Université de Montpellier 2 **Ecole Doctorale SIBAGHE**

Habilitation à Diriger des Recherches

Présentée et soutenue publiquement le 11 mai 2012

Intégration des données et des connaissances appliquée à la génomique des plantes tropicales et méditerranéennes

par

Manuel Ruiz

Jury :

Dr Quesneville Hadi, INRA, Versailles
Dr Leroy Philippe, INRA, Clermont-Ferrand
Dr Christen Richard, CNRS, Nice
Prof David Jacques, Montpellier SupAgro
Dr This Patrice, INRA, Montpellier

CIRAD-BIOS
UMR Amélioration Génétique et Adaptation des Plantes méditerranéennes et tropicales
Equipe "Intégration des données"
Montpellier

Sommaire

Curriculum Vitæ

Dr Manuel Ruiz

né le 09 mai 1971, à Perpignan

Adresse professionnelle:

CIRAD-BIOS

UMR Amélioration Génétique et Adaptation des Plantes méditerranéennes et tropicales

Equipe "Intégration des données"

TA A-108/03

Avenue Agropolis

34398 Montpellier Cedex 5

France

Tel : (33) 4 67 61 65 29

Fax : (33) 4 67 61 56 05

1) Parcours

- Depuis juillet 2007, responsable de l'équipe Intégration des Données de l'UMR DAP, puis de l'UMR AGAP
- Depuis juillet 2007, responsable scientifique de South Green Bioinformatics Platform, http://southgreen.cirad.fr
- Depuis Avril 2002, chercheur en Bioinformatique, au CIRAD
- 1997-2001, Thèse en Bioinformatique, Université des Sciences de Montpellier: "Analyse bioinformatique standardisée IMGT des relations séquence-structure des immunoglobulines et récepteurs T", dans l'équipe IMGT (international ImMunoGeneTics information system, http://imgt.cines.fr/)

2) Thèmes de recherche

- Représentation des connaissances, organisation et gestion de l'information liée aux domaines de la génétique et génomique végétale
- Intégration sémantique des données en génomique végétale
- Annotation automatique des données génomiques et génomique comparative

3) Liste des systèmes d'information dans le développement desquelles j'ai été directement impliqué :

- TropGeneDB http://tropgenedb.cirad.fr/
- CocoaGenDB http://cocoagendb.cirad.fr/
- OryGenesDB http://orygenesdb.cirad.fr/
- OryzaTagLine http://urgi.versailles.inra.fr/OryzaTagLine/
- SAT, SSR Analysis Tool http://sat.cirad.fr/sat/
- SNIPLAY, http://sniplay.cirad.fr
- SouthGreen, http://southgreen.cirad.fr

4) Vie collective

- Editeur associé de la revue BMC Genomics
- Membre du comité de programme des conférences francophones JOBIM, Journées Ouvertes en Biologie, Informatique et Mathématiques, 2010, 2011 et 2012.
- Membre du comité local d'organisation de la conférence internationale Biodiversity Information Standards, TDWG, Taxonomic Databases Working Group, 2009, Montpellier

5) Enseignement

a) Cours universitaires

- Master M2 APIMET (Amélioration des Plantes et Ingénierie végétale Méditerranéennes Et Tropicales) /SEPMET (Semences Et Plants Méditerranéens Et Tropicaux), SupAgro - TD - 4h/an
- Master M1 Biotechnologies des Plantes tropicales, UMII - cours magistral - 1h30/an

b) Cours à l'étranger

- Brésil, Bioinformatics School, Campinas, Sao Paulo, 21-26 novembre 2011
- Brésil, Workshop France-Brésil de Bioinformatique, Universidade Estadual de Santa Cruz, 08-12 novembre 2010
- Colombie, Herramientas Genómicas y Bioinformáticas para la Investigación Agrícola, Pontificia Universidad Javeriana, Bogota, 1-3 Mars 2010
- Sénégal, Module de Génomique végétale e-learning pour l'Université de Dakar, juin 2009
- Malaisie, Universiti Putra Malaysia / International Rubber Research and Development

Board, 23-25 juin 2008

- Madagascar, Formation Initiation à la bioinformatique, CIRAD Madagascar / FOFIFA, 27 novembre – 02 décembre 2006

c) Mise en place de formations professionnelles

- 2011 (6-10 juin), Bioinformatique appliquée à l'analyse de séquences biologiques, SupAgro, Montpellier
- 2011 (9-13 mai), Première session de formation ARCAD SP1-SP4, Analyse des données de polymorphisme, SupAgro, Montpellier
- 2011 (21-25 février), Bioinformatique appliquée à l'analyse de séquences biologiques, SupAgro, Montpellier
- 2010 (10-12 mai), Formation experte bioinformatique, UMR DAP, Montpellier
- 2006 (29 juin – 02 juillet), Formation Initiation à la bioinformatique, International Rubber Research and Development Board, CIRAD Montpellier
- 2005 (27-31 juin), Formation Initiation à la bioinformatique, UMR PIA, Montpellier
- 2005 (28 octobre – 04 mars), Formation Initiation à la bioinformatique, CIRAD Guadeloupe, Guadeloupe
- 2004 (4-8 octobre), Formation Initiation à la bioinformatique, UMR PIA, Montpellier

6) Encadrements

a) Thèses soutenues

- Julien Wollbrett, soutenue le 13 décembre 2011, co-encadrement 80 % avec F De Lamotte (UMR AGAP)
    - Génération semi-automatique de Services Web Sémantiques pour des bases de données relationnelles biologiques
- Pierre Larmande, soutenue le 20 décembre 2007, co-encadrement 60% avec I Mougenot (LIRMM)
    - Mutualiser et partager, un défi pour la génomique fonctionnelle végétale
- Mathieu Conte, soutenue le 21 décembre 2007, co-encadrement 10% avec C Périn (UMR AGAP)
    - Développement d'une plateforme de génomique comparative dédiée aux plantes

b) Masters/DESS : encadrements de 16 stagiaires, pour des durées de 5 à 6 mois

- Fatima Ezzahra Agharbaoui, 2010, Master 2 Intelligence Artificielle et Bioinformatique, Tanger, Maroc
    - *Développement d'une base de connaissances dédiée à l'étude des relations structure-fonction des protéines*
- Laetitia Brigitte, 2009, Master 2 Bioinformatique Toulouse
    - *Optimisation d'une interface Web de requêtes, multi-bases de données, intégrée et générée automatiquement, appliquée au système d'information TropGene*
- Julien Wollbrett, 2008, Master 2 Bioinformatique Montpellier
    - *Développement d'un système de composition automatique d'adaptateurs, basé sur une ontologie de domaine : applications à des projets d'analyse de données de diversité génétique et d'haplotypes chez les plantes*
- Keliet Aminah Olivia, 2008, Master 2 Bioinformatique Bordeaux
    - *Développement de GenDiversity, application Web intégrée pour l'analyse de données de diversité chez les plantes, basée sur la plateforme intégrée GCP Pantheon*
- Djeghem Abdelatif, 2008, Master 2 pro Informatique Montpellier
    - *Développement d'une application Web flexible, intégrée et générée automatiquement, connectée au système d'information TropGene*
- Gautier Sarah, 2007, Master 2 Bioinformatique Montpellier
    - *Développement d'une application Web pour l'analyse de données de génotypage provenant de sources hétérogènes*
- Rémi Moine, 2007, Master 2 Bioinformatique Montpellier
    - *Bibliotrop, développement d'un outil d'annotation de séquences utilisant la bibliographie scientifique*
- Aurélien De Brix, 2006, DESS Bioinformatique Montpellier
    - *Développement et exploitation d'un outil générique de recherche de régions codantes répétées*
- David Baux, 2005, DESS Bioinformatique Montpellier
    - *Développement d'une plateforme d'interopérabilité pour l'intégration de données phénotypiques, génétiques et génomiques chez le bananier*
- Karine Fayolle, 2004, DESS Bioinformatique Montpellier
    - *Développement d'un module Perl générateur d'interfaces web de consultation : application à l'intégration du programme CMAP dans Tropgene-DB*
- Alexis Dereeper, 2004, DESS Bioinformatique Montpellier

- *Mise en place d'un pipeline de traitement de séquences issues de banques enrichies en séquences microsatellites associé à une base de données*

- Matthieu Conte, 2003, DEA Bioinformatique Genève, Suisse
  - *Construction d'un pipeline de génomique comparative entre Oryza sativa et Arabidopsis thaliana*
- Gaetan Droc, 2002, DESS Bioinformatique Montpellier
  - *Mise en place d'une interface web sur la base de données FST et d'outils d'analyse bioinformatique*
- Anne Gaillard, 2002, DESS Bioinformatique Montpellier
  - *Application de gestion des OGM en vue de leur déclaration*
- Khalid El Karkouri, 2002, DESS Bioinformatique Montpellier
  - *Développement d'une Application de Gestion de Banques BAC en Perl/CGI et JSP*
- Mathieu Rouard, 2002, DESS Bioinformatique Montpellier
  - *Conception d'une interface modulable pour la base de données Tropgene DB*

c) Autres formations : encadrements de 3 stagiaires, pour des durées de 2 à 6 mois

- Dietwin Glaszmann, 2005, DUT Génie Informatique, Lyon
  - *Amélioration d'une applet Java traçant la généalogie des croisements des clones de Cacaoyer*
- Rémy Orain, 2003, Institut Ingénierie Informatique, 3 ème année, Limoges
  - *Conception d'un site web sur le cacao*
- Fleurine Pelissier, 2003, DUT Informatique, Montpellier
  - *Réalisation de l'interface Web de Bactrop-DB*

Titres et travaux

1) Activités de recherche

Ma thèse au sein de l'équipe IMGT, international ImMunoGeneTics information system, m'a permis d'aborder très tôt des thématiques de recherche essentielles en bioinformatique qui émergeaient dans les années 90 avec l'explosion des données de séquençage : représentation des connaissances en génomique, intégration des données, et annotation automatique des données.

Après ma thèse, de fin 2001 à avril 2002, au sein de l'Institut de Génétique Humaine de Montpellier, j'ai participé à la mise en place de systèmes d'informations intégrés dédiés aux analyses de corrélation génotypes-phénotypes dans le cadre d'études de maladies génétiques inflammatoires humaines (Pugnere, Ruiz et al. 2003; Sarrauste de Menthiere, Terriere et al. 2003).

Puis ma mission au sein du CIRAD, à partir de 2002, a consisté à mettre en place, dans un premier temps, une activité de support en bioinformatique pour mes collègues de l'unité. Leurs travaux génèrent des quantités souvent considérables d'informations présentant une large hétérogénéité, une forte évolutivité et qui requièrent la mise en place d'approches innovantes pour leur traitement et permettre d'en extraire du sens. J'ai donc aussi développé, en parallèle, des thématiques de recherche propres en bioinformatique.

J'ai pu valoriser mes travaux par 15 publications dans des journaux à comité de lecture depuis mon entrée au CIRAD. J'ai pris la responsabilité de l'équipe ID, Intégration des Données de l'UMR DAP, le 31 juillet 2007 en remplacement de Brigitte Courtois. J'ai été à l'origine et participé à l'émergence de la plateforme, South Green Bioinformatics Platform, http://southgreen.cirad.fr, qui est reconnue au niveau national et international.

2) Activités d'encadrement et d'enseignement

J'ai mis en place et participé à des formations en bioinformatique pour des chercheurs et étudiants, français et étrangers, dans différents contextes universitaires et professionnels.

Globalement, le nombre de bioinformaticiens dans les unités de recherche est largement insuffisant par rapport aux besoins. Par conséquent, il y a une forte demande de formation pratique en bioinformatique de la part des chercheurs et étudiants biologistes, de l'unité et d'autres instituts, et de partenaires étrangers.

Par conséquent, j'ai co-organisé avec mes collègues de l'équipe ID, et parfois en collaboration avec des chercheurs du Sud plusieurs sessions de formation d'initiation à la bioinformatique d'une ou

deux semaines. Ces formations abordent différents thèmes comme les bases de données biologiques sur le Web, les logiciels d'alignement de séquences, les logiciels de phylogénie moléculaire, la génomique comparative et l'annotation des génomes, et une introduction à UNIX. Des sessions ont été organisées à Madagascar (novembre 2006 et novembre 2008), à Montpellier (pour les partenaires du IRRDB, International Rubber Research and Development Board, 29 juin – 02 juillet 2006), en Malaisie (Bioinformatics Workshop, 23-25 juin 2008), au Sénégal (Module de Génomique végétale e-learning pour l'université de Dakar, juin 2009), en Colombie (Herramientas Genómicas y Bioinformáticas para la Investigación Agrícola, Pontificia Universidad Javeriana, Bogota, 1-3 Mars 2010), au Brésil (Workshop France-Brésil de Bioinformatique, Universidade Estadual de Santa Cruz, 08-12 novembre 2010, et Bioinformatics School, Campinas, Sao Paulo, 21-26 novembre 2011).

Je suis membre de l'équipe pédagogique en bioinformatique dans le module Génomique Végétale du Master 2 APIMET / SEPMET de Montpellier SupAgro. J'y donne environ 4h de cours/an. Pour cette formation, j'ai co-organisé, en 2011, un nouveau module optionnel, "Bioinformatique appliquée à l'analyse des séquences biologiques", ouverte à la formation doctorale et la formation continue (effectif 2011 : 42 participants, dont 9 doctorants). Je donne des cours magistraux pour le master M1 Biotechnologies des Plantes tropicales de l'Université Montpellier 2 (1h30/an).

J'accueille et encadre chaque année des étudiants en stage de Master 2 Bioinformatique ou Informatique, de différentes universités : Montpellier, Bordeaux, Orsay, Marseille, Toulouse, etc. De 2002 à 2010, j'ai encadré 19 étudiants (16 Master, 2 DUT, et 1 troisième année ingénieur).

J'ai co-encadré la thèse de Pierre Larmande, sur "Mutualiser et partager, un défi pour la génomique fonctionnelle végétale", soutenue le 20 décembre 2007, et la thèse de Julien Wollbrett sur "Génération semi-automatique de Services Web Sémantiques pour des bases de données relationnelles biologiques ", soutenue le 13 décembre 2011.

3) Animation de la recherche

a) Management l'équipe ID (Intégration des Données)

Je dirige l'équipe ID de l'UMR AGAP (anciennement UMR DAP), qui est actuellement constitué de 8 permanents, affiliés à différents instituts : 6 chercheurs CIRAD, 1 ingénieur d'étude INRA et un maître de conférence SupAgro. J'ai mené un travail de réflexion avec les chercheurs de l'équipe pour clarifier notre positionnement sur les axes de recherche en bioinformatique. J'ai ainsi monté le projet d'équipe pour l'évaluation AERES 2010 et le quadriennal de l'UMR AGAP 2011-2014.

b) Responsable scientifique de la plateforme bioinformatique South Green

Depuis 2007, nous avions dans l'équipe un certain nombre d'applications développées de

manière séparée, sans réelle intégration et coordination. J'ai donc travaillé à faire émerger une réelle plateforme bioinformatique intégrée et dédiée à la génétique et génomique des plantes tropicales. L'ensemble des développements, bases de données et logiciels, ont permis ainsi de faire émerger une plateforme en bioinformatique d'envergure nationale et internationale : South Green Bioinformatics Platform (http://southgreen.cirad.fr/).

Un certain nombre de ces développements intègrent d'autres partenaires institutionnels, comme Bioversity, l'IRD, et l'INRA, dans le cadre de consortium nationaux (partenariat avec les unités URGI de Versailles, BIO3P de Rennes) et internationaux (consortiums GMOD, Generic Model Organism Database et GCP, Generation Challenge Program). Ces collaborations permettent de mutualiser les efforts et de s'assurer que les applications développées soient assez modulaires et génériques pour être reprises et maintenues par une partie de la communauté internationale des bioinformaticiens.

J'ai œuvré à positionner durablement la plateforme bioinformatique South Green en participant au montage de collaborations régionales inter-instituts. Je suis ainsi coordinateur du sous-programme Bioinformatique du projet fédérateur ARCAD, financé par Agropolis Fondation (2010-2013). Nous avons obtenu la labellisation IBISA de la plateforme bioinformatique montpelliéraine commune IGH, ISEM, CIRAD, IRD et LIRMM. La reconnaissance IBISA nous a permis de nous rapprocher des autres plateformes du sud (IMGT, Marseille) pour constituer le nœud régional Renabi Grand-Sud (http://renabi.genouest.org/platforms/renabi-grand-sud/).

L'arrivée des larges projets de séquençage de génomes complets dans notre UMR, liée à l'évolution des techniques de séquençage, a accentué le besoin de mettre en place rapidement une plateforme capable de supporter le stockage et l'analyse de ces données. J'ai obtenu le financement par le CIRAD d'un cluster de calculs dédié à notre plateforme, qui est en production depuis début 2010, et a permis de gérer les analyses de larges projets internes à l'UMR (annotation des génomes du cacaoyer, bananier, clémentinier, etc.) et externes (analyse du génome de tilapia, de phytovirus, etc.). Une démarche de certification ISO9001 du cluster de calculs est en cours.

c) Liste des applications développées dans South Green

- TropGeneDB, Gestion des ressources génétiques et génomiques de plantes tropicales, http://tropgenedb.cirad.fr/
- CocoaGenDB, Portail Web qui croise les informations phénotypiques, génétiques et génomiques sur le cacaoyer, http://cocoagendb.cirad.fr/
- OryGenesDB, Exploration de la génétique inverse du riz, http://orygenesdb.cirad.fr/
- OryzaTagLine, Caractérisation phénotypique d'une collection de lignées d'insertion de T-DNA du riz, http://urgi.versailles.inra.fr/OryzaTagLine/

- SAT, Identification de microsatellites dans des banques enrichies, http://sat.cirad.fr/sat/
- ESTTik, Analyse de collection d'ESTs et Identification de SNPs, http://esttik.cirad.fr
- GreenPhyl, Prédiction d'orthologues par phylogénomique, http://greenphyl.cirad.fr
- GNPAnnot, Plateforme générique d'annotation structurale, fonctionnelle et comparative dédiée aux génomes de plantes et de leur bio-agresseurs, http://www.gnpannot.org/
- GenDiversity, Outil intégré pour l'analyse des données de diversité génétique, http://gendiversity.cirad.fr/Home
- Haplophyle, Réseaux haplotypiques graphiques à la lumière de données éco-géographiques, http://haplophyle.cirad.fr/
- MS-DMind, Gestion et exploration multi-échelle sequence-structure-fonction d'une famille de protéines, *en développement*
- SNIPLAY, Application intégrée pour l'analyse des données SNP, http://sniplay.cirad.fr

d) Réseaux de recherche

Mes activités de recherche m'ont permis d'être impliqué dans plusieurs réseaux nationaux et internationaux. Ainsi, j'ai été membre du comité de programme des conférences francophones JOBIM, Journées Ouvertes en Biologie, Informatique et Mathématiques 2010 et 2011, membre du comité local de la conférence internationale Biodiversity Information Standards, TDWG 2009, membre du comité scientifique du 3rd Conference of the Brazilian Association for Bioinformatics and Computational Biology 2007, membre de la cellule nationale Bioinformatique Verte en 2007 (mise en place par la direction scientifique Plantes et Produits du Végétal de l'INRA). Je suis éditeur associé de la revue BMC Genomics.

4) Publications : 22 publications dans des journaux à comité de lecture

a) Tableau de synthèse des facteurs d'impact

Revues	Nombre d'articles dans la revue	Facteur d'impact*	Rang parmi les auteurs
Nature Genetics	1	32.70	15/61
Nucleic Acids Research	8	7.31	16/18 ; 8/12 ; 2/10 ; 1/6 ; 2/3 ; 4/6 ; 2/6 ; 1/10
BMC Genomics	2	4.19	19/28 ; 4/25
BMC Bioinformatics	2	3.78	7/8 ; 5/5
Theoretical and Applied Genetics	1	3.78	6/9
Developmental and Comparative Immunology	2	3.24	8/13 ; 3/8
Rice	1	2.90	8/8
Immunogenetics	1	2.65	1/2
OMICS	1	2.27	5/11
Experimental and Clinical Immunogenetics	2	1.74**	3/4 ; 1/5
International Journal of Plant Genomics	1	-	4/19

** moyenne des facteurs d'impact sur les cinq dernières années, 2006-2010 (Thomson Reuters)*
*** facteur impact 2003, car fin de la revue en 2002 (Thomson Reuters)*

b) Liste des publications

Les étudiants en thèse que j'ai encadrés sont soulignés

- Dereeper, A., Nicolas, S., Le Cunff, L., Bacilieri, R., Doligez, A., Peros, J.P., Ruiz, M. and This, P. (2011) SNiPlay: a web-based tool for detection, management and analysis of SNPs. Application to grapevine diversity projects, BMC Bioinformatics, 12, 134.
- Argout, X., Salse, J., Aury, J.M., Guiltinan, M.J., Droc, G., Gouzy, J., Allegre, M., Chaparro, C., Legavre, T., Maximova, S.N., Abrouk, M., Murat, F., Fouet, O., Poulain, J., Ruiz, M., Roguet, Y., Rodier-Goud, M., Barbosa-Neto, J.F., Sabot, F., Kudrna, D., Ammiraju, J.S., Schuster, S.C., Carlson, J.E., Sallet, E., Schiex, T., Dievart, A., Kramer, M., Gelley, L., Shi, Z., Berard, A., Viot, C., Boccara, M., Risterucci, A.M., Guignon, V., Sabau, X., Axtell, M.J., Ma, Z., Zhang, Y., Brown, S., Bourge, M., Golser, W., Song, X., Clement, D., Rivallan, R., Tahi, M., Akaza, J.M., Pitollat, B., Gramacho, K., D'Hont, A., Brunel, D., Infante, D., Kebe, I., Costet, P., Wing, R., McCombie, W.R., Guiderdoni, E., Quetier, F., Panaud, O., Wincker, P., Bocs, S. and Lanaud, C. (2010) The genome of Theobroma cacao, Nat Genet.

- Courtois, B., Ahmadi, N., Khowaja, F., Price, A., Rami, J.-F., Frouin, J., Hamelin, C. and Ruiz, M. (2009) Rice Root Genetic Architecture: Meta-analysis from a Drought QTL Database, Rice, 2, 115-128.
- Argout, X., Fouet, O., Wincker, P., Gramacho, K., Legavre, T., Sabau, X., Risterucci, A.M., Da Silva, C., Cascardo, J., Allegre, M., Kuhn, D., Verica, J., Courtois, B., Loor, G., Babin, R., Sounigo, O., Ducamp, M., Guiltinan, M.J., Ruiz, M., Alemanno, L., Machado, R., Phillips, W., Schnell, R., Gilmour, M., Rosenquist, E., Butler, D., Maximova, S. and Lanaud, C. (2008) Towards the understanding of the cocoa transcriptome: Production and analysis of an exhaustive dataset of ESTs of Theobroma cacao L. generated from various tissues and under various conditions, BMC Genomics, 9, 512.
- Lescot, M., Piffanelli, P., Ciampi, A.Y., Ruiz, M., Blanc, G., Leebens-Mack, J., da Silva, F.R., Santos, C.M., D'Hont, A., Garsmeur, O., Vilarinhos, A.D., Kanamori, H., Matsumoto, T., Ronning, C.M., Cheung, F., Haas, B.J., Althoff, R., Arbogast, T., Hine, E., Pappas, G.J., Jr., Sasaki, T., Souza, M.T., Jr., Miller, R.N., Glaszmann, J.C. and Town, C.D. (2008) Insights into the Musa genome: syntenic relationships to rice and between Musa species, BMC Genomics, 9, 58.
- Bruskiewich, R., Senger, M., Davenport, G., Ruiz, M., Rouard, M., Hazekamp, T., Takeya, M., Doi, K., Satoh, K., Costa, M., Simon, R., Balaji, J., Akintunde, A., Mauleon, R., Wanchana, S., Shah, T., Anacleto, M., Portugal, A., Ulat, V.J., Thongjuea, S., Braak, K., Ritter, S., Dereeper, A., Skofic, M., Rojas, E., Martins, N., Pappas, G., Alamban, R., Almodiel, R., Barboza, L.H., Detras, J., Manansala, K., Mendoza, M.J., Morales, J., Peralta, B., Valerio, R., Zhang, Y., Gregorio, S., Hermocilla, J., Echavez, M., Yap, J.M., Farmer, A., Schiltz, G., Lee, J., Casstevens, T., Jaiswal, P., Meintjes, A., Wilkinson, M., Good, B., Wagner, J., Morris, J., Marshall, D., Collins, A., Kikuchi, S., Metz, T., McLaren, G. and van Hintum, T. (2008) The generation challenge programme platform: semantic standards and workbench for crop science, Int J Plant Genomics, 2008, 369601.
- Larmande, P., Gay, C., Lorieux, M., Perin, C., Bouniol, M., Droc, G., Sallaud, C., Perez, P., Barnola, I., Biderre-Petit, C., Martin, J., Morel, J.B., Johnson, A.A., Bourgis, F., Ghesquiere, A., Ruiz, M., Courtois, B. and Guiderdoni, E. (2008) Oryza Tag Line, a phenotypic mutant database for the Genoplante rice insertion line library, Nucleic Acids Res, 36, D1022-1027.
- Wanchana, S., Thongjuea, S., Ulat, V.J., Anacleto, M., Mauleon, R., Conte, M., Rouard, M., Ruiz, M., Krishnamurthy, N., Sjolander, K., van Hintum, T. and Bruskiewich, R.M. (2008) The Generation Challenge Programme comparative plant stress-responsive gene catalogue, Nucleic Acids Res, 36, D943-946.

- Dereeper, A., Argout, X., Billot, C., Rami, J.F. and Ruiz, M. (2007) SAT, a flexible and optimized Web application for SSR marker development, BMC Bioinformatics, 8, 465.
- Droc, G., Ruiz, M., Larmande, P., Pereira, A., Piffanelli, P., Morel, J.B., Dievart, A., Courtois, B., Guiderdoni, E. and Perin, C. (2006) OryGenesDB: a database for rice reverse genetics, Nucleic Acids Res, 34, D736-740.
- Bruskiewich, R., Davenport, G., Hazekamp, T., Metz, T., Ruiz, M., Simon, R., Takeya, M., Lee, J., Senger, M., McLaren, G. and Van Hintum, T. (2006) Generation Challenge Programme (GCP): standards for crop data, OMICS, 10, 215-219.
- Lefranc, M.P., Pommie, C., Kaas, Q., Duprat, E., Bosc, N., Guiraudou, D., Jean, C., Ruiz, M., Da Piedade, I., Rouard, M., Foulquier, E., Thouvenin, V. and Lefranc, G. (2005) IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains, Dev Comp Immunol, 29, 185-203.
- Ruiz, M., Rouard, M., Raboin, L.M., Lartaud, M., Lagoda, P. and Courtois, B. (2004) TropGENE-DB, a multi-tropical crop information system, Nucleic Acids Res, 32, D364-367.
- Kaas, Q., Ruiz, M. and Lefranc, M.P. (2004) IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data, Nucleic Acids Res, 32, D208-210.
- Clement, D., Lanaud, C., Sabau, X., Fouet, O., Le Cunff, L., Ruiz, E., Risterucci, A.M., Glaszmann, J.C. and Piffanelli, P. (2004) Creation of BAC genomic resources for cocoa (Theobroma cacao L.) for physical mapping of RGA containing BAC clones, Theor Appl Genet, 108, 1627-1634.
- Sarrauste de Menthiere, C., Terriere, S., Pugnere, D., Ruiz, M., Demaille, J. and Touitou, I. (2003) INFEVERS: the Registry for FMF and hereditary inflammatory disorders mutations, Nucleic Acids Res, 31, 282-285.
- Pugnere, D., Ruiz, M., Sarrauste de Menthiere, C., Masdoua, B., Demaille, J. and Touitou, I. (2003) The MetaFMF website: a high quality tool for meta-analysis of FMF, Nucleic Acids Res, 31, 286-290.
- Lefranc, M.P., Pommie, C., Ruiz, M., Giudicelli, V., Foulquier, E., Truong, L., Thouvenin-Contet, V. and Lefranc, G. (2003) IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains, Dev Comp Immunol, 27, 55-77.
- Ruiz, M. and Lefranc, M.P. (2002) IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures, Immunogenetics, 53, 857-883.

- Ruiz, M., Giudicelli, V., Ginestoux, C., Stoehr, P., Robinson, J., Bodmer, J., Marsh, S.G., Bontrop, R., Lemaitre, M., Lefranc, G., Chaume, D. and Lefranc, M.P. (2000) IMGT, the international ImMunoGeneTics database, Nucleic Acids Res, 28, 219-221.
- Scaviner, D., Barbie, V., Ruiz, M. and Lefranc, M.P. (1999) Protein displays of the human immunoglobulin heavy, kappa and lambda variable and joining regions, Exp Clin Immunogenet, 16, 234-240.
- Ruiz, M., Pallares, N., Contet, V., Barbi, V. and Lefranc, M.P. (1999) The human immunoglobulin heavy diversity (IGHD) and joining (IGHJ) segments, *Exp Clin Immunogenet*, **16**, 173-184.

5) Communications orales dans des conférences internationales

- Wollbrett, J., Larmande, P. and Ruiz, M. 2011 Towards Automatic Generation of Semantic Web Services for Relational Biological Databases. RED Fourth International Workshop on REsource Discovery. Heraklion, Greece.
- Argout X., Ruiz M., Fouet O., Lanaud C., Wincker P., Da Silva C., Courtois B. 2010. ESTtik, a semi-automatic cDNA sequence analysis and annotation pipeline including SSR and SNP search tools. In : 15th International Cocoa Research Conference : cocoa productivity, quality, profitability, human health and the environment. Cocoa Producers' Alliance, p. 501-506. International Cocoa Research Conference. 15, 2006-10-09/2006-10-14, San José, Costa Rica.
- Argout X., Ruiz M., Rouard M., Turnbull C.J., Lanaud C., Rosenquist E., Courtois B. 2010. CocoaGen DB : A web portal for crossing cocoa phenotypic, genetic and genomic data from ICGD and Tropgene DB database = CocoaGEN DB, une base de données intégrative internationale sur le cacao, avec des données phénotypiques, génétiques et génomiques. In : 15th International Cocoa Research Conference : cocoa productivity, quality, profitability, human health and the environment. Cocoa Producers' Alliance, p. 515-518. International Cocoa Research Conference. 15, 2006-10-09/2006-10-14, San José, Costa Rica.
- Risterucci A.M., Nansot G., Grangeon R., Lepitre V., Dereeper A., Argout X., Ruiz M., Billotte N. 2010. Development of guava microsatellite (SSR) markers using the SAT software. In : Rohde Wolfgang (ed.), Fermin Gustavo (ed.). Proceedings of the Second International Symposium on guava and other myrtaceae. Louvain : ISHS [Belgique], p. 113-119. (Acta Horticulturae, 849).
- Rouard M., Argout X., Bocs S., Conte M., Droc G., Guignon V., Hamelin C., Roux N., Ruiz M. 2010. Towards a bioinformatics platform for the Musa research community :. In : Abstracts of Plant and Animal Genomes XVIIIth Conference, San Diego, CA (USA), January 09-13, 2010.

[Online]. [S.l.] : s.n.. Plant and Animal Genomes Conference. 18, 2010-01-09/2010-01-13, San Diego, Etats-Unis.

- Argout X., Garcia D., Montoro P., Pujade-Renaud V., Ruiz M., Seguin M., Sidibé-Bocs S. 2009. Statement of transcriptomics and bioinformatics analyses conducted at CIRAD in rubber tree: Towards the genome analysis. In : 3rd IRRDB Workshop on the Hevea Genome and Transcriptome (Book of abstracts), 3-5 June, 2009, Montpellier, France. s.l. : s.n., p. 49. IRRDB Workshop on the Hevea Genome and Transcriptome. 3, 2009-06-03/2009-06-05, Montpellier, France.
- Ruiz M. 2009. Semantic standards for genomic analyses of the South and Mediterranean plants: the Generation Challenge Program use case : [Abstract]. In : ed. by Anna L. Weitzman. Proceedings of TDWG 2009 Annual Conference, Montpellier, France, 9-13 november 2009 . [Online]. Montpellier : Biodiversity Information Standards (TDWG), [1] p. TDWG 2009 Annual Conference, 2009-11-09/2009-11-13, Montpellier, France.
- Ahmadi N., Courtois B., Khowaja F.S., Perice A., Frouin J., Hamelin C., Ruiz M. 2007. Meta-analysis of QTLs involved in rice root development using a QTL database : [Abstract]. In : International Symposium Root Biology and MAS Strategies for Drought Resistance Imporvement in Rice, Bangalore, India, September 26-29, 2007. s.l. : s.n., 1 p. International Symposium Root Biology and MAS Strategies for Drought Resistance Imporvement in Rice, 2007-09-26/2007-09-29, Bangalore, Inde.
- Bruskiewich R., Davenport G., Senger M., Metz T., Ruiz M., Dereeper A., Takeya M., Hazekamp T., Rouard M., Simon R., Rojas E., Balaji J., Akintunde A., Costa M., Bink M., Wanchana S., Mauleon R.P., Morris J., Farmer A., Chandra S., Kikuchi S., Gaiji S., McLaren G., Van Hintum T. 2006. Crop information systems : the next generation [Abstract]. In : Plant Genomics European Meetings, Venice, 2006. Bologne : Ed. Avenue media, p. 77. Plant Genomics European Meetings. 5, 2006-10-11/2006-10-14, Venise, Italie.
- Piffanelli P., Ciampi A., Ruiz M., Rodrigues da Silva F., Lescot M., Papas G.J., Ronning C., Haas B., Wortman J., Frison E.A., Roux N., Miller R.N.G., Côte F., D'Hont A., Souza M., Glaszmann J.C., Town C. 2005. Analysis of the Musa genome from BAC sequencing and its comparison with rice [Abstract]. In : Abstracts of Plant and Animal Genomes XIIIth Conference, San Diego, CA (USA), January 15-19, 2005. [S.l.] : s.n.. Plant and Animal Genomes Conference. 13, 2005-01-15/2005-01-19, San Diego, Etats Unis.
- Ruiz M., Rouard M., Turnbull C.J., Orain R., Ford C.S., Raboin L.M., Lartaud M., Lanaud C., Clément D., Petithuguenin P., Wilkinson M.J., Hadley P., Brown S., Rosenquist E., Courtois B. 2005. A new international cocoa genetic database. In : 14th International Cocoa Research

Conference. Proceedings : hacia una economia sustentable del cacao - que estrategias para lograr este fin ?. Cocoa Producers' Alliance, p. 33-41. Conférence Internationale sur la Recherche Cacaoyère. 14, 2003-10-13/2003-10-18, Accra, Ghana.

- Piffanelli P., Ciampi A., Ruiz M., Rodrigues da Silva F., Papas G.J., Ronning C., Haas B., Wortman J., Frison E.A., Roux N., Miller R.N.G., Côte F., D'Hont A., Souza M., Glaszmann J.C., Town C. 2004. Comparative analysis of Musa and rice genome structure and organization [Abstract]. In : Picq Claudine (ed.), Vézina Anne (ed.). First International congress on Musa: harnessing research for improved livelihoods, 6-9 July 2004, Penang, Malaysia. Abstract guide. Montpellier : INIBAP, p. 21. International Congress on Musa: Harnessing Research for Improved Livelihoods. 1, 2004-07-06/2004-07-09, Penang, Malaisie.

Travaux de recherche

1) Contexte scientifique

Les projets de génétique et de génomique structurale et fonctionnelle produisent des quantités de plus en plus considérables de données. Afin de pouvoir exploiter au mieux cette masse d'informations, il est nécessaire (i) d'organiser les données et de les rendre facilement accessibles à la communauté internationale par la mise en place de systèmes d'information intégrés, (ii) d'analyser des données par le développement ou l'adaptation d'outils bioinformatiques, (iii) de valoriser l'information disponible sur les espèces modèles par l'analyse comparative, structurale et fonctionnelle, de génomes apparentés.

Au sein de l'équipe ID, j'ai développé des projets de recherche sur les thèmes suivants: (i) Représentation des connaissances, organisation et gestion de l'information liée aux domaines de la génétique et génomique végétale, (ii) Intégration sémantique des données en génomique végétale, et (iii) Annotation automatique des données génomiques et génomique comparative.

2) Représentation des connaissances, organisation et gestion de l'information liée aux domaines de la génétique et génomique végétale

Je me suis intéressé à la conception de méthodes et de développements informatiques innovants dans le domaine des Systèmes d'Information (SI) dédiés aux données génétiques, génomiques et phénotypiques végétales. Ces SI à caractère intégratif et évolutif doivent être capables de prendre en compte l'apport de nouvelles connaissances ou l'élargissement du champ d'application de nos projets.

J'ai participé au développement de systèmes d'information pour la génétique et génomique des plantes tropicales, librement accessibles sur le Web, à la communauté internationale. Mes activités vont de la conception des projets de développement, et du suivi des projets, à la valorisation des applications développées. Ces développements viennent largement en support des activités de recherche des biologistes de l'unité, mais servent aussi de support pour des projets de recherche méthodologique en bioinformatique.

J'ai participé au développement des systèmes d'information TropGeneDB (Ruiz, Rouard et al. 2004), CocoaGenDB (Argout, Salse et al. 2010), OryGenesDB (Droc, Ruiz et al. 2006), OryzaTagLine (Larmande, Gay et al. 2008), ESTtik (Argout, Fouet et al. 2008), SAT, SSR Analysis Tool (Dereeper, Argout et al. 2007), et SNIPLAY (Dereeper, Nicolas et al. 2011). Mes implications dans ces différents développements ont évolué depuis une participation directe au codage informatique jusqu'à une

coordination du projet impliquant plusieurs développeurs (ingénieurs permanents et CDD, collaborateurs étrangers, étudiants en stage).

Le caractère évolutif et multi-plantes des données gérées par notre équipe m'a conduit notamment à proposer un modèle conceptuel de données générique pour le système d'information TropGeneDB (Figure 1). Ce modèle, est actuellement appliqué à neuf modules plantes différents (TropGeneDB, http://tropgenedb.cirad.fr/). J'ai aussi conçu un générateur automatique d'interfaces Web de requêtes complexes, GenTic2 (*non publié*), utilisé actuellement pour différents bases de données : TropGeneDB, EuriGenDB (http://eurigendb.cirad.fr), et Clonothèque Café (*accès intranet*). L'ensemble de ces méthodes a pour objectifs de pouvoir proposer rapidement de nouveaux modules plante et intégrer des nouveaux types de données, et donc de pouvoir s'adapter aux données génomiques fortement évolutives.

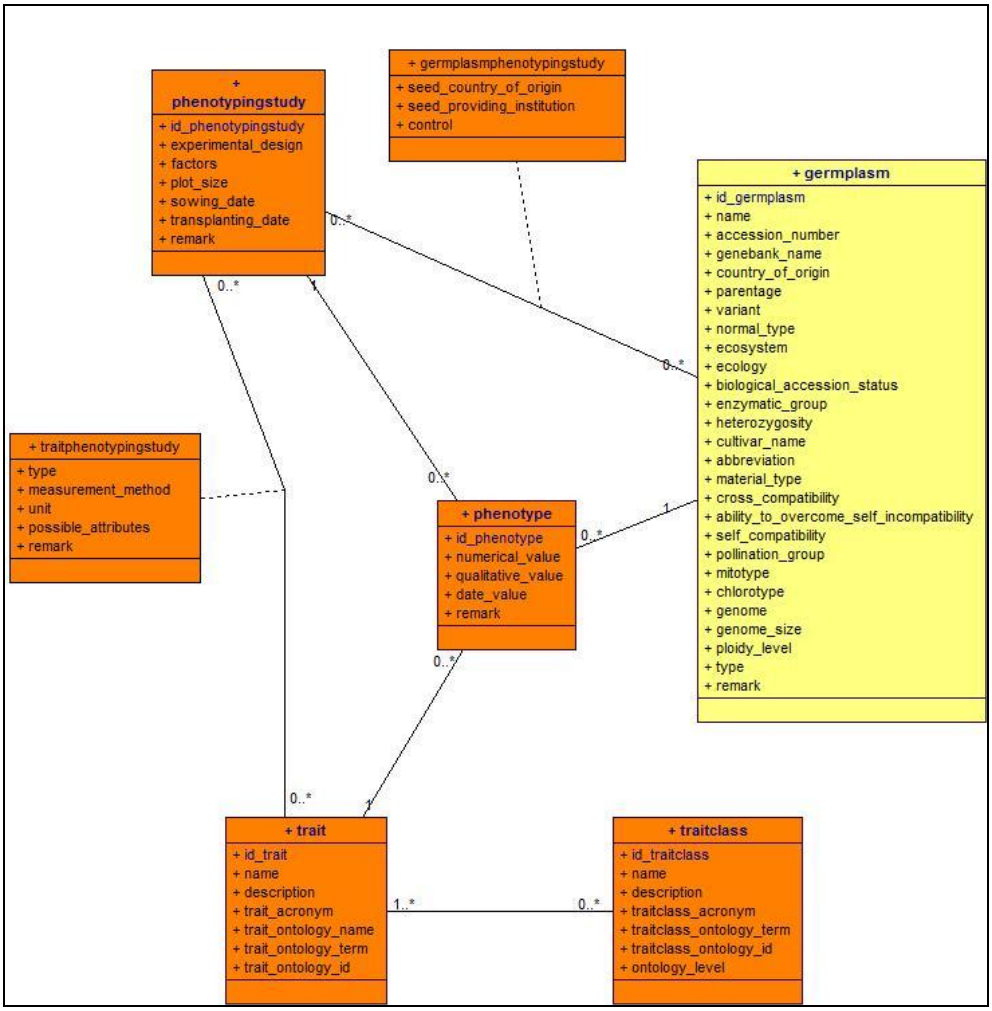

Figure 1. Extrait du modèle générique, UML ou Unified Modeling Language, de TropGeneDB concernant les études phénotypiques. Dans ce modèle, il est, par exemple, possible de rajouter

facilement tout nouveau type de caractères phénotypes et d'intégrer des correspondances avec des termes ontologiques existants.

3) **Intégration sémantique des données en génomique végétale**

Les sources de données en génomique végétale sont multiples, réparties et hétérogènes, et essentiellement accessibles par le Web. La possibilité pour les chercheurs de localiser, récupérer, intégrer et analyser, rapidement et régulièrement, l'information pertinente dans cette masse de données, reste un problème critique et important dans le domaine de la génomique (Howe, Costanzo et al. 2008). Pour résoudre cette question au niveau du développement des systèmes d'information, il est nécessaire de dépasser différents niveaux d'hétérogénéité :

- l'hétérogénéité des infrastructures logicielles, dû à l'utilisation de différents systèmes de gestion de bases de données, de différents protocoles de rapatriement des données, et de différents langages informatiques.
- l'hétérogénéité syntaxique, dans le cas des formats pour décrire le contenu des sources, et des formats d'exportation des données. Ce type d'hétérogénéité est aussi lié à la diversité des modèles de données.
- l'hétérogénéité sémantique qui recouvre plusieurs aspects : (i) le degré de spécialisation de chaque base de données, chacune se focalisant sur un type d'objet biologique et ne représentant pas l'information avec le même niveau de détails, comme par exemple les bases dites généralistes *versus* spécialisées, (ii) la diversité des modes de désignation des concepts biologiques selon différents vocabulaires entraînant des conflits sémantiques, (iii) le problème des identifiants différents pour un même objet biologique.

L'intégration sémantique des données est vraisemblablement l'approche la plus difficile à mettre en place, mais aussi la plus prometteuse en terme d'avancée des connaissances (Stein 2008). Mon projet de recherche essaie de résoudre l'hétérogénéité sémantique des données, en s'appuyant sur l'utilisation des ontologies, dont les applications n'ont cessé de progresser dans le domaine de la génomique (Rubin, Shah et al. 2008).

L'ontologie est un modèle de données représentatif d'un ensemble de concepts dans un domaine de connaissance, ainsi que des relations entre ces concepts. Elle est employée pour raisonner à propos des objets du domaine concerné. Les ontologies sont non seulement utiles dans l'intégration des données, mais aussi dans l'exploitation des données et l'extraction de nouvelles connaissances. L'utilisation des ontologies est d'autant plus pertinente que le flot des données produites ne cesse d'augmenter. Ce type d'approche est bien avancé dans le domaine biomédical (Chen, Ding et al. 2009; Holford, Khurana et al. 2010), mais encore peu développé en génomique végétale. Or les besoins sont nombreux et les spécificités du domaine sont majeures, notamment

lorsque les chercheurs essaient de comprendre par quels mécanismes génétiques une plante cultivée s'adapte à un environnement donné.

a) Le système GCP Pantheon

Dans ce contexte, mes travaux de recherche se sont orientés sur la conception d'adaptateurs sémantiques qui génèrent et exploitent les correspondances entre les schémas des bases de données biologiques et les ontologies de domaine existantes (Figure 2).

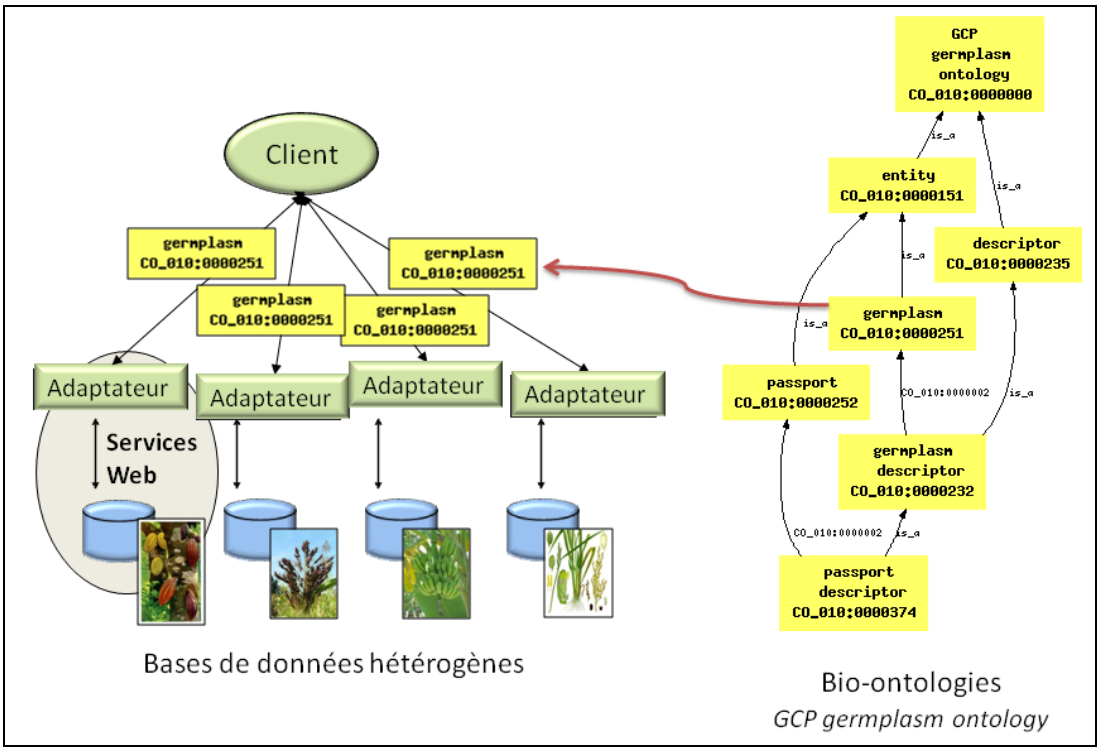

Figure 2. Les adaptateurs sémantiques génèrent et exploitent les correspondances entre les schémas hétérogènes des bases de données biologiques et les bio-ontologies existantes. Dans cet exemple, l'application cliente récupérera les données reliées au même concept *germplasm* provenant de l'ontologie *GCP germplasm ontology*. L'adaptateur peut être intégré dans une application de type Web service.

Suivant cette approche, j'ai eu une participation active au sein du projet international du Generation Challenge Program (GCP) SP4 (Subprogramme 4, Bioinformatics and Crop Information Systems), qui, tout en se basant sur les initiatives déjà existantes (Gene Ontology, Plant Ontology, Chado, etc.), a développé de nouveaux standards complémentaires pour la description des données génétiques (Bruskiewich, Davenport et al. 2006). Dans le cadre de ce projet GCP, j'ai aussi participé au développement d'un système à base de médiation (Wiederhold 1992), GCP Pantheon (http://pantheon.generationcp.org/), qui utilise les ontologies et la plateforme de Web Services BioMoby (http://biomoby.org/). Le médiateur intègre les adaptateurs à travers un schéma global, qui

dans notre cas correspond à une ontologie de domaine. Les applications développées sont librement disponibles et ont fait l'objet de publications (Bruskiewich, Senger et al. 2008; Wanchana, Thongjuea et al. 2008). Dans ce contexte j'ai coordonné le développement d'un module de GCP Pantheon, GenDiversity (Ruiz 2009), qui permet de récupérer et de filtrer des données de génotypage (SSR, SNP) et passeports, provenant de différentes sources de données distantes (TropGene, Singer, GCP central registry). L'application comprend en outre un module qui permet de contrôler la cohérence des données, et la connexion à différents outils d'analyse génétique (Structure, Darwin, Haploview). Un prototype est disponible à l'adresse http://gendiversity.cirad.fr/Home.

b) La génération automatique de Web Services Sémantiques

Une des étapes limitantes majeures pour l'exploitation optimale des ontologies dans l'intégration sémantique de données provenant de sources de données hétérogènes, est le coût en terme de temps, pour la création de correspondances entre des bases de données de type relationnelles et des ontologies stockées sous forme de DAG, Directed Acyclic Graph, ou de langages à l'expressivité plus riche comme OWL, Web Ontology Language (Moreira and Musen 2007). Ainsi, le développement d'adaptateurs sémantiques, permettant de se connecter à des sources de données hétérogènes, se heurte à un certain nombre de difficultés :

- une compréhension de l'ontologie du domaine, et le choix adéquat des concepts à utiliser,
- le degré de généralisation des concepts : quel niveau de spécialisation doit-on choisir ? Quelles règles pour faire les correspondances, ou "mapping", entre des éléments d'une source de données, comme les métadonnées associées, et des concepts d'une ontologie de domaine (Rahm and Bernstein 2001) ?
- la génération automatique des adaptateurs, et un contrôle de la cohérence sémantique entre les différents adaptateurs générés.

Pour dépasser ces limitations, je cherche à développer des méthodes originales de création automatique d'adaptateurs, basées sur une ontologie de domaine. J'ai appliqué ces méthodes dans le développement de Web Services Sémantiques. En effet, les Web Services deviennent de plus en plus utilisés dans le cadre de l'accès distant à des ressources bioinformatiques majeures telles qu'EMBL-EBI, KECG, ou NCBI. De nombreux projets apparaissent pour décrire ces Web Services avec des annotations sémantiques utilisant des ontologies, comme SSwap (Gessler, Schiltz et al. 2009), SADI (Wilkinson, McCarthy et al. 2010), BioMoby (Wilkinson, Senger et al. 2008) et BioCatalogue (Bhagat, Tanoh et al. 2010).

La thèse de Julien Wollbrett, que j'ai encadré de 2008 à 2011, a exploré les voies d'automatisation possible dans l'intégration de données végétales, notamment dans le cadre de projets d'analyse de la diversité génétique des plantes. BioMoby Converter est une première

approche pour accélérer l'utilisation des ontologies dans le développement des Web Services. C'est un module d'intégration automatique d'une ontologie de domaine dans la plateforme BioMoby (Wollbrett, Larmande et al. 2009). Mais nous avons rapidement proposé une infrastructure plus large et générique, et développé des méthodes de génération automatique de Web Services Sémantiques pour les bases de données relationnelles biologiques (Wollbrett, Larmande et al. 2011).

Nous nous sommes focalisés, dans un premier temps, sur la connexion avec des systèmes de gestion de bases de données relationnelles, qui sont largement utilisées pour stocker, gérer et interroger les données biologiques. L'implémentation de nouveaux Web Services interrogeant ces systèmes, représentent une tâche coûteuse en terme de temps de développement. Pour accélérer le développement de tels Web Services, les différentes étapes abordées ont été :

(i) la génération automatique de requêtes à partir des sources de données annotées avec les concepts ontologiques. Pour cette étape, nous nous appuyons sur l'intégration de différentes approches : la transformation automatique des schémas de bases de données sous forme de vue RDF, Resource Description Framework (W3C 2004), l'annotation des schémas avec des langages de type D2RQ (Bizer 2004), la développement d'algorithmes de parcours de graphes à la recherche du plus court chemin reliant les concepts ontologiques d'entrée et de sortie.

(ii) la génération automatique des Web Services. Cette phase consiste à intégrer dans un service Web la requête créée en amont. Les données contenues dans la source sont donc rendues accessible sur le Web via un annuaire de Web services. Ces Web services intègrent de la sémantique en étant annotés automatiquement avec les concepts ontologiques sélectionnés pour la création de la requête.

Un certain nombre de Web Services générés par notre méthode sont fonctionnels et enregistrés dans l'annuaire BioCatalogue (Bhagat, Tanoh et al. 2010).

En collaboration avec l'équipe Zenith INRIA de Patrick Valduriez (http://www-sop.inria.fr/teams/zenith/), nous commençons à travailler sur le mapping automatique des schémas, en exploitant les méthodes développées dans le cadre de la plateforme WebSmatch (ZENITH Team 2010: http://websmatch.gforge.inria.fr/). Nous utiliserons des méthodes intégrées à base de différents algorithmes de "schema matching" avec des processus d'apprentissage à partir des mappings déjà validés par des annotateurs. Ce projet nécessitera vraisemblablement l'adaptation des méthodes de mapping existantes au domaine de la biologie et, plus particulièrement, de la génomique végétale qui a ses logiques propres de terminologie et d'organisation des données.

4) Annotation automatique des données génomiques et génomique comparative

La diversité des espèces travaillées dans l'unité nous a conduit, dans l'équipe ID, à nous intéresser particulièrement à l'analyse comparative des génomes, en travaillant dans un premier temps sur les séquences de grands fragments génomiques, puis, par la suite, sur les transcriptomes et génomes complets. J'ai personnellement apporté un appui méthodologique à l'analyse bioinformatique des données issues des travaux de séquençage de grands fragments génomiques de bananier, en comparaison avec le génome du riz (Lescot, Piffanelli et al. 2008).

En terme de méthodes d'analyse, nous cherchons à dépasser la simple comparaison de séquences, en intégrant aussi les domaines protéiques, la reconstruction d'arbres phylogénétiques, la prédiction de relations d'orthologie et de paralogie, le contexte génomique, et la reconstruction de groupes de synténie. Dans ce contexte, l'équipe ID a mis en place des outils originaux d'annotation et de phylogénie moléculaire nécessaires aux analyses comparatives de séquences génomiques : GNPAnnot (Bocs 2010), et GreenPhyl (Rouard, Guignon et al. 2010).

GNPAnnot est un système communautaire pour définir la structure, la fonction d'objets génomiques de séquences eucaryotes. Ce système comprend (i) des chaînes de traitement semi-automatique de prédiction d'objets génomiques (gènes, éléments transposables), (ii) des interfaces annotateurs pour la validation manuelle des prédictions et l'exploration de données, et (iii) des entrepôts d'objets génomiques assurant l'interopérabilité vers d'autres systèmes (annotations de référence d'organisme modèle, prédictions automatiques, validations manuelles et expérimentales, outils statistiques).

GreenPhyl permet, à partir de l'ensemble des séquences protéiques déduites des génomes séquencés de plantes, de déterminer les familles de gènes par une méthode de clustering semi-automatique, suivi de la détermination des groupes orthologues par phylogénomique (Figure 3).

L'ensemble de ces développements, et l'expertise acquise, a permis à notre équipe de coordonner l'analyse bioinformatique du génome complet du cacaoyer (Argout, Salse et al. 2010). Dans ce projet je me suis particulièrement impliqué dans la coordination de l'analyse comparative des familles de gènes.

Actuellement, notre équipe participe à l'analyse bioinformatique des génomes complets du bananier, du caféier, et du clémentinier. Je suis particulièrement impliqué dans l'analyse de la densité en SNPs le long du génome du clémentinier, en comparaison avec le génome de l'oranger, afin de reconstruire l'histoire évolutive du clémentinier à partir des hybridations successives entre pamplemoussier et mandarinier.

Je coordonne les analyses bioinformatiques du sous-projet SP1 (*Comparative population genomics in wild and crop plants*) du projet ARCAD. Nous réalisons des comparaisons intra et inter-espèces sur une dizaine d'espèces cultivés, ainsi que sur leurs ancêtres sauvages, à partir de données NGS (Next Generation Sequencing) transcriptomiques. Dans le cadre de ce projet, nous avons notamment initié la mise en place d'approches phylogénétiques originales dans l'analyse des séquences "reads", de type Solexa, mappées sur des génomes de référence. Ces méthodes tentent de discriminer les erreurs de "mapping" dus aux familles multigéniques ou aux polymorphismes (Dufayard and Ruiz 2011).

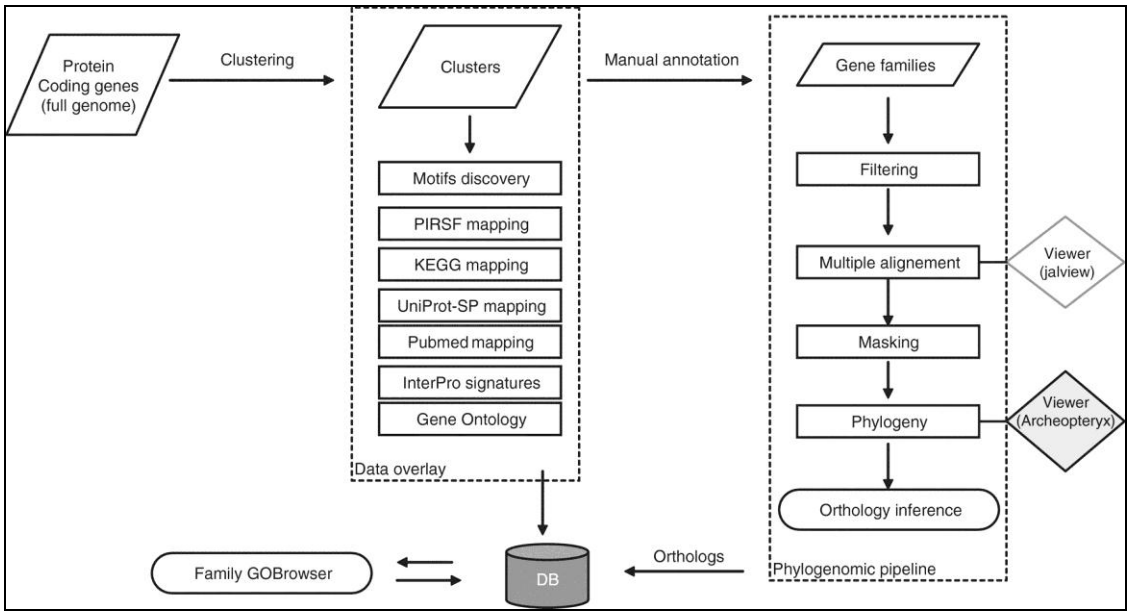

Figure 3. Les chaînes de traitement mis en place dans GreenPhyl, intègrent la détermination des familles de gènes par une méthode de clustering semi-automatique, la recherche de références croisées dans les différentes bases de données internationales, l'annotation manuelle des familles de gènes, et la détermination des groupes orthologues par phylogénomique (Rouard, Guignon et al. 2010).

5) Projet de recherche

a) Contexte scientifique

Un des apports majeurs de la génomique végétale est une compréhension accrue des évènements de domestication et d'adaptation des plantes cultivées, et notamment des plantes cruciales pour l'alimentation des populations du Sud (riz, sorgho, bananier, manioc, etc.). Les projets de génomique végétale intègrent de plus en plus les nouvelles technologies de séquençage à très haut débit (Next Generation Sequencing ou NGS). Ces technologies révolutionnent non seulement les approches expérimentales en biologie, mais aussi le traitement bioinformatique de tels volumes de

données qui exige le développement de nouvelles méthodes d'analyse suffisamment performantes pour intégrer ce changement d'échelle. L'application des NGS ne se limite pas au séquençage de nouveaux génomes, mais touche aussi le reséquençage de génomes, l'identification de variations génomiques telles que les SNPs, ou la transcriptomique. En parallèle, le développement des méthodes de phénotypage à haut-débit engendre aussi une masse d'informations considérables, qui couplées aux données de séquençage, permettent notamment des études d'association à l'échelle des génomes (Huang, Wei et al. 2010). Dans ce contexte, la mise en place d'approches innovantes pour intégrer et analyser les données biologiques végétales est plus que jamais nécessaire (Stein 2008).

b) Analyse intégrée de la diversité des plantes cultivées

Pour traiter les données génomiques, la communauté des chercheurs a créé de nombreuses méthodes d'analyse, utilisant de multiples paramètres. En général, différentes méthodes de résolution d'un même problème peuvent être proposées, ayant des niveaux de précision, des temps d'exécution et/ou des connaissances préalables associées différents. Ces analyses sont de plus en plus réalisées par des chaînes de traitement automatiques à haut-débit, ou "workflows". L'information liée à ces workflows, comme les objectifs de l'analyse, les algorithmes utilisés, les paramètres utilisés, ou les données intermédiaires produites avec leurs formats, sont actuellement difficiles à stocker, partager, et mutualiser sans l'utilisation d'approches sémantiques. Par conséquent, j'étudierai la possibilité d'étendre ma démarche, à base de Web Services Sémantiques (Shadbolt, Berners-Lee et al. 2006), dans des architectures de gestion de workflows de type Galaxy (Giardine, Riemer et al. 2005) ou Taverna (Hull, Wolstencroft et al. 2006).

J'appliquerai ces développements pour l'analyse intégrée de la diversité génétique des plantes cultivées étudiées dans l'UMR. Le cas d'utilisation est pertinent car les données sont très différentes en nature (génétiques, phénotypiques, environnements), produites par différents laboratoires et dispersées dans des bases de données "plantes" différentes. Je me focaliserai sur des plantes d'intérêt économique majeur, pour lesquelles de vastes projets de reséquençage à haut débit sont déjà prévus (riz, vigne), ou pour lesquelles les génomes sont juste séquencés ou en cours de séquençage (bananier, caféier, cacaoyer, citrus, palmier à huile, et hévéa).

L'objectif est d'intégrer des données phénotypiques, passeports, génotypiques (SNP, DarT, microsatellites, etc.), QTL, d'études d'association, et des annotations de génomes séquencés (Figure 4). Les sources de données seront connectées à des chaînes de traitements, ou "workflows", dédiés à la découverte et l'analyse de polymorphismes à partir des données de type NGS (454, Solexa, etc.). Ce travail s'inscrit dans la continuité du projet SNiPlay, application intégrée d'analyse des données SNP, que nous avons mis en place récemment (Dereeper, Nicolas et al. 2011).

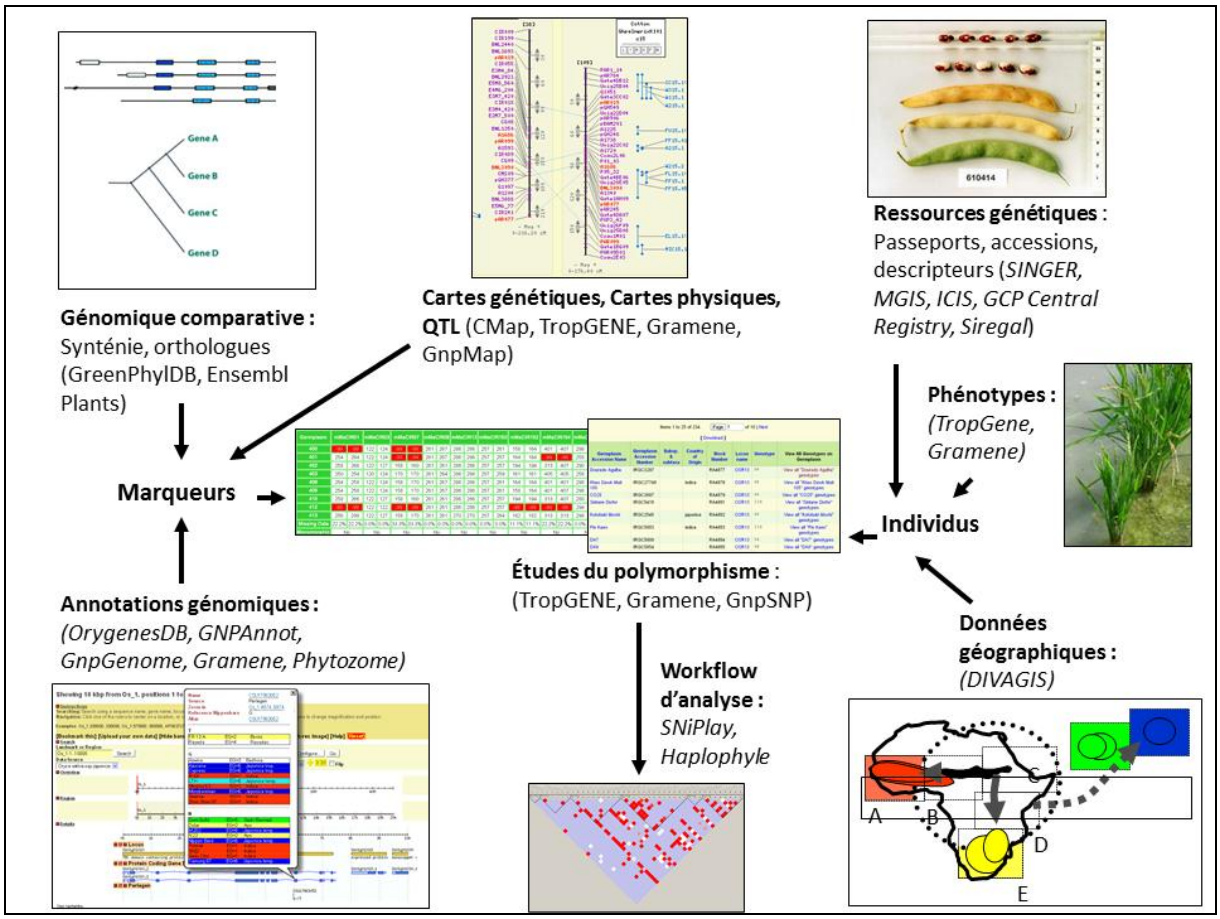

Figure 4. Dans cet exemple, nous recherchons des études de polymorphisme existantes, dans des bases de données comme TropGene, GnpSNP, ou Gramene, en fonction de critères sur les individus (ressources génétiques, phénotypes ou données géographiques), et sur les marqueurs (positions sur les cartes génétiques et physiques, annotations génomiques associées, relations d'orthologie et de synténie). Les données de polymorphismes récupérées peuvent être envoyées dans des workflows d'analyse de type SNiPlay ou Haplophyle.

Le système de création automatique d'adaptateurs sémantiques permettra à nos workflows d'évoluer et d'intégrer rapidement de nouveaux types de données et de nouvelles méthodes d'analyse. En facilitant l'intégration rapide de différents types de données, et leurs analyses, en parallèle, par différentes méthodes, avec différents paramètres, mon approche peut permettre de mieux comparer les workflows développés et de systématiser l'analyse de la performance de ces méthodes : temps de calcul, sensibilité, spécificité, etc. Actuellement, par exemple, il est très difficile de faire une analyse exhaustive de l'influence des différentes méthodes d'analyse des données NGS, comme les étapes de nettoyage des séquences courtes (reads), d'assemblage ou de mapping de ces reads sur une séquence de référence, sur les résultats des analyses ultérieures comme l'analyse du

polymorphisme, la structure et l'évolution des familles de gènes, ou le niveau d'expression des différentes séquences.

Evidemment, ce travail doit permettre d'extraire de nouvelles connaissances à partir de la masse de données disponibles. Ainsi, la corrélation de l'analyse des réseaux d'haplotypes et des informations géographiques permet de développer des analyses phylogéographiques et confère une grande puissance de résolution des événements de domestication et une meilleure compréhension de l'adaptation des plantes (Saisho and Purugganan 2007). Cependant, il n'existe pas actuellement de méthodes automatiques qui permettent une telle analyse, à haut-débit, et qui puisse gérer la masse considérable de données SNP à venir grâce notamment aux méthodes NGS. Nous nous proposons de développer un processus automatique d'analyse, qui comprend la définition des haplotypes des individus et des blocs d'haplotypes, la construction de réseaux phylogénétiques à partir des haplotypes générés, et la corrélation des données d'haplotypes avec des données extérieures, notamment géographiques. La mise en place de cette chaîne de traitements automatisée permettra d'envisager des méthodes d'optimisation dans la définition des réseaux d'haplotypes.

Enfin la mise en relation des données sur les haplotypes et des informations géographiques permettra des analyses phylogéographiques, de manière dynamique, en se connectant sur des ressources GIS (Geographic Information Systems) distantes. Ce travail s'insère dans la continuité du projet Haplophyle (http://haplophyle.cirad.fr).

c) Mise en place de bases de connaissances multi-échelles de familles de gènes

Pour mieux comprendre les rôles et l'évolution des familles de gènes impliquées dans les processus d'adaptation, de domestication et de sélection des plantes, il est souhaitable d'intégrer un certain nombre de ressources qui correspondent à différentes échelles de connaissance (Figure 5). Dans le cas de génomes comme le riz, l'ensemble des données utiles proviennent de bases de données génomiques (Genbank, RAP-DB, EnsemblPlant, OryGenesDB), de voies métaboliques (KEGG), de QTL et marqueurs (TropGene, Gramene), de protéines (UniProt), de domaines protéiques (Interpro), de mutants (OryzaTagLine), de génomique comparative (GreenPhyl), de ressources génétiques (IRIS), et de phénotypes (TropGene, Gramene).

Cependant la plupart des bases de données sont construites autour d'une entité biologique principale. Cette entité correspond à l'élément central du modèle de données utilisé dans l'implémentation de la base de données. Ce focus peut correspondre aux annotations d'un génome reliées par des coordonnées linéaires, [Ensembl (Flicek, Amode et al. 2011), Chado (Mungall and Emmert 2007)], aux séquences protéiques, [Swiss-Prot, (Schneider, Lane et al. 2009)], ou à des structures protéiques, [Columba (Trissl, Rother et al. 2005)]. Avec ce type de représentation, les

relations entre les caractéristiques moléculaires d'une protéine, sa séquence, sa structure 3D, et les fonctions biologiques dans lesquelles elle est impliquée, sont difficiles à établir. Dans les bases de données, les fonctions biologiques sont souvent décrites sous forme de langage naturel, sans réelle standardisation, ce qui accroit potentiellement les sources d'erreurs, et empêche une exploitation automatique des données fonctionnelles. A l'heure actuelle, il n'existe pas de systèmes qui permettent de relier facilement toutes les données moléculaires pertinentes pour la compréhension d'un processus biologique, depuis les annotations d'une séquence de gène dans le génome, en passant par les propriétés physico-chimiques d'un acide aminé dans son interaction avec un ligand, jusqu'aux réactions en jeu dans un stress biotique ou abiotique, par exemple.

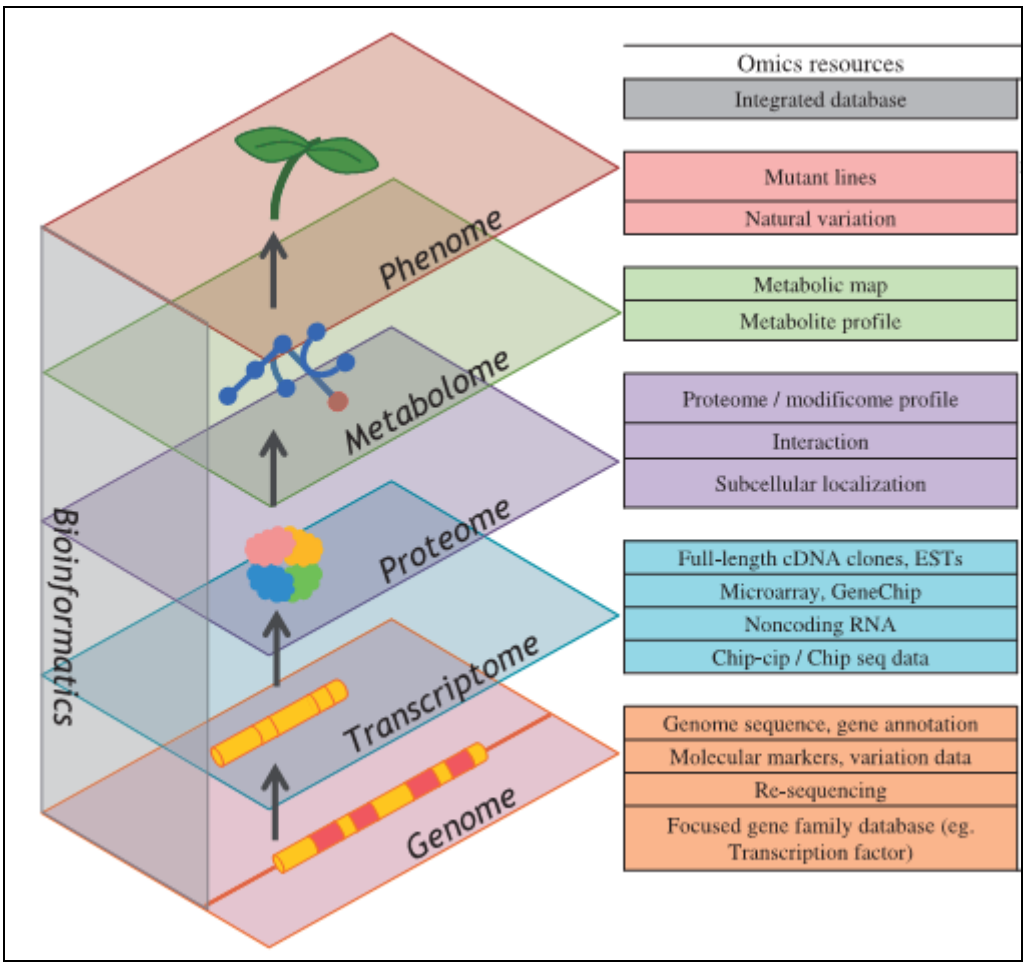

Figure 5. Représentation des différentes échelles de connaissance et des types de données associées (Mochida and Shinozaki 2010).

De nouveaux modèles de connaissance ont été développés afin de représenter les différentes échelles impliquées dans les relations séquence-structure-fonction. Il a été démontré qu'une modélisation centrée sur le positionnement standardisé des acides aminés permet une

meilleure compréhension des corrélations séquence-structure (Magdelaine-Beuzelin, Kaas et al. 2007). Parallèlement, un nouveau schéma multi échelle, BioPsi, permet la description des processus biologiques en utilisant un nombre limité d'actions élémentaires (Maziere, Parisey et al. 2007; Peres, Felicori et al. 2010). Je propose d'intégrer ces différentes approches afin de développer un nouveau type de systèmes d'information qui permet d'avoir une représentation standardisée des relations séquence-structure-fonction. Le système devra être assez générique pour être appliqué à tout type de famille protéique.

J'appliquerai notamment cette démarche dans le cadre du projet ANR MS-DMind que je coordonne, sur la période 2009-2012. Dans ce projet, nous effectuons une nouvelle classification, basée sur leurs structures 3D, de la superfamille des protéines ns-LTPs, non-specific Lipid Transfer Proteins (Fleury, Gautier et al. 2011). Les ns-LTPs sont de petites protéines, interagissant avec différents types de lipides, et présentes chez tous les végétaux. Elles sont de bons candidats pour expliquer des phénomènes aussi variés que la synthèse de la cutine, la mobilisation des lipides, ou la résistance aux pathogènes. Elles offrent un polymorphisme naturel important qui permet d'explorer comment des variations de structures primaire et tertiaire peuvent moduler leur fonctionnalité. Elles sont donc un bon modèle pour valider notre nouvelle approche d'intégration multi-échelle. Ce travail s'intègre aussi dans la continuité du projet Orylink (Droc, Périn et al. 2009) d'application de requêtes multi-bases pour la récupération de données fonctionnelles sur les gènes du riz.

Ce type d'approche devrait permettre de mieux cerner les relations complexes entre évolution, diversité et fonctions des familles de gènes, en facilitant les recoupements entre plusieurs domaines de connaissance.

d) Raisonnement automatique à partir d'une base de connaissances pour l'annotation des génomes

Avec l'arrivée des séquences complètes de nombreux nouveaux génomes de plantes, une limitation majeure de l'exploitation des données sera l'exhaustivité et la qualité des annotations produites. L'annotation de gènes est l'étape primordiale de l'analyse de séquences nucléiques et nécessite de définir la structure des gènes, la fonction des protéines et les relations d'homologie entre gènes. Même s'il existe de grandes variations entre les espèces eucaryotes, on retiendra que, dans l'ensemble, l'annotation des génomes est difficile pour plusieurs raisons : polyploïdie, proportion importante de régions non codantes et de régions répétées, structure des gènes morcelés en intron/exon. Des logiciels de prédiction automatique des gènes ont été développés mais ils ne remplacent pas l'annotation manuelle si l'on veut des annotations de haute qualité.

Nous avons mis en place dans l'équipe un système d'annotation collaboratif et intégré,

GNPAnnot (Bocs 2010), qui permet de définir la structure, la fonction d'objets génomiques de séquences eucaryotes, à la lumière des résultats de génomique comparative. Cependant, le processus d'annotation manuelle repose encore entièrement sur les épaules de l'annotateur expert qui doit valider, croiser, vérifier et intégrer les données et résultats générés par le système pour produire une annotation cohérente.

J'envisage ici la conception et la mise au point d'une base de connaissances qui reflète les connaissances des experts et qui permette l'annotation des données génomiques par une utilisation optimale des informations disponibles. Cela demande de définir une stratégie d'annotation fondée sur l'expertise des biologistes et les ontologies biologiques existantes, et la mise en place d'un ensemble de règles expertes acquises en partie par apprentissage automatique, ainsi que d'un module de raisonnement capable de mettre en œuvre cette stratégie. A chaque règle sera associé un coefficient de fiabilité qui reflètera la qualité de l'annotation qu'elle propose. Cette stratégie pourrait aider à la détection de nouveaux gènes et de nouvelles relations entre gènes. Ce travail peut être aussi très utile dans le cas de la comparaison de plusieurs annotations réalisées par différentes équipes, avec des méthodes différentes ou des versions de génomes différentes. Si l'on prend l'exemple du génome de la vigne, la communauté des chercheurs peut travailler sur les annotations de l'URGI, versions 8X et 12X (http://urgi.versailles.inra.fr/cgi-bin/gbrowse/), de l'URGV (FLAGdb++, http://urgv.evry.inra.fr/FLAGdb), du Genoscope (http://www.genoscope.cns.fr/vitis), ou du CRIBI (http://genomics.cribi.unipd.it/Grape_Genome).

Je m'appuierai sur un réseau collaboratif, mise en place depuis quelques années, qui comprend des équipes du LRI (Orsay), LORIA (Nancy), LIRMM (Montpellier), et de l'Université de Rennes, et sur des travaux déjà initiés notamment sur des génomes procaryotes (Aze, Gentils et al. 2008).

e) Conclusion

Le projet de recherche que je présente offre de nombreuses possibilités de sujets de thèse, permettant à de futurs doctorants, d'avoir la possibilité de s'investir à la fois dans le développement de méthodologies bioinformatiques innovantes, et dans l'application de ces méthodes sur des problématiques de recherche en biologie, génétique et amélioration végétale. L'interaction constante que nous mettons en place dans notre équipe entre biologistes, bioinformaticiens et informaticiens, offre un contexte multidisciplinaire stimulant et propice à ce type de recherche.

Nous pouvons aussi bénéficier de la dynamique créée par la mise en place de collaborations, dans le cadre de différents projets soumis récemment, comme l'IBC, Institut de Biologie Computationnelle de Montpellier, dans lequel je joue un rôle actif dans l'axe "Databases: Biological data and knowledge integration" qui intègre différentes équipes du LIRMM, INRIA Sophia Antipolis et

de l'IRD. Je peux citer aussi le projet plus large RENABI IFB (Infrastructure Française de Bioinformatique), où nous sommes un acteur majeur pour la mise en place d'une infrastructure en bioinformatique végétale au niveau national, notamment en coordination avec l'URGI. Le défi majeur que nous essayons de relever par la mise en place de ces programmes, étant de pouvoir extraire de la connaissance de la masse de données génomiques déjà existantes, ou en cours de production, et dont nous pouvons déjà anticiper l'augmentation considérable dans les années à venir.

Références citées

Argout, X., O. Fouet, et al. (2008). "Towards the understanding of the cocoa transcriptome: Production and analysis of an exhaustive dataset of ESTs of Theobroma cacao L. generated from various tissues and under various conditions." BMC Genomics 9: 512.

Argout, X., J. Salse, et al. (2010). "The genome of Theobroma cacao." Nat Genet.

Aze, J., L. Gentils, et al. (2008). "Towards a semi-automatic functional annotation tool based on decision-tree techniques." BMC Proc 2 Suppl 4: S3.

Bhagat, J., F. Tanoh, et al. (2010). "BioCatalogue: a universal catalogue of web services for the life sciences." Nucleic Acids Res 38(Web Server issue): W689-94.

Bizer, C. (2004). "D2RQ - treating non-RDF databases as virtual RDF graphs." IN PROCEEDINGS OF THE 3RD INTERNATIONAL SEMANTIC WEB CONFERENCE (ISWC2004).

Bocs, S. (2010). GNPAnnot Community Annotation System (CAS). Système d'annotation communautaire de génomes de plantes, d'insectes etde champignons. [Diaporama]. JOBIM 2010, Journée satellite "Annotations des génomes et génomique comparée", Montpellier, France.

Bruskiewich, R., G. Davenport, et al. (2006). "Generation Challenge Programme (GCP): standards for crop data." OMICS 10(2): 215-9.

Bruskiewich, R., M. Senger, et al. (2008). "The generation challenge programme platform: semantic standards and workbench for crop science." Int J Plant Genomics 2008: 369601.

Chen, H., L. Ding, et al. (2009). "Semantic web for integrated network analysis in biomedicine." Brief Bioinform 10(2): 177-92.

Dereeper, A., X. Argout, et al. (2007). "SAT, a flexible and optimized Web application for SSR marker development." BMC Bioinformatics 8: 465.

Dereeper, A., S. Nicolas, et al. (2011). "SNiPlay: a web-based tool for detection, management and analysis of SNPs. Application to grapevine diversity projects." BMC Bioinformatics 12(1): 134.

Droc, G., C. Périn, et al. (2009). "OryGenesDB 2008 update: Database interoperability for functional genomics of rice." Nucleic Acids Research 37(Database issue): D992–D995.

Droc, G., M. Ruiz, et al. (2006). "OryGenesDB: a database for rice reverse genetics." Nucleic Acids Res 34(Database issue): D736-40.

Dufayard, J. F. and M. Ruiz (2011). Phylogenetic analysis of mapped sequence reads. ISMB/ECCB 2011. Vienna, Austria.

Fleury, C., M.-F. Gautier, et al. (2011). Structure-based classification of the plant non-specific lipid transfer protein superfamily towards its functional characterization. ISMB/ECCB 2011. Vienna, Austria.

Flicek, P., M. R. Amode, et al. (2011). "Ensembl 2011." Nucleic Acids Res 39(Database issue): D800-6.

Gessler, D. D., G. S. Schiltz, et al. (2009). "SSWAP: A Simple Semantic Web Architecture and Protocol for semantic web services." BMC Bioinformatics 10: 309.

Giardine, B., C. Riemer, et al. (2005). "Galaxy: a platform for interactive large-scale genome analysis." Genome Res 15(10): 1451-5.

Holford, M. E., E. Khurana, et al. (2010). "Using semantic web rules to reason on an ontology of pseudogenes." Bioinformatics 26(12): i71-8.

Howe, D., M. Costanzo, et al. (2008). "Big data: The future of biocuration." Nature 455(7209): 47-50.

Huang, X., X. Wei, et al. (2010). "Genome-wide association studies of 14 agronomic traits in rice landraces." Nat Genet 42(11): 961-7.

Hull, D., K. Wolstencroft, et al. (2006). "Taverna: a tool for building and running workflows of services." Nucleic Acids Res 34(Web Server issue): W729-32.

Larmande, P., C. Gay, et al. (2008). "Oryza Tag Line, a phenotypic mutant database for the Genoplante rice insertion line library." Nucleic Acids Res 36(Database issue): D1022-7.

Lescot, M., P. Piffanelli, et al. (2008). "Insights into the Musa genome: syntenic relationships to rice and between Musa species." BMC Genomics 9: 58.
Magdelaine-Beuzelin, C., Q. Kaas, et al. (2007). "Structure-function relationships of the variable domains of monoclonal antibodies approved for cancer treatment." Crit Rev Oncol Hematol 64(3): 210-25.
Maziere, P., N. Parisey, et al. (2007). "Formal TCA cycle description based on elementary actions." J Biosci 32(1): 145-55.
Mochida, K. and K. Shinozaki (2010). "Genomics and bioinformatics resources for crop improvement." Plant Cell Physiol 51(4): 497-523.
Moreira, D. A. and M. A. Musen (2007). "OBO to OWL: a protege OWL tab to read/save OBO ontologies." Bioinformatics 23(14): 1868-70.
Mungall, C. J. and D. B. Emmert (2007). "A Chado case study: an ontology-based modular schema for representing genome-associated biological information." Bioinformatics 23(13): i337-46.
Peres, S., L. Felicori, et al. (2010). "Computing biological functions using BioPsi, a formal description of biological processes based on elementary bricks of actions." Bioinformatics 26(12): 1542-7.
Pugnere, D., M. Ruiz, et al. (2003). "The MetaFMF website: a high quality tool for meta-analysis of FMF." Nucleic Acids Res 31(1): 286-90.
Rahm, E. and P. A. Bernstein (2001). "A survey of approaches to automatic schema matching." The VLDB Journal 10(4): 334-350.
Rouard, M., V. Guignon, et al. (2010). "GreenPhylDB v2.0: comparative and functional genomics in plants." Nucleic Acids Res.
Rubin, D. L., N. H. Shah, et al. (2008). "Biomedical ontologies: a functional perspective." Brief Bioinform 9(1): 75-90.
Ruiz, M. (2009). Semantic standards for genomic analyses of the South and Mediterranean plants: the Generation Challenge Program use case. Proceedings of Biodiversity Information Standards, TDWG 2009 Annual Conference, Montpellier, France, 9-13 november 2009.
Ruiz, M., M. Rouard, et al. (2004). "TropGENE-DB, a multi-tropical crop information system." Nucleic Acids Res 32(Database issue): D364-7.
Saisho, D. and M. D. Purugganan (2007). "Molecular phylogeography of domesticated barley traces expansion of agriculture in the Old World." Genetics 177(3): 1765-76.
Sarrauste de Menthiere, C., S. Terriere, et al. (2003). "INFEVERS: the Registry for FMF and hereditary inflammatory disorders mutations." Nucleic Acids Res 31(1): 282-5.
Schneider, M., L. Lane, et al. (2009). "The UniProtKB/Swiss-Prot knowledgebase and its Plant Proteome Annotation Program." J Proteomics 72(3): 567-73.
Shadbolt, N., T. Berners-Lee, et al. (2006). "The Semantic Web Revisited." IEEE Intelligent Systems 21(3): 96-101.
Stein, L. D. (2008). "Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges." Nat Rev Genet 9(9): 678-88.
Trissl, S., K. Rother, et al. (2005). "Columba: an integrated database of proteins, structures, and annotations." BMC Bioinformatics 6: 81.
W3C (2004). "Resource Description Framework (RDF): Concepts and Abstract Syntax."
Wanchana, S., S. Thongjuea, et al. (2008). "The Generation Challenge Programme comparative plant stress-responsive gene catalogue." Nucleic Acids Res 36(Database issue): D943-6.
Wiederhold, G. (1992). "Mediators in the Architecture of Future Information Systems." Computer 25(3): 38-49.
Wilkinson, M. D., L. McCarthy, et al. (2010). "SADI, SHARE, and the in silico scientific method." BMC Bioinformatics 11 Suppl 12: S7.
Wilkinson, M. D., M. Senger, et al. (2008). "Interoperability with Moby 1.0--it's better than sharing your toothbrush!" Brief Bioinform 9(3): 220-31.

Wollbrett, J., P. Larmande, et al. (2009). Intégration automatique d'une ontologie de domaine dans un annuaire Biomoby. Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM) 2009, Nantes, France.

Wollbrett, J., P. Larmande, et al. (2011). Towards Automatic Generation of Semantic Web Services for Relational Biological Databases. RED Fourth International Workshop on REsource Discovery. Heraklion, Greece.