-Chablat M, Silar P: ***Podospora anserina AS6* gene encodes the cytosolic ribosomal protein of the *E. coli* S12 family.** *Fung Genet Newslett* 2006, **53:**26-29.
116. Tudzynski P, Esser K: **Inhibitors of mitochondrial function prevent senescence in the ascomycete *Podospora anserina*.** *Molec gen Genet* 1977, **153:**111-113.
117. Belcour L, Begel O, Duchiron F, Lecomte P: **Four mitochondrial loci in *Podospora anserina*.** *Neurospora Newsl* 1978, **25:**26-27.
118. Berteaux-Lecellier V, Picard M, Thompson-Coffe C, Zickler D, Panvier-Adoutte A, Simonet JM: **A nonmammalian homolog of the *PAF1* gene (Zellweger syndrome) discovered as a gene involved in caryogamy in the fungus *Podospora anserina*.** *Cell* 1995, **81:**1043-1051.
119. Bonnet C, Espagne E, Zickler D, Boisnard S, Bourdais A, Berteaux-Lecellier V: **The peroxisomal import proteins PEX2, PEX5 and PEX7 are differently involved in *Podospora anserina* sexual cycle.** *Mol Microbiol* 2006, **62:**157-169.
120. Schecroun J: **Sur la nature de la différence cytoplasmique entre souches s and sS de *Podospora anserina*.** *C R Acad Sci Paris* 1959, **248:**1394-1397.
121. Coustou V, Deleu C, Saupe S, Bégueret J: **The protein product of the *het-s* heterokaryon incompatibility gene of the fungus *Podospora anserina* behaves as a prion analog.** *Proc Natl Acad Sci USA* 1997, **94:**9773-9778.
122. Seshime Y, Juvvadi PR, Fujii I, Kitamoto K: **Discovery of a novel superfamily of type III polyketide synthases in *Aspergillus oryzae*.** *Biochem Biophys Res Commun* 2005, **331:**253-260.
123. **Information about Neurospora** [http://www.fgsc.net/Neurospora/neurospora.html]
124. Varela E, Jesus Martinez M, Martinez AT: **Aryl-alcohol oxidase protein sequence: a comparison with glucose oxidase and other FAD oxidoreductases.** *Biochim Biophys Acta* 2000, **1481:**202-208.
125. Zamocky M, Ludwig R, Peterbauer C, Hallberg BM, Divne C, Nicholls P, Haltrich D: **Cellobiose dehydrogenase -a flavocytochrome from wood-degrading, phytopathogenic and saprotropic fungi.** *Curr Protein Pept Sci* 2006, **7:**255-280.
126. Giffhorn F: **Fungal pyranose oxidases: occurrence, properties and biotechnical applications in carbohydrate chemistry.** *Appl Microbiol Biotechnol* 2000, **54:**727-740.
127. Whittaker JW: **Galactose oxidase.** *Adv Protein Chem* 2002, **60:**1-49.
128. Vanden Wymelenberg A, Sabat G, Mozuch M, Kersten PJ, Cullen D, Blanchette RA: **Structure, organization, and transcriptional regulation of a family of copper radical oxidase genes in the lignin-degrading basidiomycete *Phanerochaete chrysosporium*.** *Appl Environ Microbiol* 2006, **72:**4871-4877.
129. Jensen KA Jr, Ryan ZC, Vanden Wymelenberg A, Cullen D, Hammel KE: **An NADH:quinone oxidoreductase active during biodegradation by the brown-rot basidiomycete *Gloeophyllum trabeum*.** *Appl Environ Microbiol* 2002, **68:**2699-2703.
130. Baldrian P: **Fungal laccases - occurrence and properties.** *FEMS Microbiol Rev* 2006, **30:**215-242.
131. Ruiz-Duenas FJ, Camarero S, Perez-Boada M, Martinez MJ, Martinez AT: **A new versatile peroxidase from *Pleurotus*.** *Biochem Soc Trans* 2001, **29:**116-122.
132. Malagnac F, Wendel B, Goyon C, Faugeron G, Zickler D, Rossignol JL, Noyer-Weidner M, Vollmayr P, Trautner TA, Walter J: **A gene essential for *de novo* methylation and development in *Ascobolus* reveals a novel type of eukaryotic DNA methyltransferase structure.** *Cell* 1997, **91:**281-290.
133. Tamaru H, Selker EU: **A histone H3 methyltransferase controls DNA methylation in *Neurospora crassa*.** *Nature* 2001, **414:**277-283.
134. Jackson JP, Lindroth AM, Cao X, Jacobsen SE: **Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase.** *Nature* 2002, **416:**556-560.
135. Malagnac F, Bartee L, Bender J: **An *Arabidopsis* SET domain protein required for maintenance but not establishment of DNA methylation.** *EMBO J* 2002, **21:**6842-6852.
136. Cogoni C, Macino G: **Gene silencing in *Neurospora crassa* requires a protein homologous to RNA-dependent RNA polymerase.** *Nature* 1999, **399:**166-169.
137. Catalanotto C, Azzalin G, Macino G, Cogoni C: **Gene silencing in worms and fungi.** *Nature* 2000, **404:**245.
138. Cogoni C, Macino G: **Posttranscriptional gene silencing in *Neurospora* by a RecQ DNA helicase.** *Science* 1999, **286:**2342-2344.
139. Catalanotto C, Pallotta M, ReFalo P, Sachs MS, Vayssie L, Macino G, Cogoni C: **Redundancy of the two dicer genes in transgene-induced posttranscriptional gene silencing in *Neurospora crassa*.** *Mol Cell Biol* 2004, **24:**2536-2545.
140. Maiti M, Lee HC, Liu Y: **QIP, a putative exonuclease, interacts with the *Neurospora* Argonaute protein and facilitates conversion of duplex siRNA into single strands.** *Genes Dev* 2007, **21:**590-600.
141. Shiu PK, Zickler D, Raju NB, Ruprich-Robert G, Metzenberg RL: **SAD-2 is required for meiotic silencing by unpaired DNA and perinuclear localization of SAD-1 RNA-directed RNA polymerase.** *Proc Natl Acad Sci USA* 2006, **103:**2243-2248.
142. Malagnac F, Gregoire A, Goyon C, Rossignol JL, Faugeron G: **Masc2, a gene from *Ascobolus* encoding a protein with a DNA-methyltransferase activity *in vitro*, is dispensable for *in vivo* methylation.** *Mol Microbiol* 1999, **31:**331-338.

143. Saze H, Mittelsten Scheid O, Paszkowski J: **Maintenance of CpG methylation is essential for epigenetic inheritance during plant gametogenesis.** *Nat Genet* 2003, **34:**65-69.
144. Hermann A, Goyal R, Jeltsch A: **The Dnmt1 DNA-(cytosine-C5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites.** *J Biol Chem* 2004, **279:**48350-48359.
145. Bartee L, Malagnac F, Bender J: ***Arabidopsis*** **cmt3 chromomethylase mutations block non-CG methylation and silencing of an endogenous gene.** *Genes Dev* 2001, **15:**1753-1758.
146. Lindroth AM, Cao X, Jackson JP, Zilberman D, McCallum CM, Henikoff S, Jacobsen SE: **Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation.** *Science* 2001, **292:**2077-2080.
147. Cao X, Jacobsen SE: **Role of the *Arabidopsis* DRM methyltransferases in *de novo* DNA methylation and gene silencing.** *Curr Biol* 2002, **12:**1138-1144.
148. Okano M, Bell DW, Haber DA, Li E: **DNA methyltransferases Dnmt3a and Dnmt3b are essential for *de novo* methylation and mammalian development.** *Cell* 1999, **99:**247-257.
149. Aufsatz W, Mette MF, van der Winden J, Matzke M, Matzke AJ: **HDA6, a putative histone deacetylase needed to enhance DNA methylation induced by double-stranded RNA.** *EMBO J* 2002, **21:**6832-6841.
150. Earley K, Lawrence RJ, Pontes O, Reuther R, Enciso AJ, Silva M, Neves N, Gross M, Viegas W, Pikaard CS: **Erasure of histone acetylation by *Arabidopsis* HDA6 mediates large-scale gene silencing in nucleolar dominance.** *Genes Dev* 2006, **20:**1283-1293.
151. Lawrence RJ, Earley K, Pontes O, Silva M, Chen ZJ, Neves N, Viegas W, Pikaard CS: **A concerted DNA methylation/histone methylation switch regulates rRNA gene dosage control and nucleolar dominance.** *Mol Cell* 2004, **13:**599-609.
152. Hickman M, McCullough K, Woike A, Raducha-Grace L, Rozario T, Dula ML, Anderson E, Margalit D, Holmes SG: **Isolation and characterization of conditional alleles of the yeast SIR2 gene.** *J Mol Biol* 2007, **367:**1246-1257.
153. Jeddeloh JA, Stokes TL, Richards EJ: **Maintenance of genomic methylation requires a SWI2/SNF2-like protein.** *Nat Genet* 1999, **22:**94-97.
154. Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V, May B, Kasschau KD, Carrington JC, Doerge RW, Colot V, Martienssen R: **Role of transposable elements in heterochromatin and epigenetic control.** *Nature* 2004, **430:**471-476.

BMC Genomics

DATABASE Open Access

FUNGIpath: a tool to assess fungal metabolic pathways predicted by orthology

Sandrine Grossetête, Bernard Labedan, Olivier Lespinet*

Abstract

Background: More and more completely sequenced fungal genomes are becoming available and many more sequencing projects are in progress. This deluge of data should improve our knowledge of the various primary and secondary metabolisms of Fungi, including their synthesis of useful compounds such as antibiotics or toxic molecules such as mycotoxins. Functional annotation of many fungal genomes is imperfect, especially of genes encoding enzymes, so we need dedicated tools to analyze their metabolic pathways in depth.

Description: FUNGIpath is a new tool built using a two-stage approach. Groups of orthologous proteins predicted using complementary methods of detection were collected in a relational database. Each group was further mapped on to steps in the metabolic pathways published in the public databases KEGG and MetaCyc. As a result, FUNGIpath allows the primary and secondary metabolisms of the different fungal species represented in the database to be compared easily, making it possible to assess the level of specificity of various pathways at different taxonomic distances. It is freely accessible at http://www.fungipath.u-psud.fr.

Conclusions: As more and more fungal genomes are expected to be sequenced during the coming years, FUNGIpath should help progressively to reconstruct the ancestral primary and secondary metabolisms of the main branches of the fungal tree of life and to elucidate the evolution of these ancestral fungal metabolisms to various specific derived metabolisms.

Background

Currently, the Fungi have more published nuclear genome sequences than any other eukaryotic taxonomic group [1]. This relative abundance (28 genomes in May 2009) can be explained by their economic significance and their moderate genome size [Additional File 1]. Since several species are model organisms for fundamental, medical, or agronomical and industrial studies (e.g. *Saccharomyces cerevisiae*, *Candida albicans*, *Yarrowia lipolytica*), fungal genomes seem suitable for large-scale comparative studies, which will allow their evolution to be elucidated [2-4]. Several teams [5-7] have already performed extensive comparisons of a few fungal genomes to predict groups of orthologous proteins, using published methods such as Inparanoid [8], OrthoMCL [9] or TribeMcl [10].

However, current information about the number of fungal enzymes involved in metabolic pathways is rather scanty and is heterogeneously distributed in major public curated databases, both universal (Swiss-Prot [11]) and specialized (KEGG [12] and MetaCyc [13]). To perform an extensive comparison of these various databases containing enzymatic information we propose to identify each enzyme by its ID-EC, which associates its protein identifier (ID) with the EC number allocated by the IUBMB [14]. Table 1 illustrates this paucity of knowledge; it shows the respective distributions per species in both protein databases [11] and pathway databases [12,13] of ID-ECs and their respective medians in the animal, plant and fungal kingdoms. Swiss-Prot displays as many as 335 fungal species containing at least one ID-EC, but their median values are as low as two ID-ECs per species (Table 1). In contrast, only 27 fungal species are included in KEGG (which is restricted to complete published genomes), but their median values are as high as 855 ID-ECs per species (Table 1). This contrast is mainly because the public databases surveyed in Table 1 include data on *S. cerevisiae*, which is among the three best fungal genomes correctly annotated at the

* Correspondence: olivier.lespinet@igmors.u-psud.fr
Institut de Génétique et de Microbiologie, Université Paris-Sud 11, CNRS UMR 8621, Bâtiment 400, 91405 Orsay Cedex, France

Table 1 Distribution of ID-EC per kingdom in public databases

	Number of species displaying IDs annotated with EC number (ID-EC)			Median value of the set of ID-EC found per species		
	KEGG	MetaCyc	Swiss-Prot	KEGG	MetaCyc	Swiss-Prot
Animal	38	7	1252	2021	2	1
Fungi	27	14	335	855	3	2
Plant	6	158	915	1051	2	1

enzymatic level (data not shown). Most other fungi have limited or null functional annotation, explaining why the median values are so low in MetaCyc and Swiss-Prot.

This remarkable situation arises largely because there is currently no tool for large-scale analyses of fungal metabolism, except for a preliminary attempt to identify enzymes in pathogenic fungi for a limited number of metabolic pathways [15]. To cope with this major shortcoming, we designed a tool that allows us to mine genomic data by combining two complementary approaches: (i) defining reliable groups of orthologous proteins and (ii) mapping these groups on to the metabolic pathways that are described in KEGG [12] and MetaCyc [13].

Organizing relevant data for analyzing fungal metabolic pathways

Identifying enzyme activities requires relevant prediction of orthologs

As more and more genomic data become available, homology can be used to reconstruct the metabolic pathways of newly-sequenced organisms, taking the pathways of well-studied model organisms such as yeast as reference. Accordingly, one must identify the amino acid sequences encoding each step of each pathway in organisms that have not been studied experimentally [16-18]. However, there are two major drawbacks in this transfer of information. First, the accuracy of functional annotation of many fungal genomes is low because experimental data are lacking except in the case of yeast [19-21]. Secondly, it is difficult to predict reliable orthologs among all the putative homologs detected during exhaustive comparison of pairs of genomes. Numerous methods have been published but none appears completely infallible (for a recent review, see [22]). Thus, we decided to apply independent methods to the same dataset, collect as many potential orthologs as possible, and then compute their overlap. Exploring several methods raised the probability of finding consistent groups corresponding to this overlap. Accordingly, we used three different and complementary approaches based on similarity searches, and another based on the analysis of phylogenetic trees of families of homologs.

Searching pertinent orthologs

First, two published methods were used with their respective default parameters. Inparanoid [8] allows us to identify the orthologs and the inparalogs (genes duplicated since the last speciation event) during pairwise genome comparison. OrthoMCL [9] permits consistent strongly-related groups of orthologs (including inparalogs) to be identified.

Secondly, we improved the classical all-versus-all BLASTP [23] approach to identifying pairs of best reciprocal hits (BRH) [24] with a dedicated Perl script, enhancing the definition of orthologs by specifying two parameters, the alignment percentage and the score ratio, to filter the BLAST results. Local conservation was avoided by dividing the alignment length of each aligned sequence by its total length. The score ratio is defined as the ratio of the raw BLAST score computed by aligning a pair of sequences to the raw score of each sequence against itself (i.e. maximum score). Only results with score ratios over 0.2 and alignment percentages above 60% were kept for further studies.

These different methods based on sequence similarity yield various clusters of orthologous proteins that are more or less stringent depending whether single (e.g. Inparanoid) or multiple (e.g. BRH [Additional File 2]) links are used to build the orthologous protein groups.

Besides these methods based on similarity approaches, methods based on phylogenetic analysis have recently been developed to build orthologous groups [25,26]. Here we chose a phylogenetic approach we had previously developed [25] to obtain groups of orthologous proteins, using automated analysis of trees of families of homologous proteins without a reference tree. The homologous proteins were first detected using BLASTP [23] with the following constraints: an E-value less than 0.001 and an alignment extending for at least 70% of the length of the shorter matching protein. For each family, a multiple alignment was built with Muscle [27], and the phylogenetic tree deduced was reconstructed using PhyML [28]. The program Retree from the Phylip package [29] was further used to root the tree in order to distinguish orthologs from paralogs using automatic tree analysis [25].

Table 2 shows a strikingly low overlap between the results obtained by applying these four methods to the 20 fungal genomes under study. The first column shows that the highest number of groups of orthologs is obtained with the BRH method. However, this may be partly artifactual since BRH is the only method in which

Table 2 Groups of orthologous proteins for the 20 genomes available in FUNGIpath predicted by four different methods

Total Number	Relative percentage sharing between two methods				
		BRH	Inparanoid	OrthoMCL	Phylogeny
52292	BRH	-	4.8%	3.7%	5.8%
18235	Inparanoid	8.0%	-	22.4%	8.5%
20379	OrthoMCL	12.4%	16.3%	-	8.3%
12676	Phylogeny	32.4%	25.9%	32.6%	-

upper triangular matrix: percentage of identical groups
lower triangular matrix: percentage of specific groups

an amino acid sequence can belong to different groups owing to the formation of multiple links [Additional File 2]. Columns three to six show the relative percentages shared among the different methods as a double matrix. The upper matrix shows that the relative percentages of identical groups are generally low; the highest figure is 22.4% (common to the OrthoMCL and Inparanoid outputs). The lower matrix shows the percentage of groups that are unique to one of the two methods compared. The low figures obtained (ranging from 8 to 32.6%) suggest that each method brings specific information. The highest specificities are found with the phylogenetic approach, which is indeed the most distinctive of the four approaches we used.

Identifying biologically relevant groups of orthologs

Although the overlap between these different methods for detecting orthologs appears narrow, we tried to build a consensus of the groups of orthologs using both union and intersection methods. Consideration of all the orthologs found merged large numbers of proteins (2,694 proteins in the largest group), with a trend towards amalgamating sometimes quite distant groups of orthologs. On the other hand, computing the crude intersection of the different methods also seemed inadequate (32 proteins in the largest group), since the BRH approach does not detect the inparalogs found by the other methods.

To cope with these difficulties, we modified the intersection approach, using a two-step strategy based on enrichment of the reference groups, i.e. the groups of orthologs obtained by the crude intersection approach. Fig. 1 shows a flowchart of our approach. (i) For each reference group, the sequences were aligned [27] and their corresponding HMM profile was computed using the HMMER hmmbuild and hmmcalibrate programs [30]. To avoid any bias due to the numerous inparalogs present in some species, only one homologous gene per genome was conserved as the reference ortholog building the individual HMM profiles. (ii) All the computed HMM profiles were organized as a database, and each sequence not included in any reference group was further compared to the database using the HMMER hmmpfam program [30] in order to add it to a reference group using stringent threshold. Indeed, to build sound *final* groups, we limited the assignment of a sequence to a reference group if the E-value was less than a threshold of 10^{-10} [Additional File 3]. This stringent criterion allowed a good balance to be kept between sensitivity (29.2% of the sequences initially not associated with a reference group were now associated with one) and specificity (64.2% of the sequences initially found by at least one method but associated with several groups of orthologs were now associated with only one).

In total, we obtained 12,850 final groups of orthologs (size ≥ 2) that appear biologically relevant, the largest group containing 297 sequences (see the size distribution in [Additional File 4]). These figures suggest a good compromise when compared with the values obtained using the crude union and intersection methods (Table 3). With such a prediction, 57% of the total sequences were associated with a group of orthologs [Additional Files 5, 6, and 7]. The mean number of homologous proteins per genome is close to 1.3 in all final groups, suggesting that the orthology/paralogy relationships are quite well resolved by our enrichment approach. Comparison of our ortholog predictions with the four initial methods (Table 4) shows that our approach gives results different from each separate method. The highest number of identical groups with FUNGIpath is obtained with OrthoMCL, whereas the lowest percentage of specific groups is obtained with Inparanoid and BRH.

Assessing the reliability of the predicted final groups of orthologs

To ascertain the reliability of our predictions further, we computed a confidence score *S* for each final group of orthologous proteins, as follows:

$$S = \frac{10}{m} \sum_{i=1}^{m} \frac{I_{F,i}}{O_F \cdot G_i}$$

where m is the number of methods used for orthology prediction, $I_{F,\,i}$ is the number of orthologs shared (intersection) between the result of method i and the final group of orthologs, O_F is the number of orthologs in the final group and G_i is the number of groups obtained by method i for the set of proteins composing the final group. This confidence score is based on the assumption that the reliability of a final group increases with the number of independent methods that find it. Thus, if method i predicts the attested group, the score is 1. If not, the score is greater than 0 and less than or equal to 1. The average score (computed as the sum of scores for each method divided by the total number of methods m) was scaled from 0-10 by multiplying by 10; the

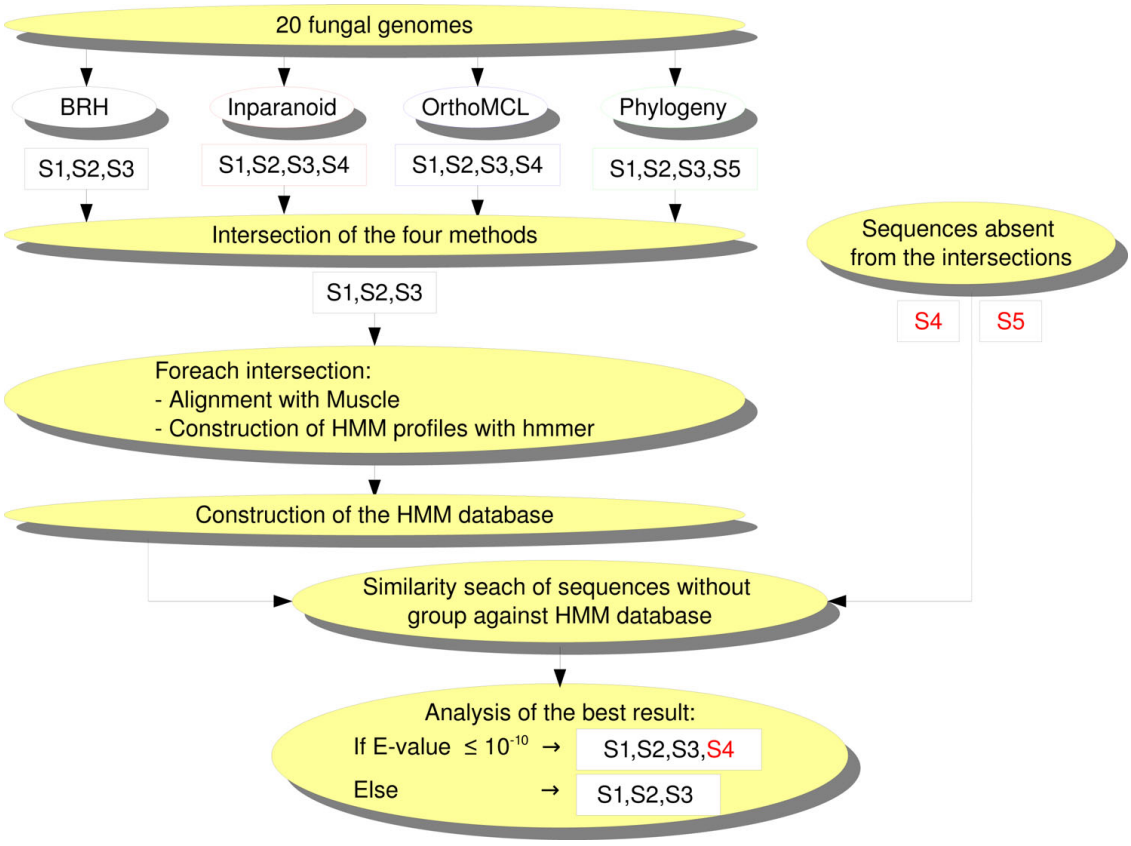


Figure 1 Flow chart of the construction of orthologous groups. The diagram represents an overview of our annotation process. We applied four methods of orthologous predictions (BRH, Inparanoid, OrthoMCL and Phylogeny) to 20 fungal genomes. We compared the different groups (represented in the diagram by the sequences S1, S2, S3, S4 and S5) and we constructed their intersections (represented in the diagram by S1, S2 and S3). For each intersection, we aligned the sequences using Muscle [27] and built a HMM profile using HMMER [30]. All these HMM profiles were merged in one file to create a HMM database. Thus, we searched all sequences without a group (S4 and S5 for instance) for shared similarity with one HMM profile (HMMER hmmpfam program) [30]. According to the E-value, we attributed the sequence (E-value ≤ 10^{-10}) to the corresponding group or excluded it from that group (E-value > 10^{-10}).

higher the score, the better the agreement among the four methods. With this scoring approach, the user of FUNGIpath can evaluate the reliability of each predicted group of orthologs at any time.

Reconstructing pathways

Transferring EC number annotations to predicted groups of orthologs

Once the final groups of orthologs have been defined and attested, the functional annotations defined for well-studied proteins referenced in reliable public databases can be transferred to homologous unannotated amino acid sequences. For that purpose, an HMM profile was built for each final group of orthologs after multiple alignment of their sequences [27] and use of the HMMER programs (hmmbuild and then hmmcalibrate [30]). We then searched all the HMM profiles against the sequences annotated with a valid four-digit EC number available in Swiss-Prot release 56.7 using the HMMER hmmsearch program [30]). The Swiss-Prot functional annotation was transferred to all members of a group of orthologs displaying a best hit E-value ≤ 10^{-80}. The E-value threshold was lowered to 10^{-20} if at least one sequence of the group of orthologs was already endowed with the same Swiss-Prot annotation.

This approach allows fungal annotation to be improved by using the enzymatic annotation of any protein, irrespective of the phylum in which it was first described. Accordingly, we could transfer 864 EC

Table 3 Sampling the orthologs in relevant groups

Method	Union	Intersection	HMM profile and enrichment
Total number of groups	12985	12985	12850
Size of the largest group	2694	32	297

Table 4 Comparing the orthologous groups predicted by FUNGIpath and by the four methods initially used

	BRH	Inparanoid	OrthoMCL	Phylogeny	Average
Percent of groups identical with FUNGIpath	2.8%	18.6%	18.8%	10.7%	**12.7%**
Percent of groups specific in FUNGIpath	10.6%	10.6%	23.4%	24.6%	**17.3%**

numbers to 1399 of the 12850 groups of orthologs; if the fungal Swiss-Prot annotations were directly transferred, the number of groups would be only 935. This allowed 160 EC numbers to be added that were not present in fungal genomes in Swiss-Prot [11].

Note that as many as 349 EC numbers (40% of the total of 864) are present in the 20 genomes.

Numbering pathways defined by KEGG and/or MetaCyc

Once the different putative orthologs had been annotated as described above, we used them to predicting the different metabolic pathways exhaustively in the completely sequenced fungi under study. To do that, we used two reliable public databases, KEGG [12] and MetaCyc [13], which differ in the way they define pathways.

KEGG [12] defines so-called reference pathways, agglomerating related elementary pathways, while MetaCyc [13] is a universal metabolic database that presents the elementary pathways encoded by various organisms (1,500) separately, including variants (similar biochemical functions using different biochemical routes or similar sets of reactions). KEGG [12] was used to extract useful information from the reaction file and to download all corresponding GIF maps. BIOPAX (BIOlogical PAthway eXchange) files defined in MetaCyc [13] were downloaded and we automatically generated map pictures by directed graph building. We thus collected 154 reference pathways in KEGG and 1386 elementary pathways in MetaCyc, which define the main anabolic and catabolic routes.

Challenging the FUNGIpath predictions

To test the soundness of the data computed in FUNGIpath, we compared the predictions made for the model organism *S. cerevisiae* with the information published for the same ID-EC in four curated public databases: Swiss-Prot (release 56.7) [11], KEGG (version 2009-02-02) [12], MetaCyc (release 12.5) [13], and SGD (version 2009-02-10) [31]. Table 5 compares each database against the four others. Each public database appears to have its own specificity and the overlaps between pairs of the databases compared are significantly low, especially in respect of the large differences between the total numbers of ID-ECs (e.g. 1,101 in KEGG versus 527 in SGD). Table 5 also shows that the percentage of ID-ECs that are identical between public databases is at best 60% (KEGG versus Swiss-Prot). Although we mainly used Swiss-Prot data to predict enzymatic annotation in FUNGIpath, the relative percentage of identical ID-ECs was only 68%: 16% of Swiss-Prot annotations were not confirmed by the experimental strategy we used to build FUNGIpath, while 16% of FUNGIpath predictions were absent from Swiss-Prot.

To understand these differences better, we looked more closely at the similarities of EC numbers between FUNGIpath and the four public databases. Table 6 shows the distribution of identities at each digit of the shared EC numbers. It appears that the FUNGIpath predictions correspond to more than 80% of the EC numbers found in the other databases. In addition, it can be seen that almost all the differences are limited to the fourth digit, corresponding mainly to the nature of the substrate of the enzyme compared. If we compare our

Table 5 Comparing the *S. cerevisiae* enzymatic data published in four different databases with those predicted in FUNGIpath

				Distribution of ID-EC (percentage of larger database content)		
Database 1	**Total ID-EC**	**Database 2**	**Total ID-EC**	**Identical**	**Specific to database 1**	**Specific to database 2**
KEGG	1101	MetaCyc	155	127 (11%)	974 (86%)	28 (2%)
KEGG	1101	SGD	527	409 (34%)	692 (57%)	118 (10%)
KEGG	1101	Swiss-Prot	1261	889 (60%)	212 (14%)	372 (25%)
KEGG	**1101**	**FUNGIpath**	**1261**	**844 (56%)**	**417 (27%)**	**257 (17%)**
MetaCyc	155	SGD	527	132 (24%)	23 (4%)	395 (72%)
MetaCyc	155	Swiss-Prot	1261	136 (11%)	19 (1%)	1125 (88%)
MetaCyc	**155**	**FUNGIpath**	**1261**	**134 (10%)**	**21 (2%)**	**1127 (88%)**
SGD	527	Swiss-Prot	1261	433 (32%)	94 (7%)	828 (61%)
SGD	**527**	**FUNGIpath**	**1261**	**419 (31%)**	**842 (62%)**	**108 (8%)**
FUNGIpath	**1261**	**Swiss-Prot**	**1261**	**1024 (68%)**	**237 (16%)**	**237 (16%)**

Table 6 Comparing the *S. cerevisiae* enzymatic data predicted in FUNGIpath with public databases

		Number of ID-EC in FUNGIpath		Number of differences at digit position			
Public Database	Total ID-EC in FUNGIpath	identical	different	1st	2nd	3rd	4th
KEGG	1012	844 (83.4%)	34 (3.4%)	1	1	5	27
MetaCyc	148	134 (90.5%)	8 (5.4%)	2	0	1	5
SGD	504	419 (83.1%)	32 (6.3%)	5	2	3	22
Swiss-Prot	1055	1024 (97.1%)	27 (2.6%)	1	0	3	23

predictions with a predictor such as PRIAM [32], we note that 50.9% of the enzymatic annotations are identical and only 6.8% are different (the difference being mainly in the last EC number digit). The remaining 42.3% are specific to PRIAM (18.6%) or FUNGIpath (23.7%). Thus, the reliability of the automatic approach used by FUNGIpath, predicting groups of orthologous proteins and annotating their enzymatic function, seems comparable with that of other tools or the independently curated public databases. This is true whether the functional annotation is based mainly on experimental data (e.g. SGD) or on sequence similarity (e.g. KEGG).

Moreover, Table 7 shows the level of agreement when functional annotations for 12 species established by KEGG [12] and FUNGIpath are compared. Strikingly, the average number of specific ID-EC predictions is larger in FUNGIpath (1,551) than in KEGG (879) and their distribution is unexpected. Only 647 (38%) are strictly identical and 30 more are nearly identical, mostly differing only at the level of the last EC number digit, suggesting that we predicted the right reaction but the substrate is uncertain [Additional File 8]. Four times more predictions are specific to FUNGIpath (48%) than to KEGG (12%). This result is probably not due to any overprediction effect. Indeed, many enzyme predictions have been curated manually in *S. cerevisiae* and in this case the results are fairly close (15% for KEGG against 26% for FUNGIpath). Moreover, the corresponding figures for Swiss-Prot and FUNGIpath are 17.6% and 13.7%, respectively [Additional Files 9, 10]. To check whether there is any correlation, we plotted the genome size and the number of sequences with enzymatic annotations predicted respectively by FUNGIpath and KEGG (Fig. 2). We obtained a better correlation for the FUNGIpath data (R^2 = 0.28), and the slope of the tendency curve was positive with the FUNGIpath predictions but negative with the KEGG predictions. Thus, there seems to be no strong methodological bias that could explain why the predictions of FUNGIpath are generally far better than those of KEGG and close to those of the well-curated Swiss-Prot database. In fact, we observed that a significant number of the Swiss-Prot-specific IDs have no orthologs in other genomes, explaining why they are not detected in FUNGIpath. Thus, the high number of specific FUNGIpath predictions obtained is probably due to neither under-prediction by KEGG nor over-representation by FUNGIpath. Indeed, the average numbers of proteins that are annotated for an enzymatic reaction in KEGG and FUNGIpath are quite close (respectively 9.5 and 13.5% [Additional File 11]). The main reason for the better performance of FUNGIpath is probably our choice to work only with complete EC numbers [33], allowing a significant portion of the incomplete KEGG EC numbers to be recovered. For instance, 92 (25%) of the 388 incomplete EC numbers in KEGG have been completed in FUNGIpath. This enrichment by FUNGIpath is illustrated by comparing the information given by the different databases for the KEGG reference pathway 'terpenoid biosynthesis' (Fig. 3). When the level of pathway conservation is compared among the FUNGIpath, KEGG and Swiss-Prot predictions, we observe that this level is globally lowest with the Swiss-Prot data and higher in KEGG, but the highest conservation is obtained with FUNGIpath. These differences can be explained by the better annotation of fungal genomes in FUNGIpath.

Using FUNGIpath

FUNGIpath http://www.fungipath.u-psud.fr has been designed as a user-friendly website. Implemented in PHP, HTML and Javascript, it allows various aspects of fungal cell biology to be studied by performing specific predetermined queries on a PostgreSQL [34] database containing primary (genome sequences, metabolism information) and secondary (orthology) data. The sources of the fungal genomes are indicated in [Additional File 12]. An overview of the database is available in [Additional File 13].

A few examples of the proposed queries are given below.

Querying orthologs

It is possible to seek out orthologs present in the full set of genomes or to restrict queries on specific subsets defined by taxonomic or other criteria. One can use either a sequence or its sequence identifier (if available). Fig. 4a shows a typical output of such queries. Each resulting group of orthologs is associated with its confidence score (computed as described supra), a putative function (if any), an EC number (if available), the group

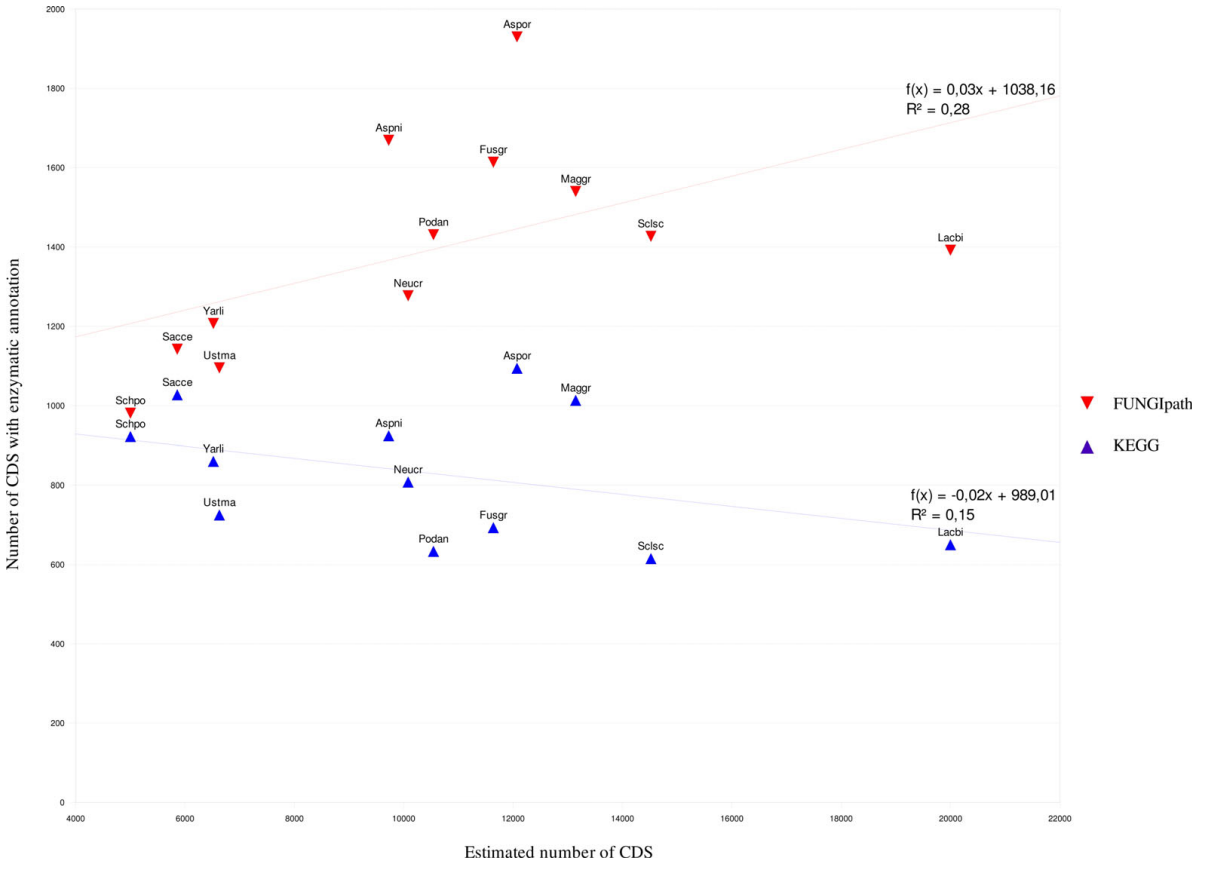


Figure 2 Correlation between genome size and number of proteins annotated with a four-digit EC number. The x-axis represents the genome size computed as the estimated number of CDS and the y-axis is the number of CDS endowed with enzymatic annotation. The red triangles correspond to the FUNGIpath data and the blue triangles to the KEGG data.

size and the conservation profile for the previously selected species. The list of orthologous (including inparalog) IDs belonging to the selected species can also be displayed. Moreover, as shown in Fig. 4b, its multiple sequence alignment can be computed in the process, and the topology of its deduced phylogenetic tree can then be examined, in order to evaluate the predicted group and to assess its relevance in terms of range of sequence identity and functional annotation.

For instance, querying the sequence UM03237.1 belonging to the *Ustilago maydis* genome defines a final group that is found whichever method is used (confidence score is maximal) and displays an alignment of quite good quality. Thus, the likelihood of this group of orthologous proteins seems quite reasonable if we combine the high-level quality of the score and the suitability of its alignment.

Exploring pathways

FUNGIpath further allows the conservation of pathways between different fungi to be checked and visualized. This can be done either at the level of a particular step (corresponding to a defined EC number) in a pathway or by considering all the steps of a complete pathway. Figs. 5 and 6 detail the different strategies used by FUNGIpath (see below). Moreover, one can handle a user-defined pathway delineated in a simplified BIOPAX format (data not shown).

Searching a specific step in a pathway

Searching a specific EC number (Fig. 5a) allows the level of conservation of this enzyme activity in each taxonomic group to be assessed; also the full list of pathways to which this EC number is predicted to belong can be obtained directly (Fig. 5b). For instance, Fig. 5 shows that acylamide amidohydrolase (EC 3.5.1.4) is very well conserved in fungi and is involved in at least six different pathways in both the KEGG and MetaCyc databases (Fig. 5b). Since this activity is used in so many pathways of both primary and secondary metabolisms, it is not surprising to find this EC number in ten distinct groups of orthologous proteins ranging in size from 4 to 25 members (data not shown). The distribution of the

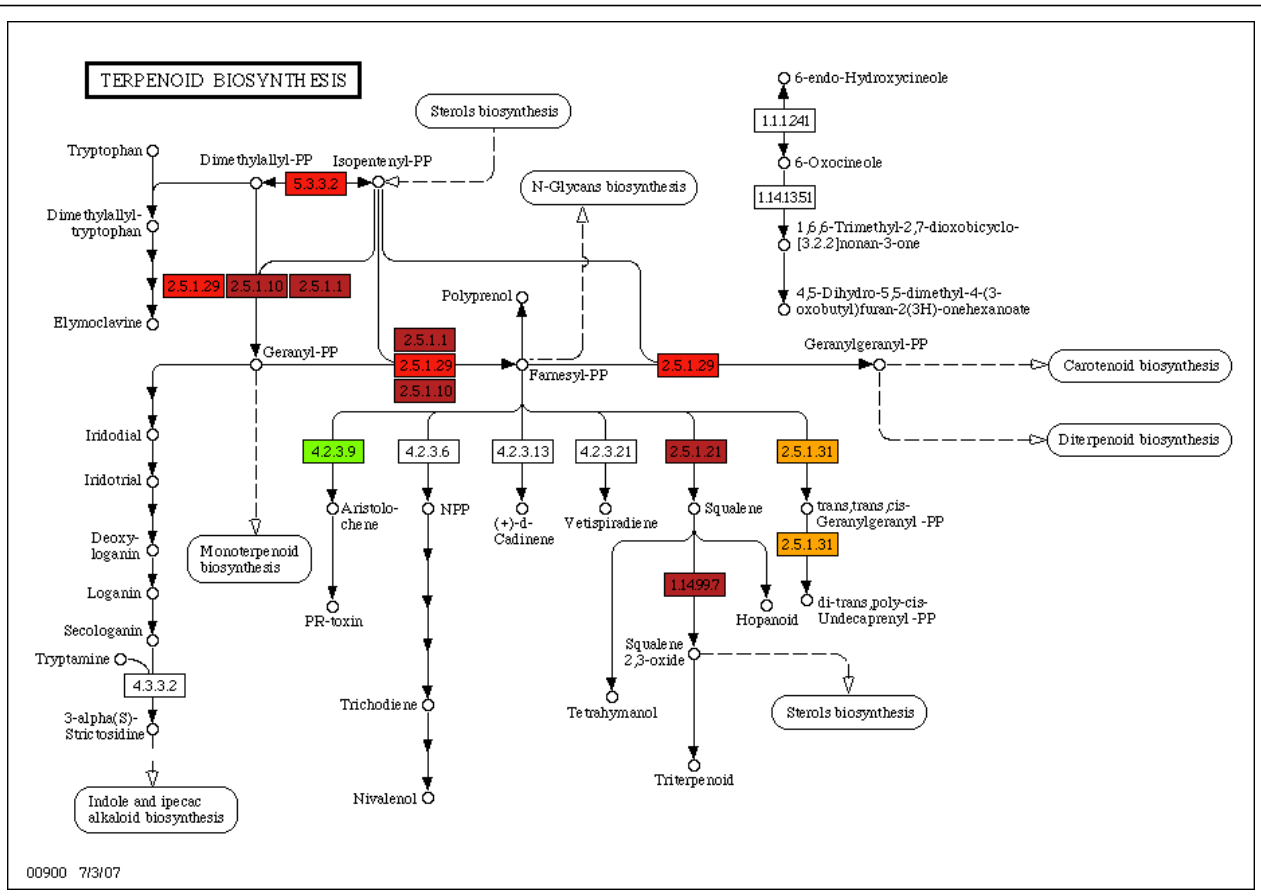


Figure 3 Comparison of the levels of conservation of the 'terpenoid biosynthesis' pathway according to different sources (Swiss-Prot, KEGG and FUNGIpath). The level of conservation of each EC number involved in the 'terpenoid biosynthesis' pathway was computed in FUNGIpath (rectangle) and two public sources. The coloured triangles represent the Swiss-Prot data for the 17 species shared with FUNGIpath. The coloured circles stand for the KEGG data for the 12 species shared with FUNGIpath.

Table 7 Comparison of enzymatic data between KEGG and FUNGIpath based on the 12 species they share

	Number of ID-EC		Number of			
Genome	**KEGG**	**FUNGIpath**	**Identical ID-EC**	**Same ID with different EC**	**KEGG specific ID-EC**	**FUNGIpath specific ID-EC**
Aspergillus nidulans	967	1890	675 (31%)	30 (1%)	262 (12%)	1185 (55%)
Aspergillus oryzae	1142	2148	853 (36%)	45 (2%)	244 (10%)	1250 (52%)
Fusarium graminearum	725	1786	535 (27%)	26 (1%)	164 (1%)	1225 (63%)
Laccaria bicolor	684	1536	472 (27%)	31 (2%)	181 (11%))	1033 (60%)
Magnaporthe grisea	1070	1801	749 (36%)	39 (2%)	282 (14%))	1013 (49%)
Neurospora crassa	852	1407	658 (42%)	26 (2%)	168 (11%)	723 (46%)
Podospora anserina	665	1594	473 (27%)	18 (1%)	174 (10%)	1103 (62%)
Saccharomyces cerevisiae	1101	1261	844 (57%)	35 (2%)	222 (15%)	382 (26%)
Schizosaccharomyces pombe	1009	1073	752 (58%)	33 (3%)	224 (17%)	288 (22%)
Sclerotinia sclerotiorum	651	1601	493 (28%)	16 (1%)	142 (8%)	1092 (63%)
Ustilago maydis	772	1206	546 (39%)	35 (3%)	191 (3%)	625 (45%)
Yarrowia lipolytica	909	1311	710 (48%)	27 (2%)	172 (12%)	574 (39%)
Average	879	1551	647 (38%)	30 (2%)	202 (12%)	874 (48%)

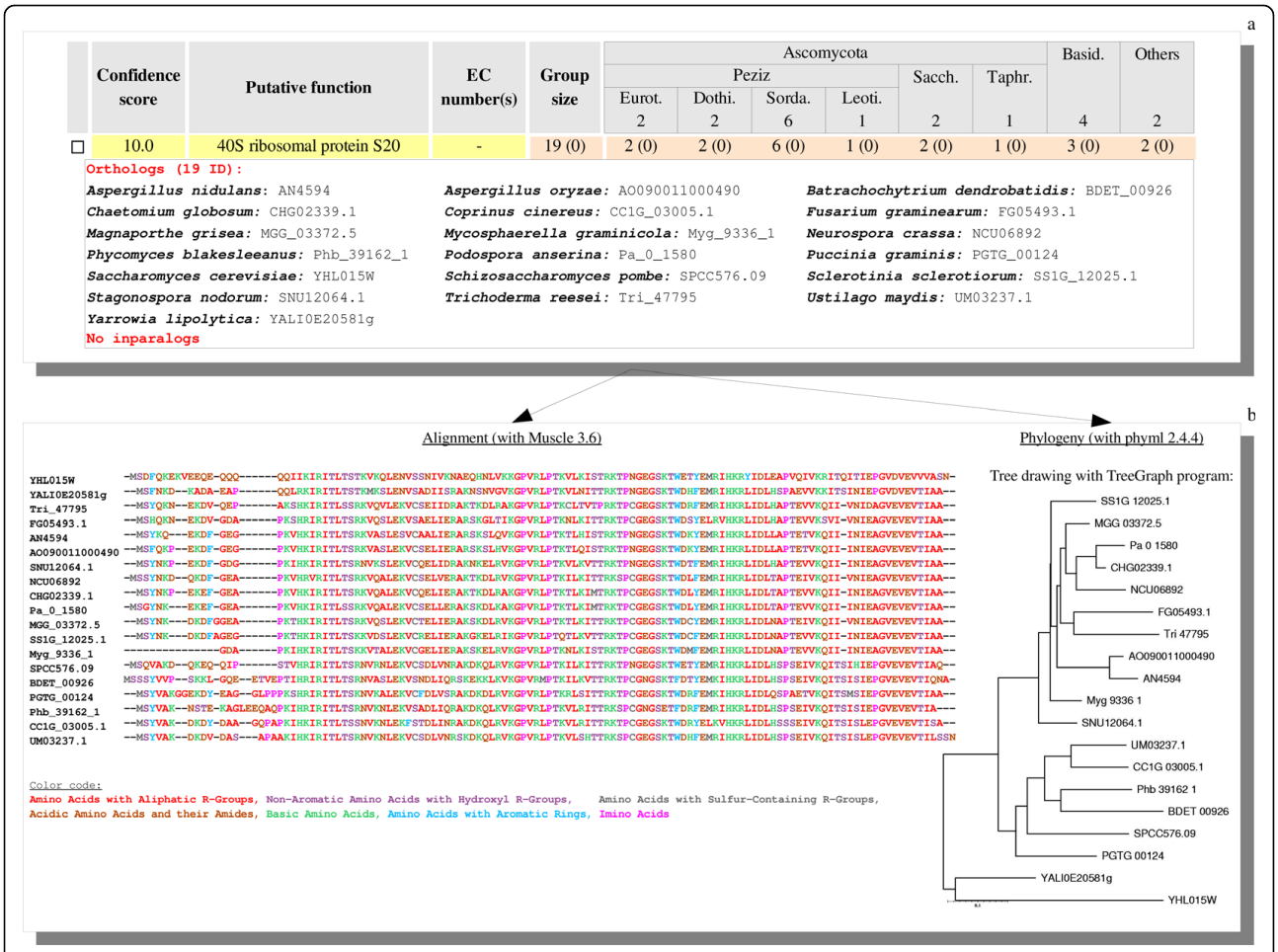


Figure 4 Searching groups of orthologous proteins. **a**: Results obtained when searching groups containing the protein ID UM03237.1. The first four columns give general information about the group and the later columns indicate the profile of conservation for the different groups of fungi. Putting the mouse on the group size displays a tooltip with its IDs and their species names. **b**: Alignment [27] and phylogeny [28] of the 20 sequences of this group of orthologs are immediately available by clicking on the group size.

different orthologs and inparalogs present in these groups can be further used to study the evolution of these different pathways using the approaches described in Fig. 4.

Searching a complete pathway

It is further possible to assess the level of conservation of each EC number in a complete pathway. Figs. 6 and 7 illustrate the available queries we propose in order to analyze primary (e.g. biotin metabolism) and secondary (e.g. terpenoid biosynthesis) pathways, respectively. The results are presented as both a KEGG gif map (Figs. 6a and 7a) and a table listing the presence/absence of each step in the pathway in the various fungal species (Figs. 6b and 7b), examining the EC numbers associated with each step. The conservation level of the different steps in the pathway is indicated by a colour code from dark red (100%) to white (0%). Groups of orthologous proteins associated with the conserved EC numbers are listed in the genome features table (Fig. 6c). Note that rich information is available and can be viewed using mouse-over facilities on many - 'explicit' and 'implicit' - links; for example, protein sequences can be downloaded for further study.

Fig. 6 shows that only two of the five steps in biotin biosynthesis are highly conserved. EC 2.8.1.6 is detected in all the species compared except *Aspergillus oryzae* and *Magnaporthe grisea*. EC 2.6.1.62 is absent from several species (*Coprinus cinereus*, *Puccinia graminis*, *U. maydis*, *Batrachochytrium dendrobatidis* and *Phycomyces blakesleeanus*). Thus, the KEGG reference pathway 'biotin metabolism' (Fig. 6a) appears to be incomplete in many fungi, since several of its specific enzyme activities (EC 2.3.1.47, 3.5.1.12, 6.2.1.14, 6.2.1.11 and 6.3.3.3) are not found. We may suppose that either these EC numbers exist in the fungi but are not currently detectable, or the fungi use other enzyme

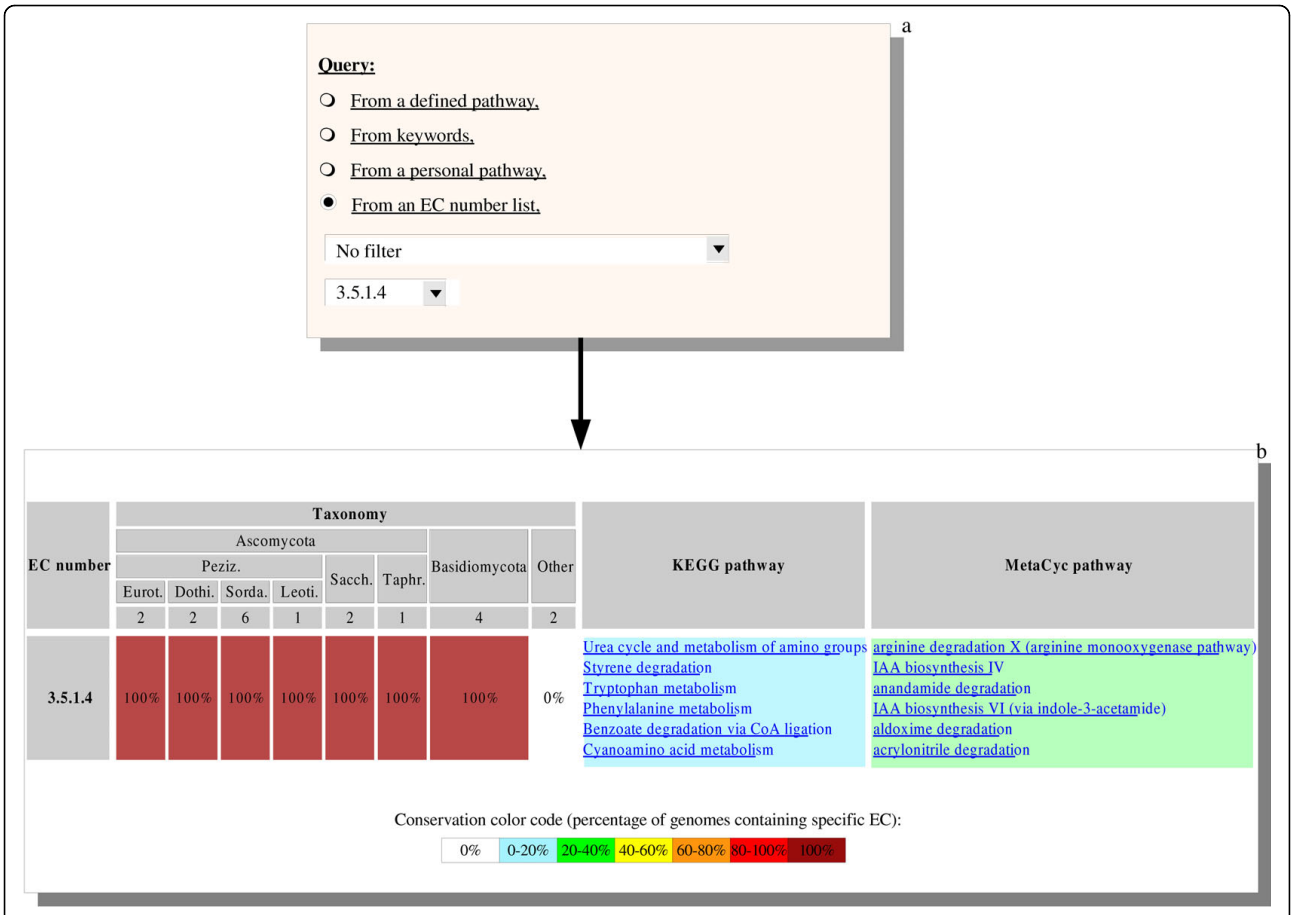


Figure 5 Exploring pathways using a specific EC number. **a**: The EC number 3.5.1.4 was searched against all 20 genomes available in the database. **b**: The level of conservation among the different groups belonging to different taxonomic groups of fungi is indicated with a colour code (from white (0%) to dark red (100%)). The pathways that contain the requested EC number are listed in the KEGG and MetaCyc columns, respectively. Note that the pathway names are different in these two databases.

activities to catalyse these reactions (see below). Moreover, the further steps in biotinylation catalyzed by the ligases EC 6.3.4.9, 6.3.4.10, 6.3.4.11, and 6.3.4.15 are fully conserved in all the main taxonomic groups of fungi.

Fig. 6c further shows that most of the EC numbers (blue text) correspond to proteins that have no EC number assignment in Swiss-Prot but have been annotated in FUNGIpath by orthology prediction. Only two of the twenty genomes (*S. cerevisiae* and *Schizosaccharomyces pombe*) have an annotation in these two databases (bold text).

Fig. 7 shows that only six of the 13 EC numbers involved in the KEGG reference pathway 'terpenoid biosynthesis' appear to be conserved among the fungi analyzed. Of these six EC numbers, four (EC 1.14.99.7, 2.5.1.1, 2.5.1.10 and 2.5.1.21) are found in all the species present in FUNGIpath. Some EC numbers are missing from only one fungal group: this seems to be the case for EC 5.3.3.2, which is absent in the Taphrinomycotina group. Note, however, that this group is represented by only one species, namely *S. pombe*. Two EC numbers (4.2.3.9 in green and 2.5.1.31 in orange) seem to be specific to certain fungi.

Discussion

Fungal metabolism is exceptionally rich and complex [35], generating a wide variety of secondary metabolic pathways as these organisms progressively evolved to invade new ecosystems. Except in a few model organisms, very few reactions have been studied experimentally. The present-day facility in obtaining complete genome sequences for organisms that have never been experimentally studied has revealed a wide gap between the knowledge gained by disclosing full repertoires of putative amino acid sequences and ignorance of their actual function.

To close this gap, one needs to transfer functional annotation to putative sequences by homology using inductive instead of hypothetico-deductive approaches (holism versus reductionism) [36]. For metabolism, this

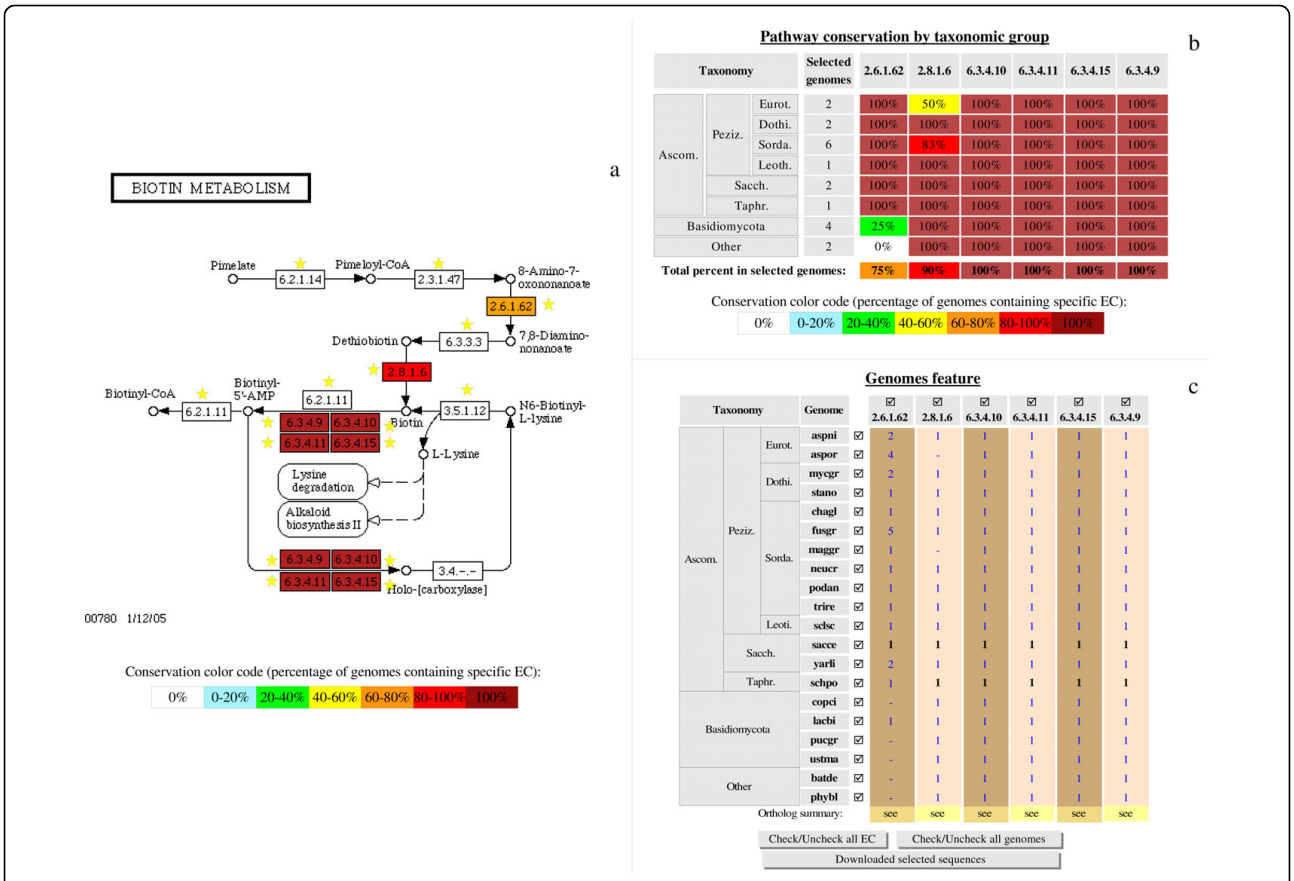


Figure 6 Analysis of 'biotin metabolism' defined by KEGG. The 'biotin metabolism' reference pathway was searched against all 20 genomes available in the database. The conservation colour code is the same as in Fig. 5. **a**: The EC numbers specific to this pathway are indicated by a yellow star. **b**: The table lists the percentages of conservation of this pathway in each species forming the different taxonomic groups of fungi. c: The table lists the presence (number of sequences) or absence (dash) of orthologs for each enzyme. The species are sorted by taxonomic group.

allows entire pathways to be reconstructed [37]. In order to facilitate the study of fungal metabolism and its evolution, we have created the tool FUNGIpath, which makes the predictions made on this homology basis publicly available. Thus, it was necessary to design new experimental approaches in order to obtain reliable and sound predictions.

Collecting reliable orthologs

The first requirement was to detect sound orthologs, knowing that there is no uniquely reliable way to do so [22]. The most commonly-used approach is bidirectional best hits (BRH) of BLAST alignment, with imposition of strict criteria on discriminating E-value over a given alignment length, but various more sophisticated approaches have also been developed [22]. Selecting the best method(s) is not easy. For instance, benchmarking tests suggested that Inparanoid performs best while BRH is good for closely-related species [38]. More recently, BRH was found to give results comparable to the more sophisticated methods [39], but it is limited to finding only a single hit among the multiple possible links between paralogs.

We therefore preferred to use several different approaches simultaneously, three based on sequence similarity and one on phylogeny, to obtain robust results. Since the overlap between the outputs of these four methods is very narrow (a result underlining how conflicting these orthology methods are), we enriched the data found in the intersection of the different methods with a HMM approach. This allowed us to obtain fairly coherent sets of reliable orthologs forming well-defined groups that are of adequate size (the largest containing only 297 sequences) and biologically relevant.

Using reliable orthologs to improve functional annotation

The second requirement for exploiting these orthology data to predict metabolic pathways in fungal species that have never been studied experimentally was to assign a functional annotation to each group of orthologous proteins. To do that, a correspondence was established between a group and an EC number, defining an

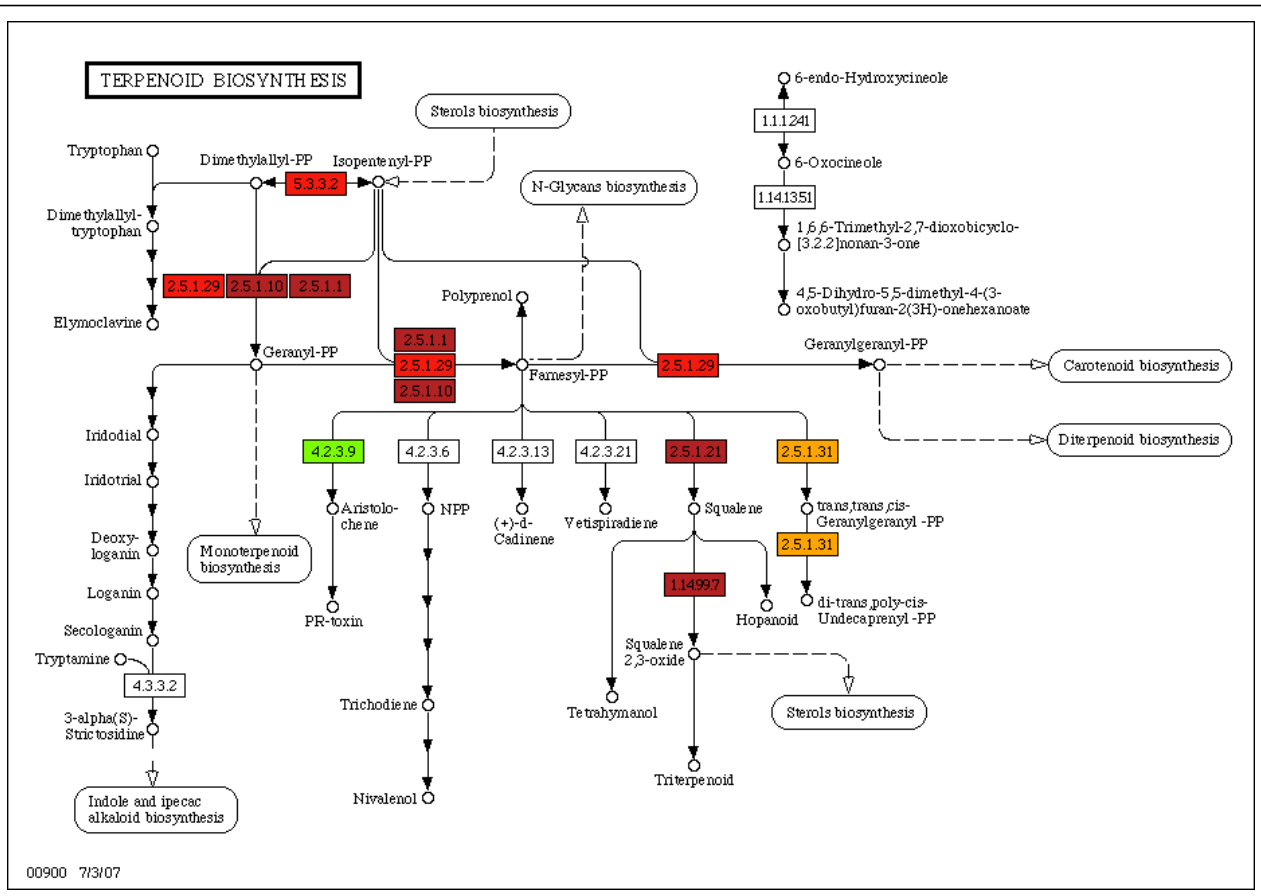


Figure 7 Exploring pathways using a specific pathway name. The 'terpenoid biosynthesis' reference pathway defined by KEGG was searched against all 20 genomes available in the database. **a**: Each EC number has been coloured according to its global level of conservation as in Figs. 5 and 6. Two EC numbers (2.5.1.31 and 4.2.3.9) specific to this pathway (indicated by a yellow star) are not detected in all species studied. **b**: This table lists the percentage conservation of each EC number in this pathway among all taxonomic groups of fungi. Its global presence in all taxonomic groups is given in the last line of this table.

enzyme catalyzing a specific step in a known pathway included in the KEGG and MetaCyc databases. Figs. 5 and 6 show how group(s) of orthologs responsible(s) for a specific enzyme can be found and how this EC number is distributed in the different genomes. *Inter alia*, the multiple sequence alignment and the deduced phylogenetic tree can be obtained for each family of orthologs and inparalogs encoding this EC number in the fungi compared. We have provided evidence that FUNGIpath is a reliable tool for annotating enzyme function in an automatically predicted group of orthologous proteins. It gives data that are either comparable to those of the independently curated public databases or, in many cases, better (see Table 6). At any rate, most of the differences appear to be limited to the fourth digit, corresponding mainly to the nature of the substrates of the enzymes compared.

FUNGIpath is also useful for finding the set of orthologs that constitutes an entire pathway. This allows us to determine whether all the steps of the pathway have been predicted and, if so, in how many of the genomes compared that pathway is complete. Indeed, one of the main problems encountered in trying to reconstruct entire pathways from orthology data is the occurrence of missing data [40] such as pathway holes [41]. The absence of an EC number (orphan metabolic activities [42]) may be due to a low percentage identity of the corresponding amino acid sequence or to its replacement with another protein. Alternatively, the simultaneous absence of several EC numbers that belong to a specific pathway would suggest that the entire pathway is absent from the species concerned. This is the case, for instance, in the later steps in the KEGG reference pathway 'terpenoid biosynthesis', where the last three EC numbers are missing (Fig. 7a). However, it is possible that this absence may simply be due to a major annotation problem or to the replacement of this pathway with an alternate, undetected, one.

Overall, FUNGIpath appears to be a useful and innovative tool for helping to resolve some artifactual pathway holes. For instance, it is unique in annotating a group of orthologs found in six species as EC 4.2.3.9 (Fig. 3), aristolochene synthase. No such amino acid sequences are predicted in Swiss-Prot, KEGG or MetaCyc, but the presence of aristolochene synthase has been demonstrated experimentally in two fungi not included in FUNGIpath [43,44], supporting our prediction.

Conclusions

FUNGIpath appears to be a reliable tool for the analysis of fungal metabolism. It will be especially useful for annotating newly-sequenced genomes of poorly-studied organisms.

Moreover, it allows the respective metabolisms of various taxa to be compared easily. For instance, 101 EC numbers are found uniquely in ascomycetes (data not shown) and may help to delineate the metabolic specificities of the last common ancestor of this group.

As more and more genomes are expected to be decrypted in the near future, tools such as FUNGIpath will be very useful for the progressive reconstruction of primary and secondary metabolisms in the ancestors of the main branches of the present-day fungal tree and for elucidating the evolution of various specific derived metabolisms. FUNGIpath will be updated regularly (at least twice a year) with newly published fungal genomes.

Availability and requirements

The database is available at http://www.fungipath.u-psud.fr. This web site is optimized for Firefox 2.x and has been successfully tested for Safari 2.0.3 and Internet Explorer 7.0.

Lists of abbreviations

BRH: Best reciprocal hits; CDS: Coding sequences; EC number: Enzyme Commission number; ID-EC: a unique protein identifier (ID) and EC number pair.

Additional file 1: Number of sequencing projects by kingdom. The table shows the number of published and ongoing genomes for the three kingdoms.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-11-81-S1.PDF]

Additional file 2: Example of ID that may be present in several groups determined by the BRH method. BRH pairs define multiple links between the different orthologous proteins (A) as indicated by bi-directional arrows. Accordingly, the lack of BRH link between proteins A_2 and A_4, leads to building two different groups of orthologs.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-11-81-S2.PDF]

Additional file 3: Influence of the E-value threshold on the association of a sequence and an HMM profile. The first table shows the sequence distribution after comparison with the HMM profile database according to the number of methods that initially assign a protein sequence (ID) to an group of orthologs. The second table shows the same results for different E-value thresholds.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-11-81-S3.PDF]

Additional file 4: Distribution of orthologous group sizes. The graph represents the distribution of the group size with the number of sequences in a group on the x-axis and the number of groups on the y-axis.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-11-81-S4.PDF]

Additional file 5: Number of sequences per genome. The table shows, for each genome, the total numbers (and their percentages) of protein sequences, of proteins belonging to groups of orthologs and of proteins endowed with an enzymatic activity (annotated with a EC number).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-11-81-S5.PDF]

Additional file 6: Comparison of the percentages of annotated sequences for the 20 fungal genomes. The graph represents, for each genome, the genome size on the x-axis and the percentage of annotated CDS on the y-axis. Genome abbreviations: AspNi for *A. nidulans*, AspOr for *A. oryzae*, BatDe for *B. dendrobatidis*, ChaGl for *C. globosum*, CopCi for *C. cinereus*, FusGr for *F. graminearum*, LacBi for *L. bicolor*, MagGr for *M. grisea*, MycGr for *M. graminicola*, NeuCr for *N. crassa*, PhyBl for *P. blakesleeanus*, PodAn for *P. anserina*, PucGr for *P. graminis*, SacCe for *S. cerevisiae*, SchPo for *S. pombe*, SclSc for *S. sclerotiorum*, StaNo for *S. nodorum*, TriRe for *T. reesei*, UstMa for *U. maydis* and YarLi for *Y. lipolytica*.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-11-81-S6.PDF]

Additional file 7: Distribution of sequence lengths. The graph represents the distribution of sequence lengths (x-axis) with the number of sequences (y-axis). The red point corresponds to all the sequences and the blue point to the sequences assigned to an orthologous group.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-11-81-S7.PDF]

Additional file 8: Analysis of the different enzymatic annotations in KEGG. The table provides, for each genome, the numbers of ID-EC that diverge and the positions that differ between KEGG and FUNGIpath.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-11-81-S8.PDF]

Additional file 9: Comparison of enzymatic data between Swiss-Prot and FUNGIpath (based on 17 shared species). The table provides, for each genome, the numbers of ID-EC that are common, divergent or specific between Swiss-Prot and FUNGIpath.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-11-81-S9.PDF]

Additional file 10: Analysis of the different enzymatic annotations in Swiss-Prot. The table provides, for each genome, the numbers of ID-EC that diverge and the positions that differ.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-11-81-S10.PDF]

Additional file 11: Comparison of data for FUNGIpath genomes. The table provides, for each genome, the number of sequences with complete enzymatic annotation for FUNGIpath, KEGG and Swiss-Prot.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-11-81-S11.PDF]

Additional file 12: Sources of the genomic data used in FUNGIpath. The respective sequencing centers of each fungal genome are indicated by the url we used to download the primary genomic data.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-11-81-S12.PDF]

Additional file 13: Database schema. The various tables (schematized as rectangles) are coloured in red (genomic data), yellow (predictions of orthologs), and blue (pathway data). Links between tables are indicated by lines. Foreign key names are displayed in italics.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-11-81-S13.PDF]

Acknowledgements
We are grateful to Philippe Silar for his help during the process of designing the web site and his helpful comments about this work. The computations were performed on the MIGALE platform (INRA, Jouy-en-Josas, France). Special thanks to the JGI, Broad Institute, NITE, Stanford University, Sanger Institute, Genoscope and Génolevures for the fungal genomes, Swiss-Prot for the protein annotations, and KEGG and MetaCyc for the metabolic pathways. SG is a PhD student supported by a 'Doctorant CNRS' fellowship. We thank the Université Paris-Sud (PPF Bioinformatique et Biomathématiques), and the Agence Nationale de la Recherche (ANR-05-MMSA-0009 MDMS_NV_10) for support. Finally, we thank two anonymous reviewers whose inputs have led to significant improvements in this work.

Authors' contributions
SG built the database and developed the website. OL supervised the work and tested the tool. All authors (SG, BL, and OL) drafted, read and approved the final manuscript.

Received: 25 June 2009
Accepted: 1 February 2010 Published: 1 February 2010

References

1. Liolios K, Mavrommatis K, Tavernarakis N, Kyrpides NC: **The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata.** *Nucleic Acids Res* 2007, **36**: D475-479.
2. Lutzoni F, Kauff F, Cox CJ, McLaughlin D, Celio G, Dentinger B, Padamsee M, Hibbett DS, James TY, Baloch E, Grube M, Reeb V, Hofstetter V, Schoch C, Arnold AE, Miądlikowska J, Spatafora J, Johnson D, Hambleton S, Crockett M, Shoemaker R, Sung GH, Lücking R, Lumbsch T, O'Donnell K, Binder M, Diederich P, Ertz D, Gueidan C, Hansen K, Harris RC, Hosaka K, Lim YW, Matheny B, Nishida H, Pfister D, Rogers J, Rossman A, Schmitt I, Sipman H, Stone J, Sugiyama J, Yahr R, Vilgalys R: **Assembling the fungal tree of life: progress, classification, and evolution of subcellular traits.** *American Journal of Botany* 2004, **91**:1446-1480.
3. Hibbett DS, Binder M, Bischoff JF, Blackwell M, Cannon PF, Eriksson OE, Huhndorf S, James T, Kirk PM, Lücking R, Thorsten Lumbsch H, Lutzoni F, Matheny PB, McLaughlin DJ, Powell MJ, Redhead S, Schoch CL, Spatafora JW, Stalpers JA, Vilgalys R, Aime MC, Aptroot A, Bauer R, Begerow D, Benny GL, Castlebury LA, Crous PW, Dai YC, Gams W, Geiser DM, Griffith GW, Gueidan C, Hawksworth DL, Hestmark G, Hosaka K, Humber RA, Hyde KD, Ironside JE, Kõljalg U, Kurtzman CP, Larsson KH, Lichtwardt R, Longcore J, Miadlikowska J, Miller A, Moncalvo JM, Mozley-Standridge S, Oberwinkler F, Parmasto E, Reeb V, Rogers JD, Roux C, Ryvarden L, Sampaio JP, Schüssler A, Sugiyama J, Thorn RG, Tibell L, Untereiner WA, Walker C, Wang Z, Weir A, Weiss M, White MM, Winka K, Yao YJ, Zhang N: **A higher-level phylogenetic classification of the Fungi.** *Mycol Res* 2007, **11**:509-547.
4. Soanes DM, Alam I, Cornell M, Wong HM, Hedeler C, Paton NW, Rattray M, Hubbard SJ, Oliver SG, Talbot NJ: **Comparative Genome Analysis of Filamentous Fungi Reveals Gene Family Expansions Associated with Fungal Pathogenesis.** *PLoS ONE* 2008, **3(6)**:e2300.
5. Kriventseva EV, Rahman N, Espinosa O, Zdobnov EM: **OrthoDB: the hierarchical catalog of eukaryotic orthologs.** *Nucleic Acids Res* 2008, **36**: D271-275.
6. Hedeler C, Wong HM, Cornell J, Alam I, Soanes DM, Rattray M, Hubbard SJ, Talbot NJ, Oliver SG, Paton NW: **e-Fungi: a data resource for comparative analysis of fungal genomes.** *BMC Genomics* 2007, **8**:426.
7. Marthey S, Aguileta G, Rodolphe F, Gendrault A, Giraud T, Fournier E, Lopez-Villavicencio M, Gautier A, Lebrun MH, Chiapello H: **FUNYBASE: a FUNgal phYlogenomic dataBASE.** *BMC Bioinformatics* 2008, **9**:456.
8. Remm M, Storm CE, Sonnhammer EL: **Automatic clustering of orthologs and in-paralogs from pairwise species comparisons.** *J Mol Biol* 2001, **314**:1041-1052.
9. Li L, Stoeckert CJ Jr, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13**:2178-2189.
10. Enright AJ, Kunin V, Ouzounis CA: **Protein families and TRIBES in genome sequence space.** *Nucleic Acids Res* 2003, **31(15)**:4632-8.
11. The UniProt Consortium: **The universal protein resource (UniProt).** *Nucleic Acids Res* 2008, **36**:D190-195.
12. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res* 1999, **27**:29-34.
13. Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C, Walk TC, Zhang P, Karp PD: **The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases.** *Nucleic Acids Res* 2008, **36**:623-31.
14. **Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB).** http://www.chem.qmul.ac.uk/iubmb/enzyme/index.html.
15. Giles PF, Soanes DM, Talbot NJ: **A relational database for the discovery of genes encoding amino acid biosynthetic enzymes in pathogenic fungi.** *Comp Funct Genomics* 2003, **4(1)**:4-15.
16. Boyer F, Viari A: **Ab initio reconstruction of metabolic pathways.** *Bioinformatics* 2003, **19(Suppl 2)**:ii26-34.
17. Overbeek R, Larsen N, Maltsev N, Pusch GD, Selkov E: **WIT/WIT2: Metabolic reconstruction system.** *Bioinformatics, database and systems* Kluwer academic Publisher, Boston, USA 1999.
18. Paley S, Karp PD: **Evaluation of computational metabolic-pathway predictions for Helicobacter pylori.** *Bioinformatics* 2002, **18**:715-724.
19. Brenner SE: **Errors in genome annotation.** *Trends Genet* 1999, **15**:132-133.
20. Janssen P, Goldovsky L, Kunin V, Darzentas N, Ouzounis CA: **Genome coverage, literally speaking. The challenge of annotating 200 genomes with 4 million publications.** *EMBO Rep* 2005, **6**:397-399.
21. Ouzounis CA, Karp PD: **The past, present and future of genome-wide re-annotation.** *Genome Biology* 2002, **3**:comment2001.1-2001.6.
22. Kuzniar A, van Ham RC, Pongor S, Leunissen JA: **The quest for orthologs: finding the corresponding gene across genomes.** *Trends Genet* 2008, **24**:539-551.
23. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-10.
24. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling.** *Proc Natl Acad Sci USA* 1999, **96**:2896-2901.
25. Lemoine F, Lespinet O, Labedan B: **Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data.** *BMC Evol Biol* 2007, **7**:237.
26. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E: **EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates.** *Genome Res* 2009, **19**:327-35.
27. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32**:1792-7.
28. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol* 2003, **52**:696-704.
29. Felsenstein J: **PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author.** Department of Genome Sciences, University of Washington, Seattle 2005.

30. Eddy SR: **Profile Hidden Markov models.** *Bioinformatics* 1998, **14(9)**:755-63.
31. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, Weng S, Botstein D: **SGD: Saccharomyces Genome Database.** *Nucleic Acids Res* 1998, **26**:73-79.
32. Claudel-Renard Clotilde, Chevalet Claude, Faraut Thomas, Kahn Daniel: **Enzyme-specific profiles for genome annotation: PRIAM.** *Nucleic Acids Research* 2003, **22**:6633-6639.
33. Green ML, Karp PD: **Genome Annotation Errors in Pathway Databases Due to Semantic Ambiguity in Partial EC Numbers.** *Nucleic Acids Research* 2005, **33**:4035-4039.
34. **PostgreSQL database management system.** http://www.postgresql.org/.
35. Vaidyanathan S, Harrigan GG, Goodacre R: **Metabolome analyses: strategies for systems biology.** *Metabolites and Fungal Virulence* SpringerDriggers EM, Brakhage AA 2005, 367-381.
36. Maynard Smith J: **The Problems of Biology.** Oxford University Press 1986.
37. Oliver SG: **From DNA sequence to biological function.** *Nature* 1996, **379**:597-600.
38. Hulsen T, Huynen M, de Vlieg J, Groenen P: **Benchmarking ortholog identification methods using functional genomics data.** *Genome biology* 2006, **7(4)**:R31.
39. Altenhoff A, Dessimoz C: **Phylogenetic and Functional Assessment of Orthologs Inference Projects and Methods.** *PLoS Computational Biology* 2009, **5(1)**:e1000262.
40. Osterman A, Overbeek R: **Missing genes in metabolic pathways: a comparative genomics approach.** *Curr Opin Chem Biol* 2003, **7**:238-251.
41. Green ML, Karp PD: **A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases.** *BMC Bioinformatics* 2004, **5**:76.
42. Chen L, Vitkup D: **Predicting genes for orphan metabolic activities using phylogenetic profiles.** *Genome Biol* 2006, **7**:R17.
43. Proctor RH, Hohn TM: **Aristolochene synthase. Isolation, characterization, and bacterial expression of a sesquiterpenoid biosynthetic gene (Ari1) from *Penicillium roqueforti*.** *J Biol Chem* 1993, **268**:4543-4548.
44. Cane DE, Kang I: **Aristolochene synthase: purification, molecular cloning, high-level expression in Escherichia coli, and characterization of the *Aspergillus terreus* cyclase.** *Arch Biochem Biophys* 2000, **376**:354-64.

doi:10.1186/1471-2164-11-81
Cite this article as: Grossetête *et al.*: FUNGIpath: a tool to assess fungal metabolic pathways predicted by orthology. *BMC Genomics* 2010 **11**:81.