



Belgian Bioinformatics Conference

April 23, 2004

Brussels, Belgium

An EMBnet meeting sponsored under the European
Commission QLTR-2000-00472 EMBCORE
program

Contents

<i>Program</i>	3
<i>Opening and keynote talks</i>	5
<i>Oral contributions</i>	7
<i>Software presentations</i>	13
<i>Poster presentations</i>	17
<i>List of participants</i>	47

Program

8:30 Registration + poster setup

9:15 Welcome address by local organizer

9:20 Opening talk by Denis Thieffry (Université de Marseille)
Qualitative modelling, analysis and simulation of genetic regulatory networks

9:45 Contributed talks

chairman Gianluca Bontempi

1. Fred Oppendoes & J.P.Szikora - *In silico prediction of the glycosomal enzymes of Leishmania major*
2. Jan Wuyts et al. - *Comparative detection of microRNAs in the Arabidopsis and rice genomes*
3. Ari Loytynoja - *A hidden Markov model of evolution and structure for multiple sequence alignment*

11:00 Coffee break, software demonstrations and first poster session

11:45 Keynote presentation by Alfonso Valencia (Centro de Biotecnología, U. Autónoma de Madrid, Protein design group) Title to be communicated

12:30 Lunch and coffee

13:30 Second poster session and software presentation

14:30 Keynote presentation by Peter Ghazal (Scottish Centre for Genomic Technology & Informatics University of Edinburgh Medical School) *An integrative approach to post-genomics and Bioinformatics*

15:15 Contributed talks

Chairman Eric Depierreux

4. L.Duchateau et al. - *Estimating the variability of expression of a randomly inserted gene in rice plants by mixed model methodology*
5. Stein Aerts et al. - *Computational discovery of cis-regulatory modules in animal genomes*

Chairman Jacques van Helden

6. Klaas Vandepoele et al. - *Major events in the genome evolution of vertebrates*
7. Gert Thijs et al. - *Detection of cis-regulatory elements in sets of coregulated genes by phylogenetic footprinting*

17:30 Discussion and plans for next meeting

18:00 Closing of the conference

Opening talk and keynote presentations

Qualitative modelling, analysis and simulation of genetic regulatory networks

Denis Thieffry - LGPD-IBDM, Marseille, France

A proper understanding of the mechanisms controlling gene expression requires the integration of molecular and genetic data into full fledged mathematical models. As most available data are qualitative, we rely on a multi-level, logical approach, which enables a flexible dynamical modeling of complex regulatory networks. Our approach encompasses the development of a dedicated software suite (GIN-sim), and will be illustrated by applications to cell cycle, cell differentiation and pattern formation during *Drosophila* development.

Title to be communicated

Alfonso Valencia - Centro de Biotecnologia, U. Autonoma de Madrid, Protein design group

An integrative approach to post-genomics and Bioinformatics

Peter Ghazal - Scottish Centre for Genomic Technology & Informatics University of Edinburgh Medical School

New post-genomic technologies enable high throughput analysis of a huge spectrum of genetic content, gene expression and protein activity. Accordingly, biology has become a rapidly expanding domain of information science. In this context, the integration of the qualitative and quantitative genomic and proteomic measuring technologies with appropriate statistical and computational systems is a fundamental challenge. Professor Ghazal will discuss his collaborative efforts to evolve new biologic and informatic structures for accelerating biomarker discovery and a systems-level understanding of biology for medicine.

Oral contributions

1. In silico prediction of the glycosomal enzymes of Leishmania major

Fred Opperdoes and Jean-Pierre Szikora

Almost 8000 protein sequences available through the "Leishmania genome project" were used to predict the glycosomal proteome of *Leishmania major*. All protein sequences were analyzed for the presence of either a C-terminal (PTS1) or an N-terminal (PTS2) peroxisomal targeting sequence. 178 potential PTS1 proteins and 83 potential PTS2 proteins were identified. About 50% of them were hypothetical proteins for which no function was attributed. From those proteins with known function it appears that the glycosomes of *L. major* strongly resemble those of *T. brucei* with respect to their enzyme content. Glycosomes are not only involved in glycolysis, but also carry out reactions of the hexose-monophosphate pathway, purine salvage and pyrimidine biosynthesis, β -oxidation of fatty acids, fatty acid elongation and the biosynthesis of ether lipids. In addition they catalyze several reactions of the isoprenoid biosynthetic pathway and of protection against oxidant stress.

2. Comparative detection of microRNAs in the Arabidopsis and rice genomes

Jan Wuyts, Eric Bonnet, Pierre Rouz , Yves Van de Peer

MicroRNAs (miRNAs) are an extensive class of tiny RNA molecules that are thought to regulate the expression of target genes via complementary base-pair interactions with the mRNAs of these genes. Although the first miRNAs were discovered in the worm *Caenorhabditis elegans*, during the last years more than 200 miRNAs were discovered in diverse Eukaryotic organisms, most of the time by direct cloning methods. This approach is obviously biased in favor of the most abundant miRNAs. Different research groups have developed computational approaches to identify new miRNA genes in animals, using methods based on comparative genomics. Here, we present a genome-wide computational approach to detect new miRNA genes in the *Arabidopsis* genome. Our method is centered around the fact that miRNAs are conserved between different species. (We use *Oryza sativa* as second species) However, since these molecules are very small (20- 24nt.), this property alone is insufficient for their detection. To solve this problem, we take into account the characteristic properties of the secondary structure of the miRNA precursor molecules. Our method was explicitly designed to accommodate the variable length of plant miRNA precursors. This plant specific property makes it largely impossible to use miRNA detection programs developed for animal genomes on plant data. To further minimize the number of false positives we have implemented a number of filtering steps based on GC content, sequence complexity and minimum folding energy. The sensitivity of our procedure is demonstrated by the identification of 6 out of 8 miRNAs that were previously identified and experimentally validated as conserved between *Arabidopsis* and *Oryza*. In total we have identified a set of approximately 300 potential miRNA genes. In this set, 86 candidates have at least one nearly-perfect match with an *Arabidopsis* messenger RNA, thus constituting the potential target genes of those miRNAs.

3. A hidden Markov model of evolution and structure for multiple sequence alignment

Ari Loytynoja

The two most widely used heuristic alignment methods, progressive pairwise alignment algorithms and profile hidden Markov models (HMM), have different strengths: progressive methods take into account the hierarchical relationships among the sequences by using a tree to guide the alignment, whereas profile-based methods, based on linear models, allow for variable processes across (functionally different) sequence sites. I describe a novel multiple alignment algorithm that combines ideas from both methods, and simultaneously models (1) the evolution, and (2) the structure of multiple sequences related by a binary tree. The method is based on a HMM describing the match/gap-process as well as structural regions within sequences, progressive algorithms defining ancestral sequences from the pairwise alignments of their child nodes, and a probabilistic evolution model describing the character substitution processes in different structure states. The structure within sequences, e.g., a genes structure with introns and exons, a protein secondary structure with either buried and exposed sites, or alpha, beta and coil elements, or just slowly and fast evolving sequence regions, can be pre-defined based on external knowledge, or estimated during the alignment process. The method, being based on a stochastic model, also allows for defining the posterior probabilities of sequence sites being aligned and the process being in a given structure state at a given position, as well as the estimation of the model parameters from the data. Although this is still work in progress, I show that the algorithm works in practice: a model with 11 structural states (including two duration states, 3-state "codon" and 6-state "motif") can simultaneously (1) align genomic DNA sequences, (2) predict the gene structure (protein coding exons), and (3) find short, highly-conserved sequence regions. The algorithm is flexible and can easily be adjusted to more complex tasks.

4. Estimating the variability of expression of a randomly inserted gene in rice plants by mixed model methodology

L Duchateau (RUG), J De Wolf (CropDesign) and E Schrevens (KUL)

Rice plants of 5 related lines, obtained by insertion of the same gene on different locations on the genome, have been grown. Within a line, 2 types of sister plants are cultivated: those containing (I=Insertion) and those lacking (W=Wild type) the trans-gene. At 7 time points in their life cycle, the biomass of 9 replicate plants has been estimated using leaf area measurements through 2D imaging based on 6 pictures per plant and per time point. Randomization is done within lines, but the 5 lines were kept as separate blocks. The main interest of this particular analysis lays in predicting the effect of the 5 different insertions with respect to biomass increase as measured by leaf area and in assessing the variability between the 5 different lines. Both objectives can be dealt with in the context of the mixed model. Furthermore, this model accommodates for the repeated-measure structure of the data. An auto-regressive correlation structure between repeated measures was assumed for a first order linear model. The first model includes insertion as a random effect, resulting in an estimate of the variance component corresponding to the variability between the insertions and predictions of the effect of the different insertions by the best linear unbiased predictors (BLUP). This model assumes no interaction between type and line. In the second model both line and type within line are incorporated in the model as random

effects, in order to adjust for a possible interaction between type and line. If wild types of the different lines behave differently with respect to biomass increase, the second model is to be preferred. The heterogeneity between the wild types of the different lines can also be assessed by the mixed model and model choice is based on these results, together with the likelihood ratio test comparing the two models above.

5. Computational discovery of cis-regulatory modules in animal genomes

Stein Aerts, Bert Coessens, Yves Moreau, Gert Thijs, Peter Van Loo, and Bart De Moor

The transcriptional regulation of metazoan genes is governed by combinations of transcription factor binding sites in cis-regulatory modules. Their central role in gene regulatory networks makes their detection and characterization of great importance for the understanding of the genetic programs encoded in the genome. Here we present several methods and strategies for the detection of significant transcription factor binding sites (TFBS) and combinations thereof as modules, in animal genomes. A first strategy, that is implemented as a software tool called TOUCAN [1], allows for the retrieval of genomic sequences from Ensembl, the detection of conserved non-coding sequences (CNS) between orthologous sequences, the scoring of the CNSs with position weight matrices from TRANSFAC within a probabilistic framework, and the detection of over-represented motif instances using the binomial formula. The methods are accessed remotely via SOAP web services. Next, we move one step further and search for the optimal combination of TFBSs using both an A* [2] and Genetic Algorithm [3] based search algorithm called ModuleSearcher. The newly found putative modules are then used to scan all human CNSs to find putative target genes in the genome. By measuring the functional coherence of the top scoring target genes using Gene Ontology associations, the modules are validated. Finally, we present a new and unpublished integrated genomics approach to validate and prioritise the putative target genes using microarray data, Gene Ontology associations, textual data from LocusLink and Medline abstracts, KEGG pathway membership, EST-based expression data, and InterPro protein domains. The prioritisation is based on order statistics and can also be applied to the prioritisation of candidate disease genes.

[1] Stein Aerts, Gert Thijs, Bert Coessens, Mik Staes, Yves Moreau and Bart De Moor (2003) TOUCAN: deciphering the cis-regulatory logic of coregulated genes *Nucleic Acids Research*, 31(6), 1753-1764.

[2] Stein Aerts, Peter Van Loo, Gert Thijs, Yves Moreau and Bart De Moor (2003) Computational detection of cis-regulatory modules *Bioinformatics*, 19 Suppl. 2, ii5-ii14. [3] Stein Aerts, Peter Van Loo, Yves Moreau and Bart De Moor (2004) A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes. *Bioinformatics*, in press.

6. Major events in the genome evolution of vertebrates: paranome age and size differs considerably between ray-finned fishes and land vertebrates

Klaas Vandepoele, Wouter De Vos, John S. Taylor, Axel Meyer and Yves Van de Peer

It has been suggested that fish have more genes than humans. Whether most of these additional genes originated through a complete (fish-specific) genome duplication or through many lineage specific tandem gene or smaller block duplications and family expansions continues to be debated. We analyzed the complete genome of the pufferfish *Takifugu rubripes* (Fugu) and compared it to the paranome of humans. We show that most paralogous genes of Fugu are the result of three complete genome duplications. Both relative and absolute dating of the complete predicted set of protein-coding genes suggest that initial genome duplications, estimated to have occurred at least 600 million years ago, shaped the genome of all vertebrates. In addition, analysis of more than 150 block duplications in the Fugu genome clearly supports a fish-specific genome duplication (about 320 million years ago) that coincided with the vast radiation of most modern ray-finned fishes. Unlike the human genome, Fugu contains very few recently duplicated genes; hence, many human genes are much younger than fish genes. This lack of recent gene duplication or, alternatively, the accelerated rate of gene loss, is possibly one reason for the drastic reduction of the genome size of Fugu observed during the last 100 million years or so, subsequent to the additional genome duplication that ray-finned fishes, but not land vertebrates, experienced.

7. Detection of cis-regulatory elements in sets of coregulated genes by phylogenetic footprinting

Gert Thijs, Pieter Monsieur, Kathleen Marchal, Bart De Moor

Given the increasing availability of newly sequenced genomes, cross-species comparison becomes a more important aspect of bioinformatics research. Phylogenetic footprinting is the methodology for the discovery of transcription factor binding sites in a set of orthologous regulatory regions from multiple species. The basic idea is that selective pressure causes functional elements to evolve at a slower rate than nonfunctional sequences. This means that unusually well conserved sites among a set of orthologous regulatory regions might be functional regulatory elements. Applying motif search algorithms to such a set of orthologous promoter sequences should reveal these conserved patterns. To further strengthen this methodology phylogenetic footprinting can be combined with the motif finding in sets of co-regulated genes. The goal is then to find common motifs that are not only conserved among the set of co-regulated genes but also among the related species. Here we present a two-step approach comparable to the method of Wang and Stormo [1]. The input is a set of coregulated genes for which the corresponding orthologs are identified. In the first step of the analysis, larger conserved blocks will be detected among the orthologous genes with BlockSampler, an adapted version of our Gibbs sampling method for motif finding MotifSampler [2,3]. The most important extensions to BlockSampler are the definition of a seed or reference sequence, that is always included when building the conserved block, the usage of a species-specific background model for each individual sequence, and a method to extend motifs based on the level of conservation. In the second step, a local alignment algorithm is used to

find the largest motif common to two conserved blocks. As a distance measure we use the Kullback-Leiber distance between two distributions and we have derived a method to easily estimate the statistical significance of the retrieved motifs. After aligning all blocks from the set of co-regulated genes to each other, clustering is applied to find the common motifs. As a test example we have chosen two regulons, *pmrA* and *phoP*, and also a set of random genes from *Salmonella thypimurium* [4]. Orthologous genes were selected from several *Escherichia coli* and *Yersinia pestis* strains. First results show that we are able to detect the common motifs in the two regulons with this strategy and distinguish them from the random genes.

[1] Wang T., Stormo G., 2003. *Bioinformatics*, 19(18), 2369-2380.

[2] Thijs G. et al., 2001. *Bioinformatics*, 17(12), 1113-1122.

[3] Thijs G. et al., 2002. *J.Comp.Biol.* 9(2), 447-464. [4] Marchal K., et al., 2004. *Genome Biology*, 5(2), R9.

Software presentations¹

The Metapopulation Genetic Algorithm: an efficient Solution for the problem of large phylogeny estimation

Alan R. Lemmon & Michel C. Milinkovitch

Large phylogeny estimation is a combinatorial optimization problem that no future computer will ever be able to solve exactly in practical computing time. The difficulty of the problem is amplified by the need to use complex evolutionary models and large taxon samplings. Hence, many heuristic approaches have been developed, with varying degrees of success. Here, we report on a novel heuristic approach, the “metapopulation genetic algorithm” (metaGA), involving several populations of trees that are forced to cooperate in the search for the optimal tree. Within each population, trees are subjected to evaluation, selection, and mutation events, which are directed using inter-population consensus information. The method proves to be both very accurate and vastly faster than existing heuristics, such that data sets comprised of hundreds of taxa can be analyzed in practical computing times under complex models of maximum likelihood evolution. Branch support values produced by the metaGA might closely approximate the posterior probabilities of the corresponding branches. A demonstration of the version 1 of the software will be given.

The Belgian Biodiversity Information Facility

Wautelet F., Duflost J., Mergen P., Herzog R.

BeBIF (<http://www.be.gbif.net>) is the Belgian National Participant Node of the worldwide biodiversity information network Global Biodiversity Information Facility (GBIF, <http://www.gbif.org/>). GBIF's main goal is to integrate and make worldwide biodiversity related data freely available to all. BeBIF fulfils the role and the tasks assigned to a national node by GBIF, is member of the GBIF NODES Committee, the GBIF Data Access and Data Interoperability sub-committee (DADI) and the Multilingual Working group. BeBIF plays an active role in related international organisations and projects like the European Network for Biodiversity Information (ENBI) and the Taxonomical Data Working Group (TDWG)

At the national level, there is an endeavour to build a bioinformatics infrastructure to integrate "Belgian" biodiversity resources. We play the role of national data viewing unit and gateway of biodiversity data within the Belgian Federal Science Policy Office initiative "Biodiversity.be". The central portal is combining interrogations of distributed and centralized databases. BeBIF plays the role of a Information Technology Center at national level having its IT specialist sharing their expertise with the other partners and recommend technical solutions according to their needs as well as guarantee the necessary interoperability between national initiatives to expose and exchange Belgian Biological data according to international standards (DarwinCore, ABCD) using IT tools (DIGIR, BioCASE data providers) developed in collaboration with GBIF and TDWG.

¹ Abstracts of software presentations are reproduced twice for clarity : in this section and in the Poster presentations section

The main goal of BeBIF is that metadata and data listed in the definition of a node according to the Memorandum of Understanding (MOU) of GBIF can be found and provided in the appropriate standards to expose “Belgian” biodiversity information to its end-user and transmit them in the appropriate standard to the GBIF central portal in a distributed network. BeBIF goes also further in its investments by developing software and data validation tools in accordance with our user needs. Our mission is also to encourage and assist Belgian data nodes to share data with GBIF under common standards. We can also play the role of IT advisors and to some extent fulfill helpdesk activities.

BEN –The National Biocomputing Support for Belgian Scientific Researchers in Molecular Biology, Biotechnology and Biomedical Sciences, since 1993.

*Guy Bottu, David Coornaert, Valérie Ledent, Marc Colet & Robert Herzog
ULB, Campus de la Plaine, Bat. NO-CP 257, Boulevard du Triomphe, 1050
Bruxelles, Belgium*

BEN - the Belgian EMBnet node provides to scientific researchers in Belgium, both academic and industrial, an access to permanently updated databanks of nucleic acid and protein sequences, as well as the necessary software to search these databanks and to perform various sequence analysis tasks. BEN maintains a SRS server offering access to more than 40 databanks.

BEN sequence analysis services are centred on the freeware package EMBOSS as well as on other useful software such as BLAST, Clustal, HMMER, Phylip, Primer3.... These different tools have been integrated in order to make their use easier. BEN ensures a technical support service that users can call in case of problems and provides users with freeware of interest for the molecular biologist through its anonymous ftp server.

BEN has developed several graphical interfaces for various software (GCG, Blast ...). Recently, a new wEMBOSS interface is born from a coordinate effort from Prof. Marc Colet from BEN and Martin Sarachu from the Argentinian EMBnet Node. wEMBOSS allows web-users to access their personal data files and to save their sequence analysis results on the remote computer

BEN is a joint initiative of the ULB and the VUB. It is financially supported by the Belgian Science Policy. The computers of BEN are located at the VUB/ULB Computing Centre. The functioning of BEN is largely depending on the existence of the Belgian national research network, BELNET. BEN is member of the EMBnet (European Molecular Biology Network) established in 1988 to link European laboratories distributing bioinformatics services to their national and/or regional scientific community.

Web site : <http://www.be.embnet.org>

Ftp site : <ftp://ftp.be.embnet.org>

A web site for path finding in metabolic pathways

Didier Croes, Fabian Couche, Jacques van Helden and Shoshana Wodak

We developed a web tool (<http://www.scmbb.ulb.ac.be/~didier/pathfinding/>) which runs a path finding algorithm in the whole metabolic network. We tool takes as input two compounds, reactions, or EC numbers, and calculates the k shortest paths between these seed nodes. The result is presented in both textual and graphical form. During the demonstration, we will show the results obtained with typical examples, and highlight the crucial importance of the choice of parameters. In particular, we will compare the results effect of excluding pool metabolites from the complete network, and the improvements obtained by assigning a specific weight to each compound. The validation of the tool will be presented in an accompanying poster (Validation of pathways inferred by path finding in metabolic networks, Croes et al.).

A Snow/aMAZE demo.

Olivier Sand, Christian Lemer, Frédéric Fays, Erick Antezana, Fabian Couche, Simon De Keyzer, Olivier Hubaut, Jesintha Mary Maniraja, Hassan Anerhour, Xavier Santolaria, Jean Richelle, Shoshana Wodak

We will present a demonstration of Snow, a user-friendly interface for querying and browsing databases of networks (see the poster "The Snow system, a tool for representation and analysis of networks"). Snow supports iql, a simple language which enables users to perform complex queries knowing only the database conceptual model. Results can be viewed together or separately with a by-default or chosen set of their attributes. The set can be expanded to all attributes later if necessary. Some attributes are themselves expandable, allowing to navigate through the whole content of the database. Results can also be displayed as graphs generated on the fly, showing connections between related entities. For this demonstration, Snow will be interfacing aMAZE, our database for molecular interactions and cellular processes (see the poster on aMAZE). We will execute simple and complex interrogations of aMAZE running live on the remotely connected database server.

Poster presentations

1. The Metapopulation Genetic Algorithm: an efficient Solution for the problem of large phylogeny estimation

Alan R. Lemmon & Michel C. Milinkovitch

Large phylogeny estimation is a combinatorial optimization problem that no future computer will ever be able to solve exactly in practical computing time. The difficulty of the problem is amplified by the need to use complex evolutionary models and large taxon samplings. Hence, many heuristic approaches have been developed, with varying degrees of success. Here, we report on a novel heuristic approach, the “metapopulation genetic algorithm” (metaGA), involving several populations of trees that are forced to cooperate in the search for the optimal tree. Within each population, trees are subjected to evaluation, selection, and mutation events, which are directed using inter-population consensus information. The method proves to be both very accurate and vastly faster than existing heuristics, such that data sets comprised of hundreds of taxa can be analyzed in practical computing times under complex models of maximum likelihood evolution. Branch support values produced by the metaGA might closely approximate the posterior probabilities of the corresponding branches. A demonstration of the version 1 of the software will be given.

2. Protein-DNA Interactions: Exploring the Nonadditive Effects in Stair motifs involving H-bond, Cation-pi and Stacking Interactions

Biot, C., Wintjens, R. and Rooman M.

At the interface between protein and double-stranded DNA, stair motifs simultaneously involve three different types of pairwise interactions: aromatic base stacking, hydrogen bonding and cation-pi. The relative importance of these interactions is studied in the stair motif occurring in the 1Tc3crystal structure, which involves an arginine and two stacked guanines, by means of Hartree Fock (HF) and Møller-Plesset energy and free energy calculations, including vibrational, rotational, translational contributions, both in vacuum and various solvents. The results obtained show an anti-cooperative tendency of the HF energy and vibrational free energy terms, and the cooperativity of the rotational, translational and solvation free energies. Hence, the cooperativity of the stair motif interactions, in the context of protein-DNA recognition, can be viewed as arising from the environment.

3. Web based automated arrayCGH analysis: 'arrayCGHbase'

Björn Menten, Filip Pattyn, Geert Mortier, Anne De Paepe, Frank Speleman, Stefan Vermeulen, Jo Vandesompele

Although the principle of array comparative genomic hybridization (arrayCGH) has already been described for more than 5 years, its application for genome wide analysis of gene copy number quantification has only recently come of age. One of the challenges of this new technology is the handling of 1000 s of data points generated in each individual assay. We developed a comprehensive and MIAME (Minimal Information About Microarray Experiment) compliant MySQL-based database and data mining webtool in order to store, compare, analyze, interpret and graphically display arrayCGH results in a uniform and user friendly format. Important features of arrayCGHbase are the chromosome based visualization of arrayCGH results, enabling rapid visualization of the genomic location and content of each spotted BAC, cDNA or oligonucleotide on the array. ArrayCGHbase can be used either online (<http://medgen.ugent/be/arraycghbase/>), or can be installed on a local server running the free MySQL database and PHP scripting language. This database will allow investigators to interpret and compare large clinical and research arrayCGH datasets and incorporates the essential links to investigate genomic regions of interest.

4. Building genomic profiles for uncovering segmental homology in the twilight zone

Cedric Simillion, Klaas Vandepoele, Yvan Saeys and Yves Van de Peer

The identification of homologous regions within and between genomes is an essential prerequisite for studying genome structure and evolution. Different methods already exist that allow detecting homologous regions in an automated manner. Those methods are either based on finding sequence similarities at the DNA level or on identifying chromosomal regions showing conservation of gene order and content. Especially the latter approach has proven useful for detecting homology between highly divergent chromosomal regions. However, until now, such map-based approaches required that candidate homologous regions show significant colinearity with other segments in order to be considered as being homologous. Here, we present a novel method that creates profiles combining the gene order and content information of multiple mutually homologous genomic segments. These profiles can be used to scan one or more genomes to detect segments that show significant colinearity with the entire profile but not necessarily within individual segments. When applying this new method to the combined genomes of Arabidopsis and rice, we find additional evidence for ancient duplication events in the rice genome.

5. The Snow system, a tool for representation and analysis of networks

Christian Lemer, Erick Antezana, Fabian Couche, Simon De Keyzer, Frédéric Fays, Olivier Hubaut, Jean Richelle, Shoshana Wodak

The Snow system is mainly constituted of a workbench (Snow) and a database management system (Igloo). It has been developed to represent and analyze networks in the context of the aMAZE project for the representation and the analysis of biological processes. Snow adopts the Generalized Entity-Relationship (GER) conceptual data modeling approach. It allows the designer to work at a level of concepts very close to the domain expert's view. In the context of the amaze project, it gives the biologists an efficient tool to model the rich set of intricate biological networks. The data modeling is done using DB-MAIN, a computer aided software engineering tool (CASE tool) developed at the Laboratoire d'Ingénierie des Bases de Données (LIBD/FUNDP). For efficiency the data are stored in a relational database (PostgreSQL) and accessed through a database layer for user-friendliness. Igloo, as database layer, is in charge of the queries, the edition (creation and modification) of the entities stored in the database, and in that way of the loading and annotation of the data. It provides a GER application programming interface (API) and hides the underlying data storage organization. The database management system Igloo is available as a Java library. The Snow workbench is the main user interface. It provides interactive tools to interrogate the database, navigate through the web of relations between entities, analyze and visualize the data. It has been built on top of Eclipse, a Rich Client Platform. Igloo has been packaged as an Eclipse plug-in and a companion plug-in, Snow, provides the user interface: query launcher, object browser, diagram editor. Beside Snow as the main user interface, other applications may access the database using the Igloo Java library. Even if it has been developed for the aMAZE project, the Snow system is domain independent: as the data model is not hard coded in the application and is obtained from an external repository, Snow can be used for many other application domains.

6. The aMAZE database goes public

Christian Lemer, Hassan Anherour, Jesintha Mary Maniraja, Olivier Sand, Jean Richelle, Shoshana Wodak

The goal of the aMAZE project is to develop a workbench for the representation and analysis of networks of molecular interactions and cellular processes, genetic expression and regulation, enzymatic transformations and regulation, metabolic pathways, signal transduction, etc... The project has reached its first milestone. A first version of the data model has been finalized. It contains all the basic entities used in the description of biological processes: Genes, Polypeptides, Compounds. The elementary transformations involved in metabolic processes and regulation are also described: Reaction and Reaction Catalysis, Gene Expression and Transcriptional Regulation. Last but not least, the description of biological processes has been defined: Process, Process Step and Process Intermediate. This model has been implemented using the Snow system, for analysis and representation of networks (see other poster). The database has been populated from external sources (Kegg, Swissprot, Genbank,...) for the basic entities. Elementary transformations also come from external sources (Kegg, Swissprot) but with some additional processing required

by the semantic differences between our data model and the one from the source. The metabolic processes come exclusively from in-house annotation. Both the software and the database access are publicly available. The status of both the data model and the database content will be presented and discussed.

7. Statistical mean-force potentials dependent on protein size

Dehouck Yves, Gilis Dimitri, Rooman Marianne

Knowledge-based potentials are widely used in simulations of protein folding, structure prediction and protein design. Their advantages include limited computational requirements and the ability to deal with low-resolution protein models compatible with long-scale simulations. Their drawbacks comprehend their dependence on specific features of the dataset from which they are derived such as the size of the proteins it contains, and their physical meaning is still subject of debate. We address these issues by probing the theoretical validity of these potentials as mean force potentials that take the solvent implicitly into account and involve entropic contributions due to atomic degrees of freedom and solvation. The dependence on the size of the system is checked distance-dependent amino acid pair potentials, derived from six protein structure sets containing proteins of increasing length N . For large inter-residue distances, they are found to display the theoretically predicted $1/N$ behavior weighted by a factor depending on the boundaries and the compressibility of the system. For short distances, different trends are observed according to the nature of the residue pairs and their ability to form, for example, electrostatic, cation- $\{\pi\}$ or $\{\pi\}$ - $\{\pi\}$ interactions, or hydrophobic packing. The results of this analysis are used to devise a novel, protein size-dependent, distance potential, which displays an improved performance in discriminating native sequence-structure matches among decoy models.

8. Validation of pathways inferred by path finding in metabolic networks

Didier Croes, Fabian Couche, Jacques van Helden and Shoshana Wodak

Biochemical databases (such as LIGAND, ENZYME, ...) contain information on thousands of reactions. Graph-based representations of the metabolic network have been used to describe global properties of the metabolic network, such as connectivity, small world properties, modularity (Jeong et al., 2000; Fell and Wagner, 2000; Wagner and Fell, 2001; Ravasz et al., 2002; van Helden et al., 2002). Path finding has also been used to measure a metabolic distance between pairs of enzymes, and to compare this metabolic distance with structural and genomic features (Rison et al., 2002; Simeonidis et al., 2003). Surprisingly, none of these papers attempted to compare the paths found with the metabolic pathways previously characterized by biochemists. In this paper, we performed an evaluation of the path finding results by comparing, for each pathway from the aMAZE database (Lemer et al., 2004), the shortest paths between the first and the last reaction, and the annotated pathway. This evaluation clearly shows that the paths found in the complete metabolic graph have not much to do with biochemical pathways: the inferred paths generally borrow shortcuts through pool metabolites such as H_2O , ATP, NADH, ... When these trivial compounds are discarded from the metabolic graph, the results are more relevant, but

the inference is correct only for very short pathways (1 to 3 intermediate reactions). A drastic improvement is obtained by assigning a specific weight to each compound, on the basis of its connectivity.

References

- Fell DA, Wagner A. The small world of metabolism. *Nat Biotechnol.* 2000 Nov;18(11):1121-2.
- Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. The large-scale organization of metabolic networks. *Nature.* 2000 Oct 5;407(6804):651-4.
- Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL. Hierarchical organization of modularity in metabolic networks. *Science.* 2002 Aug 30;297(5586):1551-5.
- Rison SC, Teichmann SA, Thornton JM. Homology, pathway distance and chromosomal localization of the small molecule metabolism enzymes in *Escherichia coli*. *J Mol Biol.* 2002 May;318(3):911-32.
- Simeonidis E, Rison SC, Thornton JM, Bogle ID, Papageorgiou LG. Analysis of metabolic networks using a pathway distance metric through linear programming. *Metab Eng.* 2003 Jul;5(3):211-9.
- van Helden J, Wernisch L, Gilbert D, Wodak SJ. Graph-based analysis of metabolic networks. *Ernst Schering Res Found Workshop.* 2002;(38):245-74.
- Wagner A, Fell DA. The small world inside large metabolic networks. *Proc R Soc Lond B Biol Sci.* 2001 Sep 7;268(1478):1803-10.
- Lemer C, Antezana E, Couche F, Fays F, Santolaria X, Janky R, Deville Y, Richelle J, Wodak SJ. The aMAZE LightBench: a web interface to a relational database of cellular processes. *Nucleic Acids Res.* 2004 Jan 1;32 Database issue:D443-8.
- van Helden J, Naim A, Lemer C, Mancuso R, Eldridge M, Wodak SJ. From molecular activities and processes to biological function. *Brief Bioinform.* 2001 Mar;2(1):81-93.
- van Helden J, Naim A, Mancuso R, Eldridge M, Wernisch L, Gilbert D, Wodak SJ. Representing and analysing molecular and cellular function using the computer. *Biol Chem.* 2000 Sep-Oct;381(9-10):921-35.

9. A web site for path finding in metabolic pathways

Didier Croes, Fabian Couche, Jacques van Helden and Shoshana Wodak

We developed a web tool (<http://www.scmbb.ulb.ac.be/~didier/pathfinding/>) which runs a path finding algorithm in the whole metabolic network. We tool takes as input two compounds, reactions, or EC numbers, and calculates the k shortest paths between these seed nodes. The result is presented in both textual and graphical form. During the demonstration, we will show the results obtained with typical examples, and highlight the crucial importance of the choice of parameters. In particular, we will compare the results effect of excluding pool metabolites from the complete network, and the improvements obtained by assigning a specific weight to each compound. The validation of the tool will be presented in an accompanying poster (Validation of pathways inferred by path finding in metabolic networks, Croes et al.).

10. Gene duplication and biased functional retention of paralogs in bacterial genomes

Dirk Gevers, Klaas Vandepoele, Cedric Simillion and Yves Van de Peer

Microbial genomes have a considerable fraction of genes that are homologous to other genes within the same genome. Gene duplication and retention has been observed as an evolutionary response in bacteria exposed to different selection pressures, such as for example starvation conditions and thermal stress. When the selective pressure is removed the duplicates can be rapidly lost, thereby forming a reversible adaptive mutation that alters gene dosage without really altering genetic information. In addition to this short-term evolutionary advantage of paralogs, gene duplication and consequent functional divergence is considered an important evolutionary step towards diversity in the functional repertoire of an organism, presumably enabling the

organism to adapt to varying environmental conditions and broadening the phenotypes. Detailed analysis of the duplication events that have occurred is essential for the understanding of the evolutionary dynamics of prokaryotic genomes. Most of the duplicated genes in Bacteria seem to have been created by small gene duplication events. Evidence for large-scale gene duplications such as those observed in eukaryotic genomes could not be detected in any of the 106 bacterial genomes investigated. Nevertheless, paralogous genes comprise a significant fraction of the bacterial genome coding capacity. Interestingly, there seems to be a clear correlation between the number of retained duplicates and the functional class to which they belong. Indeed, particularly genes involved in the adaptation to a constantly changing environment seem to have been preserved, which again shows the importance of gene duplication for biological evolution. The majority of the duplicated genes occur dispersed over the genome, whereas the rest was found as potential retained operon duplications, i.e. blocks of 3 to 4 genes. The question remains whether the block duplications that we identified are under a strong selective pressure preserving gene co-expression and co-regulation in duplicate, or, alternatively, whether these represent recently duplicated gene strings as yet spared from rearrangements and disruption.

11. In silico analysis of gene co-regulation: exploiting whole-genome hyper-geometric word probabilities in a combined enumerative-alignment method

Dominique Vlieghe, Pieter J De Bleser, Frans Van Roy

The methods for the detection of consensus patterns common to the promoter sequences of co-expressed genes can be divided into enumerative and alignment methods. Enumerative methods count exhaustively all possible words and yield significant results because of the background model. An important limitation, however, is the reliance on the detection of non-degenerated words resulting in the failure of detection of motifs that deviate from the consensus. Alignment methods rely on the identification by significant local alignment but the statistical approaches are sensitive to noise and local conditions. Nevertheless, their representation of the profile, in the form of a weight matrix, deals better with the variability naturally observed with regulatory elements in promoter sequences. We present a new, combined approach that uses three selection steps and eliminates most of the limitations of the individual methods. First, a set of promising putative words are obtained using a hyper-geometric distribution approach, wherein the (over-)representation of words in the set of co-regulated promoters is compared to their presence in the full genome promoter set. Next, this restricted set of candidate motifs is used to determine profiles by optimizing its information content. Finally, the best profiles are used to extract additional motifs from the unaligned promoter sequences. Our combined approach detects consensus sequences of experimentally verified transcription factor binding sites from published data set of *Saccharomyces cerevisiae*, even if the component words deviate from the consensus.

12. RMAGEML: integrating MAGE-ML format microarray data and Bioconductor

Durinck Steffen, Joke Allemeersch, Vincent J. Carey, Yves Moreau, Bart De Moor

The microarray gene expression markup language (MAGE-ML) is a widely used XML standard for describing and exchanging information about microarray

experiments. It can describe microarray designs, microarray experiments designs, gene expression data and data analysis results. Bioconductor is an open source project that provides a framework for the statistical analysis of genomic data in R. These R packages provide a wide range of microarray data analysis tools. Up till now it was not possible to import data stored in MAGE-ML format in Bioconductor. Because of the importance of both MAGE-ML and Bioconductor in the field of microarray data analysis, this acute gap had to be filled. We describe RMAGEML, a new Bioconductor package that provides a link between MAGE-ML format microarray data and Bioconductor. The current version enables MAGEML-import to the *limma* and *marray* Bioconductor packages. RMAGEML is available at <http://www.bioconductor.org>

13. Planet, a network of european plant databases

Eric Bonnet, Stephane Rombauts & Pierre Rouzé

The future development of agricultural and environmental research relies strongly on plant gene data. The compilation of information resources requires dynamic information acquisition, expert curation and the integration of bioinformatic methods. PlaNet aims to overcome the limitations of individual efforts as well as the limitations of heterogeneous, independent data collections. PlaNet is a distributed effort among bioinformatics groups and plant molecular biologists to establish a comprehensive integrated database in a collaborative network. PlaNet creates a nucleus for other European and International groups and consortia to join and utilise the network.

Objectives

- To create a network of dynamically interconnected European plant databases
- Development of new methods for data exchange, database integration and access
- High quality integrated data resources for research
- High availability of data generated by European laboratories and plant research consortia (data platform)
- Focussing direct contribution by regional plant research communities (expert annotation system)
- Systematic classification of plant genes and regulators
- Development of standards for data representation and nomenclature Partners
 - o JIC (John Innes Centre) Norwich, United Kingdom
 - o NASC (Nottingham Arabidopsis Stock Centre) Nottingham, United Kingdom
 - o CNB/CSIC (Centro Nacional de Biotecnologia) Madrid, Spain
 - o VIB (Flanders Interuniversity Institute for Biotechnology) Gent, Belgium
 - o PRI (Plant Research International) Wageningen, The Netherlands
 - o MIPS (Munich Information Centre for Protein Sequences) Neuberberg, Germany Technology
- Multi-layer architecture, XML The partners implement local data collections within their special fields of interest or expertise and provide access to their databases via web services
- XML, SOAP, WSDL The connection between the individual resources is realized with BioMoby (biomoby.org). BioMoby provides an architecture for the discovery and distribution of biological data through web services

- Ontologies and data models Nomenclature and data representation are standardized using ontologies and generic data models. Standards are defined in interaction with the relevant communities
- Annotation tools to keep databases curated and allow expert annotation, local and remote annotation interfaces are created. They provide easy and direct access to all data.
- Data integration tools. External data is gathered into PlaNet using integration tools that allow flexible migration from various representations. Consistency of data is realized by 8 / 30 data synchronization between the individual resources.

14. Computational approach of the fertilization calcium waves

Geneviève Dupont & Rémi Dumollard

Fertilization triggers repetitive waves of cytosolic Ca^{2+} in the egg of many species. The mechanism involved in the elevation of intracellular Ca^{2+} has been studied in much detail in mature ascidian eggs, by submitting those to artificial stimulations by InsP3 or its poorly- metabolizable analog, gPIP2 (Dumollard and Sardet, 2001). Here, we use these experimental results to develop a realistic theoretical model for Ca^{2+} oscillations in mature eggs. The model takes into account the spatial heterogeneity of the endoplasmic reticulum (ER) distribution. Our results corroborate the hypothesis that Ca^{2+} wave pacemakers are associated with accumulations of cortical ER. The validity of the model is moreover confirmed by the adequacy of its theoretical predictions as to the effect of localized injections of massive amounts of InsP3 analogs. In a second step, we use the model to make some propositions about the possible characteristics of the sperm factor (SF). We found that to account for the spatial characteristics of the first series of Ca^{2+} oscillations seen at fertilization in ascidian eggs, it has to be assumed that the SF is a Ca^{2+} -sensitive, highly diffusible phospholipase C. Although the actual state of knowledge does not allow us to explain the relocalization of the initiation site of the successive waves observed experimentally, the model corroborates the assumption that PIP2, the substrate for PLC is distributed all over the egg. We also predict that the dose of SF injected into the egg might modulate the temporal shape of the first, long-lasting fertilization wave.

15. AssocioportDB, a Database System for protein-protein interactions of membrane protein

Richard Kamuzinzi, Valérie Ledent, Robert Herzog, Petr Obrdlík, Heinz Ellerbrock, Jose L. Revuelta, Eckhard Boles, Bruno André and Wolf B. Frommer

The EU Associoport project is dedicated to the systematic identification of protein-protein interactions in the specific case of membrane proteins. To achieve this, a method based on the split-ubiquitin two-hybrid system is applied on a large scale. Associoport will exploit this system in exhaustive searches for membrane protein interactions. This will be done in two model organisms for which the complete genome sequence is available: the yeast *Saccharomyces cerevisiae* and the plant *Arabidopsis thaliana*.

In the Associoport project, a major contribution of Bioinformatics laboratory of the ULB is the analysis of user's requirements and the design of a database system whose purpose is to harbor the large amount of experimental data produced by the different partners of the project. The database will centralize and provide the necessary information that must be shared among the partners of the project.

The current implementation of the Associoport Database System follows an N-Tier architecture built with Java technologies. The design of the web application part is ruled by the MVC design pattern that will be, in this project context, implemented by the Struts framework. Struts is an open source framework that is most widely adopted for building web applications based on JSP/Servlet technologies.

16. Prophage detection in bacterial genomes

Gipsi Lima, Raphael Leplae, Shoshana Wodak and Ariane Toussaint

Sequencing of bacterial genomes confirmed the crucial role that bacteriophages play in bacterial evolution and the divergence between closely related bacterial strains and species. Temperate phages, while residing in their host as latent prophages, either integrated in the host genome or as a circular or linear plasmid, become a repository for the shuffling of genetic information by recombination with temporary resident phage, plasmid or other mobile element. ACLAME is a database dedicated to a reticulate classification of prokaryotic Mobile Genetic Elements including phages and prophages (Leplae et al. 2004).

To populate the database with additional phage proteins, we developed a system to identify prophages in sequenced bacterial genomes. The procedure includes:

- The detection of integrases in the bacterial genome based on sequence similarity with the integrases classified in the ACLAME database; indeed prophage integrase genes usually define one prophage end
- Similarity evaluation of 100 contiguous protein sequences on both sides of the integrase with phage proteins stored in ACLAME; definition of putative prophages
- Search for a direct repeat of at least 10bp between a 250 kb stretch including the putative prophage and a 300 bp DNA fragment next to the integrase, that does not

include prophage genes. This repeat, generated upon integration of the viral into the host genome, assigns the prophage boundaries . Applied on 145 bacterial genomes, the strategy identified ~700 putative prophages in 102 genomes. These results have been compared to 320 complete and defective prophages manually assigned (Casjens 2003) and residing in 123 bacterial genomes. Hundred and eighty of those have been detected by our system . Most of the prophages our system missed were filtered out in the comparison with phage proteins stored in ACLAME.

17. Ambiguity Analysis in Large-Scale Protein Identification

Grégoire R. Thomas, Kris Gevaert, Lennaert Martens, Joel Vandekerckhove

High throughput protein identification techniques based on tandem-mass spectrometry allow the identification of highly complex protein mixtures. The accuracy of protein identification greatly depends upon reliable identification of peptide sequences from spectra. We present here Pixit, a software tool that analyses and reports protein identification from MS/MS-spectra using results of probability-based scoring algorithms such as Mascot. By processing the various types of ambiguities that occur following identification, Pixit optimises the use of scoring results and increases the accuracy of identification.

18. Fast fitting of atomic structures to low resolution electron density maps by surface overlap maximization

Hugo Ceulemans and Robert B. Russell

Structural genomics and modelling offer the prospect of a comprehensive knowledge of the structure of individual proteins at atomic resolution. However, extending this knowledge to (larger) protein complexes by X-ray crystallography or NMR spectroscopy has proven difficult due to technical limitations. 3D electron density maps of such complexes can be reconstructed from electron micrographs, but the resolution of these maps is typically insufficient for direct atomic modelling. A detailed understanding of the structure of macromolecular complexes therefore often requires fitting of atomic-resolution models of parts of the complex into comprehensive low-resolution density maps. The most widely used automated fitting methods aim to maximize the cross-correlation between the EM density map and a pseudo-map computed by convolving the atomic structure with a point-spread function. The main disadvantage of this approach is its computational cost, even after Fourier acceleration of the translational search. Run times of an hour or more become prohibitive if an analysis requires multiple runs, for instance with alternative model structures. The key role of surface information in visual fitting inspired us to devise a method that maximizes surface overlap. This approach allowed the formulation of an algorithm that is typically two orders of magnitude faster than publicly available general fitting methods, yet performs equally well on targets with sufficient surface exposure. Briefly, the template (a low-resolution density map) and the convolved target (a structure model) are described as 3D iso-surfaces. The vast search space is reduced by considering only those transformations that superimpose a pair of vectors

capturing local surface information. For all these transformations, the goodness-of-fit is quantified as surface overlap, i.e. the portion of the target surface that is projected onto the template surface. The set of best scoring transformations is then re-ranked by using more sophisticated scores including direct and Laplacian cross-correlation. An implementation of this method, 3SOM, has been written in ANSI-C and is freely available for all academic use at <http://www.russell.embl.de/3SOM>.

19. Align-m - a new algorithm for multiple alignment of highly divergent sequences

Ivo Van Walle, Ignace Lasters, Lode Wyns

We present a new program for multiple sequence alignment, called Align-m¹, which focuses on highly divergent sequences. It is freely available for academic use and can be downloaded from <http://bioinformatics.vub.ac.be>. We compare its performance with that of the well-known algorithms ClustalW, T-Coffee and DiAlign on 3 different measures: fraction of correctly (f_D) and incorrectly ($1 - f_M$) aligned residues, and running time.

For this purpose, 2 large testsets were used that represent the entire SCOP classification, and cover sequence similarities between 0 and 50% identity. In general, Align-m has comparable or slightly higher f_D , especially for distantly related sequences. Importantly, it aligns much less residues incorrectly, with significant average increases in f_M of over 15% compared to some of the other algorithms. The memory requirement is $O(N^2L^2)$ where N is the number of sequences and L their average length, and the time requirement is $O(N^2L^2 + N^3L)$. On average, for both testsets, ClustalW was the fastest, with DiAlign, T-Coffee and Align-m resp. 6x, 20x and 33x slower than ClustalW.

Align-m calculates a multiple alignment in 3 steps:

- (i) Based on the FASTER² optimiser, high scoring local alignments (columns of 1 residue width) are computed. These are expected to contain correctly aligned residues, but, depending on the similarity of the sequences, may also contain much incorrect alignments.
- (ii) This information is used to guide a global alignment of each pair of sequences, by dynamic programming with affine gap penalties. This is done by not taking all residue similarity scores from a substitution matrix but rather derived from (i) whenever a residue pair occurs in a high scoring local alignment.
- (iii) Regions of each pairwise alignment that are sufficiently consistent with other alignments are kept as the final alignment³. These pruned pairwise alignments are output, and can optionally be converted to a single multiple alignment by selecting 1 sequence as a reference to which all the others are aligned.

¹Van Walle I, Lasters I and Wyns L (2004) Align-m - a new algorithm for multiple alignment of highly divergent sequences *Bioinformatics*, in press

²Desmet J, Spriet J and Lasters I (2002) Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization *Proteins* 48, 31-43

³Van Walle I, Lasters I and Wyns L (2003) Consistency matrices - quantified structure alignments for sets of related sequences *Proteins* 51, 1-9

20. Identification of non coding RNA genes in the genome of the algae *Ostreococcus tauri*

Jan Wuyts, Pierre Rouzé, Yves Van de Peer

When annotating a genome, different properties of protein coding genes like codon bias and sequence conservation between species serve as guidelines in the hunt for genes. However, a number of genes are transcribed to RNA molecules but do not encode a protein sequence. These non coding RNA (ncRNA) genes often go undetected in a genome, especially when they show too little sequence conservation to be detected by programs as BLAST and FastA. Most of these ncRNA molecules have short stretches of complementary sequences which allows them to fold into characteristic secondary structure pattern. This secondary structure pattern is often conserved between different species even though the sequence itself is often very diverged. To annotate the different ncRNA genes we have to look for these conserved secondary structure patterns in the entire genome. Although software is available to do this, these algorithms are very CPU intensive, making it often impractical to look for these genes. Our group is responsible for the annotation of the genome of the algae *Ostreococcus tauri* and since this genome is very small (approximately 12Mb) we had a unique opportunity to scan for ncRNA genes in this genome, despite the slowness of the algorithms. Two of the ncRNA genes we detected are the Selenocysteine (Sec) insertion sequence and the Sec-tRNA gene. Until recently the Eukaryotic Sec protein insertion machinery was thought to be restricted to animals. The only other plant known to contain these genes is *Chlamydomonas reinhardtii*, another algae. In this organism, the homolog genes were discovered just last year.

21. Validation of the Complete Arabidopsis Transcriptome MicroArray, CATMA v1

Joke Allemeersch, Steffen Durinck, Yves Moreau, Bart De Moor, Pierre Hilson, and Martin Kuiper

The CATMA project aimed at the design and production of a collection of high quality Gene Sequence Tags (GSTs), covering most Arabidopsis genes and with a minimal homology (less than 70%). This work resulted in the first version of the Complete Arabidopsis Transcriptome MicroArray (CATMA), a novel microarray platform for Arabidopsis transcriptome analysis. Here a benchmarking of this CATMA array against a short- and long-oligo microarray alternative is presented. The coverage, sensitivity and reproducibility of each platform is assessed and compared by a carefully designed spike mix experiment.

22. The role of non-linear DNA structures in transcription

Kobe Florquin, Sven Degroeve Yvan Saeys, Yves Van de Peer

The assembly of the stereo specific DNA polymerase nucleoprotein requires proteins to bind to DNA in a sequence specific manner and the stereo specific assembly of this nucleoprotein complex necessitates that the DNA at least facilitate the architectural complex to be built. Because of the requirement of this physical support it is likely that specific structural features are present within a promoter. The role of non-linear

DNA structures in transcription is not directly linked to the proteins encoded or bound by the nucleotide sequence. DNA is unique in this respect; it is able to encode in its sequence at least two independent levels of functional information. The first one is for encoding proteins, whether structural or regulatory, and sequence targets for DNA-binding factors. The second level is contained in the physical and structural properties of the DNA-molecule itself. Although the physical and structural properties are ultimately determined in each case by the nucleotide sequence, these properties are exploited by the cell in a fashion in which the sequence itself plays no role other than to support or facilitate certain spatial structures. In this view bendability, intrinsic or protein induced curvature, propeller twist and etc become regulatory assets as building blocks for a productive promoter geometry. Our current research aims at studying the structural features in greater detail and tries to implement these different structural features in a general promoter model, which we then can use to find back new promoters on a genomic scale.

23. Design and implementation of a database for genome annotation

L. Lahlimi, C. Lambert, C. Lemer, J. Richelle, E. Depiereux

Following the complete sequencing of the genome of *Brucella militensis*, the Unité de Recherche en Biologie Moléculaire (URBM) was involved in the annotation of the predicted coding sequence (pCDS). Results lead to build a dedicated database system for storing functional and structural information about *Brucella militensis*. The technique used to build this database didn't allow the evolution of its structure to include other genomes and particularly those phylogenitically related to *Brucella militensis*. In addition, this database has become cumbersome to manage due to the continuous increase of knowledge and analysis features. To overcome these problems, we decided to implement a new database, with a more general schema, under a DBMS (DataBase Management System) that allows complex queries. To ensure the good course of this project, we have initiated collaboration with experienced teams: the aMAZE team at the Service de Conformation des Macromolécules Biologiques et de BioInformatique (SCMBB/ULB) and the DB-Main project at the Laboratoire d'Ingénierie des Bases de Données (LIBD/FUNDP). To carry out this project, our development plan defines several milestones. The first one consists of the definition of a conceptual model driven by the analysis of the available data and adapted to the application domain. The conceptual modeling will be achieved using DB-MAIN, a dedicated computer aided software engineering tool (CASE tool) from the DB-Main project. The second milestone will be the translation of the conceptual model into a physical schema. This step will be carried out using the Snow system developed by the aMAZE team. The third milestone will be the implementation of the database and the loading of the data currently stored in the initial database. A further milestone will be the development of dedicated tools for the exploitation of the database

24. Annotation of the poplar tree genome

Lieven Sterck, Stephane Rombauts, Sven Degroeve, Pierre Rouze, and Yves Van de Peer

The Poplar tree is a perennial plant with a great economical value, used for construction works and by the paper industry, but is also rapidly becoming the model organism for tree biotechnology. It has an estimated genome size of ~520 Mbp divided into 19 chromosomes, which is four times larger than the genome of Arabidopsis, yet 40 to 50 times smaller than the genome of pine, making the poplar an ideal model system for trees. As partner of the International Populus Genome Consortium (IPGC), in the cooperative effort to sequence and annotate the *Populus balsamifera* ssp. *trichocarpa* (black cottonwood), we are involved in the annotation of the nuclear genome. A whole-genome shotgun approach has been the adopted strategy to sequence the *Populus* genome. The JGI is responsible for sequencing the genome and provided a 3 x draft coverage of the female poplar tree clone Nisqually-1 in late-2003, while a second 3x draft will be provided in early 2004. Currently, the genome is being assembled using physical and genetic maps. Our research group, who gained an excellent reputation in genome annotation over the past years, will produce - together with two American groups - a high quality annotation for this new genome. Producing a high-quality annotation is a daunting task that needs to be accomplished as a prerequisite to publish a genome. We are currently building the gene data set necessary to train our software tool called EuGene. This training will provide a more poplar-specific prediction platform capable of predicting genes on this new genome with a higher accuracy.

25. BEN –The National Biocomputing Support for Belgian Scientific Researchers in Molecular Biology, Biotechnology and Biomedical Sciences, since 1993.

*Guy Bottu, David Coornaert, Valérie Ledent, Marc Colet & Robert Herzog
ULB, Campus de la Plaine, Bat. NO-CP 257, Boulevard du Triomphe, 1050
Bruxelles, Belgium*

BEN - the Belgian EMBnet node provides to scientific researchers in Belgium, both academic and industrial, an access to permanently updated databanks of nucleic acid and protein sequences, as well as the necessary software to search these databanks and to perform various sequence analysis tasks. BEN maintains a SRS server offering access to more than 40 databanks.

BEN sequence analysis services are centred on the freeware package EMBOSS as well as on other useful software such as BLAST, Clustal, HMMER, Phylip, Primer3.... These different tools have been integrated in order to make their use easier. BEN ensures a technical support service that users can call in case of problems and provides users with freeware of interest for the molecular biologist through its anonymous ftp server.

BEN has developed several graphical interfaces for various software (GCG, Blast ...). Recently, a new wEMBOSS interface is born from a coordinate effort from Prof. Marc Colet from BEN and Martin Sarachu from the Argentinian EMBnet Node. wEMBOSS allows web-users to access their personal data files and to save their sequence analysis results on the remote computer

BEN is a joint initiative of the ULB and the VUB. It is financially supported by the Belgian Science Policy. The computers of BEN are located at the VUB/ULB

Computing Centre. The functioning of BEN is largely depending on the existence of the Belgian national research network, BELNET. BEN is member of the EMBnet (European Molecular Biology Network) established in 1988 to link European laboratories distributing bioinformatics services to their national and/or regional scientific community.

Web site : <http://www.be.embnet.org>

Ftp site : <ftp://ftp.be.embnet.org>

26. wEMBOSS a web interface to the popular EMBOSS software package for sequence analysis

Martin Sarachu & Marc Colet

wEMBOSS is a free Open Source web interface for the EMBOSS free Open Source software analysis package. wEMBOSS requires the organization of all tasks into projects and subprojects, which provides an easy way to review all project related data at a later time. The user can load the data he wants to work with into each project, either by creating a file with the tools provided by the interface, uploading it from his local computer, or retrieving it from the databases available in EMBOSS. The results of each job are saved under the current project with indication to the name of the executed program and the exact date and time it was run, in a way they can be reviewed later. The results data can be also used as input for other programs and/or projects. wEMBOSS is a simple but powerfull interface, which can be installed in just a couple of minutes on top of an existing and functional EMBOSS installation. For most popular Linux distributions and Unix systems, wEMBOSS is an out-of-the-box solution. wEMBOSS is a coordinated effort from Martin Sarachu of the Argentinian EMBnet Node and Marc Colet from the Belgian EMBnet node. The work is based on EMBOSS-GUI an initial initiative of Luke McCarthy. wEMBOSS is released under the GPL license as well as EMBOSS. *wEMBOSS portal:* <http://ben.vub.ac.be:6080/wEMBOSS>

27. Information systems for automated functional analysis of gene expression data

Michal Okoniewski, Koen Van Leemput, Marleen Maras, Bart Naudts

Ontologies such as Gene Ontology (GO) or KEGG, together with accompanying databases of annotations are intended to organise existing biological knowledge, and group together genes with similar function. They are a good "prior knowledge" basis for qualitative or mixed analyses performed on microarray data. Ontologies are basically information systems that combine the functionality of taxonomies and thesauri, grouping together terms in a given area of knowledge in a form of a directed acyclic graph, which nodes are interpreted as various lexical relations (broader-narrower term, synonym, is a, etc.). GO includes ca. 20k terms grouped hierarchically in ontologies for molecular function, biological process and cellular component. Annotations are additions to the ontology, having the form of gene-term relationships (e.g. MITF <-> melanocyte differentiation). For human genes that we consider in our research, major sources of annotation with GO are databases available at Ensembl, Swiss Prot and GeneCards. We have combined these databases using approved gene names (HUGO) as identifiers, obtaining a repository of ca. 60k annotations. Other identifiers, such as GeneBank accession number are too ambiguous to be used for human annotation at the moment. Among genes on the 22k human VIB array used for toxicity marker genes selection in our project, almost 18k have a HUGO name assigned. The number of annotations in each of the databases was similar, but the content was in some 15% different. Most of existing systems using ontology annotations has only one input repository, so combining the three makes our searches more reliable and robust. The use of such database allows to analyse functionally lists of over- and under- expressed genes from experiments with human microarrays, as well as more intelligent applications that could enable analysts to combine prior biological knowledge with knowledge discovery techniques. This should result in functional association rules, functional clustering and classification, also for genes that do not have HUGO names assigned yet.

28. Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction

Nathalie Pochet, Frank De Smet, Johan A.K. Suykens and Bart L.R. De Moor

Microarrays are capable of determining the expression levels of thousands of genes simultaneously. In combination with classification methods, this technology can be useful to support clinical management decisions for individual patients in for example oncology. Our objective is to systematically benchmark the role of nonlinear versus linear techniques and dimensionality reduction methods. A systematic benchmarking study is performed by comparing linear versions of standard classification and dimensionality reduction techniques with their nonlinear versions based on nonlinear kernel functions with a radial basis function (RBF) kernel. Nine binary cancer classification problems, derived from seven publicly available microarray data sets, and twenty randomizations of each problem are examined. Three main conclusions can be formulated based on the performances on independent test sets.

1. When performing classification with least squares support vector machines (LS-SVM) (without dimensionality reduction), RBF kernels can be used without risking too much overfitting. The results obtained with well-tuned RBF kernels

are never worse and sometimes even statistically significantly better compared to results obtained with a linear kernel in terms of test set ROC and test set accuracy performances.

2. Even for classification with linear classifiers like LS-SVM with linear kernel, using regularization is very important.
3. When performing kernel principal component analysis (kernel PCA) before classification, using an RBF kernel for kernel PCA tends to result in overfitting. Kernel PCA with linear kernel gives better results in this case.

29. Computing action for storage information on micro-organisms hosted by the Belgian Co-ordinated Collections of Micro-organisms (BCCM)

Ndimubandi A., Herzog R.

The BCCM consists in a consortium of four Belgian culture collections of micro-organisms. The BCCM is a member of international projects and then, its data must be easily exchanged and universally available. The main goal assigned to this project is a development of a common computerized database containing the whole information available in the 4 collections and which makes it easy to manage. The Entity-Relation Model was used to create the relational scheme and the database structure was created using DBMain (a database conception software, Hainaut J.L., Computer Science Institute of the University of Namur, 1991-2002).

Four main entities constitute the core of the database: taxa, synonyms and strains. Each entity contains an identifier (primary key) and specific information. The additional information such as mode of preservation, literature, media, etc. is stored in new entities which are linked to the main entities by the mean of a one-to-many or a many-to-many relation. For this reason, the response to the query of a user is a combination of information from one, two or more entities. Because it was necessary to provide a friendly and full multi users system for end users, the implementation of the structure was made using Linux (free, stable and secure, Linux also offers multi-user and multitasking functions across a network), PostgreSQL (it is free, reliable and a powerful DBMS), Perl (a server-side scripting language, therefore no problem of browser compatibility: no dependence to the client browser), Zope (this open source application server enables teams to collaborate while creating and managing dynamic web sites).

About security, users are given access to the database according to tasks they have to perform and nobody can access the system without login. Different categories of users were defined according to the tasks they have to perform : the curator is able to insert new data, modify, validate new data and visualize them but he can not modify the structure. This task is to be accomplished by an informatician. The typist insert, modify and visualize data. A simple internal user can only visualize data.

The system also allows to keep track of modifications made in the database. Different information is saved when a modification occurred: the user, the date of modification the old value, the new value and the table containing data which have been modified. All public data are visible on line on the official BCCM site (<http://www/belspo.be/bccm>) and on BeBIF site (<http://www.be.gbif.net>) which integrates Belgian Biodiversity resources within a unified environment.

30. BIGRE : Bioinformatics Grid Ressources and Environments

Dugas O. (1) , Major J. (1) , Colet M. (1) Buyle P. (2), Dalon Q. (2), Englebert V. (2)

Mantrach A. (3), Salihoglu U. (3), Bersini H. (3)

The BIGRE project is a bioinformatics tool for the integration of bioinformatics data services. It is part of a regional-government program named WIST (Walloon Information Company Technologies) that is financed by the Walloon regional-government and involves researchers from three laboratories. Bioinformatics a complex field for developing an application that provides the integration of data services. This is due to the variable levels of bioinformatics applications, heterogeneity on the level of the quantities and sources of data as well as the diversity of trace programs. This heterogeneity forces BIGRE to be a generic "framework" which can therefore also be used in fields having similar constraints. BIGRE users have transparent access to various bioinformatics services (algorithms, data banks, documentations, etc) via a service directory. The BIGRE graphical user interface displays the description of the available services as well as plug-ins that allow the visualization and editing of data provided by these services.

BIGRE offers its users access to diverse bioinformatics tools by using a higher level of abstraction than what is currently available. This abstraction is obtained by using an ontology that provides the description of the services and data on the basis of their semantics. This allows BIGRE to offer a unique service by way of the construction of "workflows" that permit the fast expression of the possibilities (service-data combinations) offered by this abstraction. The distributed architecture of BIGRE takes into account various technical limitations: passage through firewalls, availability of services, replication and administration (management of users, groups, encryption, certificates, etc). Bioinformatics services (and services in general) are integrated via "wrappers" specific to each service. Access to each service is governed by policies defined by the service provider via an administration service. A service provider can create policies for free or paid access according to users or groups and policies that limit the access to certain resources. The development of each "wrapper" is assisted by a "Wrapper Development Toolkit".

- (1) Service de Bioinformatique – ULB IBMM Gosselies
- (2) Pôle Systèmes d'Information FUNDP Namur
- (3) IRIDIA ULB Bruxelles

31. A Snow/aMAZE demo.

Olivier Sand, Christian Lemer, Frédéric Fays, Erick Antezana, Fabian Couche, Simon De Keyzer, Olivier Hubaut, Jesintha Mary Maniraja, Hassan Anerhour, Xavier Santolaria, Jean Richelle, Shoshana Wodak

We will present a demonstration of Snow, a user-friendly interface for querying and browsing databases of networks (see the poster "The Snow system, a tool for representation and analysis of networks"). Snow supports iql, a simple language which enables users to perform complex queries knowing only the database conceptual model. Results can be viewed together or separately with a by-default or chosen set of their attributes. The set can be expanded to all attributes later if necessary. Some attributes are themselves expandable, allowing to navigate through the whole content of the database. Results can also be displayed as graphs generated on the fly, showing connections between related entities. For this demonstration, Snow will be interfacing aMAZE, our database for molecular interactions and cellular processes (see the poster on aMAZE). We will execute simple and complex interrogations of aMAZE running live on the remotely connected database server.

32. methBLAST and methPrimerDB: web-tools for PCR based methylation analyses of cancer-related genes

Pattyn F, Robbrecht P, Hoebeek J, De Paepe A, Speleman F, Vandesompele J

Epigenetic modification by DNA methylation plays an important role in tumorigenesis. Detection of abnormal methylation patterns provides general insights in tumour biology and is expected to be applied in cancer diagnosis or early detection of recurrence.

Among the DNA methylation detection techniques, PCR-based methods are a breakthrough in speed and sensitivity. Another step forward was the introduction of sodium bisulphite treatment of the DNA to introduce methylation-dependent sequence differences. Nowadays, the most frequently used DNA methylation analysis methods employ a combination of bisulphite treatment and PCR, and have proven to be extremely useful in validation of array based results from genome wide scanning of aberrant promoter methylation.

Although the PCR-based DNA methylation analysis methods are easy to use and sensitive, the design and experimental validation/optimisation of the primers is often difficult, labour intensive, and excludes a certain level of standardization and uniformity. Therefore, we developed a public repository holding essential assay information for the four major PCR-based methods for DNA methylation analysis: methylation-specific PCR (MSP), combined bisulphite restriction analysis (COBRA), bisulphite-PCR-SCCP (BiPS), and methylation-sensitive single-nucleotide primer extension (Ms-SNuPE). MethPrimerDB is a web based database with free public access to perform queries and to submit user based information. This tool stores its data in an Oracle database and is developed to contain validated primer pairs annotated with submitters details, gene information linked to LocusLink, assay information, and a direct link via PubMed to the article in which the PCR based methylation assay is described. To allow the evaluation of designed PCR primers, we implemented a similarity search program to compare primers against in silico bisulphite modified DNA. The tool is mainly developed to find primer binding sites and hence addresses specificity for PCR based assays that use bisulphate converted DNA as input material. The methBLAST web tool is based on NCBI's stand alone WWW BLAST server, and custom databases generated by an in house developed Perl script, which simulates CpG methylation and bisulphite modification. MethBLAST is available for human, mouse and rat genomes.

MethPrimer : DB <http://medgen.ugent.be/methprimerdb/>

methBLAST : <http://medgen.ugent.be/methblast/>

33. The Complete GENome Tracking database: an update

Paul J. Janssen, Anton J. Enright, Benjamin Audit, Ildefonso Cases, Leon Goldovsky, Nicola Harte, Victor Kunin, Christos A. Ouzounis

The constantly updated Complete Genome Tracking (COGENT) database of fully sequenced and published genomes holds, as of March 2004, the genomes of 129 Bacteria, 16 Archaea, and 12 eukarya (up from 89, 15, and 10, respectively, only 11 months ago). Taken together this represents a vast body of protein sequences (611,181) as compared to the nearly 1.32 million entries in the non-redundant protein database (July 2003) and the 143,790 entries in the highly curated SwissProt (release 42.9) database. The core of the database is composed of just two SQL tables that can be accessed via Perl DBI based scripts and modules allowing basic operations as well as complex querying. The genomes table contains genome related information such as the date of publication, source of data, proteome size, etc., while the proteins table holds the actual amino acid sequence data. Each genome is automatically assigned to a mnemonic species_code (e.g. HINF-KW2-01 for Haemophilus influenza strain KW2, version 01) and for each protein a unique identifier protein_id is produced, existing of the species_code followed by a dash and a number. This standard nomenclature contributes to a better consistency and reproducibility for large-scale computational analyses and improves the sharing of data and results. The core database is used as a front end in HTML [2] allowing selective downloads of protein sequence sets. The database is updated using DBI functionality but outdated genome information – including the older protein sequences - are retained in a separate table. Cogent can also be extended by a number of additional tables, for instance GeneQuiz derived annotations, cellular localizations, chromosomal positions, expression profiles, etc. For those that want to build a local SQL database a complete SQL dump (taking about 180 Mbytes of memory) of the core database is been made available.

[1] Janssen et al. (2003). COmplete GENome Tracking (COGENT): a flexible data environment for computational genomics. *Bioinformatics* 19 (11): 1451-1452.

[2] <http://maine.ebi.ac.uk:8000/services/cogent/>

34. Discovery of sequence variations with novoSNP

Peter De Rijk, Stefan Weckx, Christine Van Broeckhoven and Jurgen Del-Favero

Resequencing in different individuals followed by comparison of sequences traces is the standard and most reliable technique for discovering sequence variations. The technique is typically used to discover mutations by comparing DNA sequences of patients and controls. Another use of the technique is the discovery of Single Nucleotide Polymorphisms (SNPs) for the generation of high-density SNP based genetic maps: SNPs are the DNA sequence variations with frequent occurrence in the human genome. Because SNPs have a mean distance of 1-2 kb and can easily be used in automated high-throughput genotyping systems, they are becoming the markers of choice in genetic studies of complex diseases. The number of SNPs found in public databases like HGVbase is often too small for detailed genetic studies so that SNP discovery is the most obvious next step. Because of high-throughput sequencing techniques, data analysis is becoming the bottleneck for this approach. To analyse the sequence data in a fast but reliable automated way we developed novoSNP, a flexible expert system for the automated detection of SNPs and insertion-deletion

polymorphisms (INDELs). novoSNP aligns the traces to a reference sequence and scores each position for a set of features. A weighted global score is calculated based on the scores per feature. The results can be evaluated by means of an intuitive graphical environment where the potential polymorphisms are displayed, optionally filtered based on e.g. quality score. In case of SNP discovery, a high cutoff score can be used in order to reduce the number of false positives. For mutation analysis, a low cutoff score will be used to avoid false negatives. In a comparative study of novoSNP, PolyPhred and PolyBayes with human curated polymorphisms, novoSNP could correctly predict all SNPs and INDELs, where both other programs missed several.

35. The vector model for information retrieval and its application to the classification of the promoters of human and mouse caspase genes

Pieter De Bleser, Mohamed Lamkanfi, Michael Kalai, Bram De Craene, Geert Berx, Frans Van Roy and Peter VandenAbee

The computational studies of factors that orchestrate transcriptional regulation at the level of the DNA sequence typically follow a similar scheme: co-expressed genes are identified by cluster analysis of their expression data, while searching for motifs that are statistically over-represented, does the actual detection of potential transcription factor binding sites (TFBS). This focus on canonical TFBS has its limitations since transcription factors usually do not act alone: it is assumed that genome-wide control is achieved by combinatorial use of multiple sequence elements, including enhancers, silencers and insulators, and, not to forget, the role of post-translational modifications such as phosphorylation and acetylation. In order to circumvent these TFBS-associated limitations, we consider promoter sequences as bags of words words being the abstraction of regulatory sequence features -. According to this view, we propose the hypothesis that genes with similar expression patterns are likely to share a similar promoter context (have more words in common). The problem of predicting which genes have the potential to be co-regulated, based on the word content of their promoters, now becomes imilar to the problem of the automatic classification of the online textual information that has been growing explosively on the WWW in the last years. Solutions to this problem, such as the vector model for information retrieval, have been implemented successfully in search engines such as Google in order to help the people accessing the information. We have adapted this vector model for information retrieval in an attempt to solve the problem of predicting which genes in a given set have the potential to be co-regulated. In short, unique words are extracted from each promoter sequence and treated as features. Next, every promoter sequence is represented as a vector of the counts (or Boolean presences) of these words in this feature space. By using a distance metric, the distance between any pair of promoter sequences is evaluated and a distance matrix is constructed. Unsupervised clustering subsequently reveals whether any natural classification of promoter sequences occurs. In the present study, we have applied this approach to predict the classification of the promoters of the human and mouse families of caspase (cysteine aspartate protease) genes. Using cumulative hyper-geometric statistics, we show that the correlation between the predicted groups and experimental data is highly significant.

36. Comparison of the PhoPQ regulon in *Escherichia coli* and *Salmonella typhimurium*

Pieter Monsieurs, Kathleen Marchal, Sigrid Dekeersmaecker, Gert Thijs, Jos Vanderleyden, Bart De Moor

The PhoPQ system is a relatively ubiquitous pleiotropic transcriptional regulator that responds to external Mg²⁺ in both *Escherichia coli* and *Salmonella typhimurium*. Mutations in the PhoPQ system in pathogenic species such as *Salmonella* result in an attenuated virulence phenotype showing an increased sensitivity to antimicrobial peptides and acid pH, a decreased resistance to bile salts, deficiency in epithelial cell invasion and the inability to survive within macrophages. As compared to the non-pathogenic strains such as *E. coli*, the PhoPQ regulon in *Salmonella* seemingly has obtained novel targets that allow bacterial survival in the intracellular environment of the host. From this observation, we hypothesized that the PhoPQ regulon in both related species, *E. coli* and *S. typhimurium* must have besides a core of common target genes, a set of genes that has been specifically acquired during evolution in each of the species. These specific gene sets would then contribute to the specificities of the phenotype in each of the organisms. To verify this hypothesis, the composition of the PhoPQ regulon was compared between *E. coli* and *S. typhimurium* using a combination of expression- and motif data. To determine the overlap in the expression domain of the PhoPQ regulatory system of *Salmonella* and *E. coli*, PhoPQ-dependent genes were isolated from microarray datasets obtained in the appropriate conditions from respectively *E. coli* (described by Minagawa et al. [3]) and *S. typhimurium* (kindly provided by Bader et al. [1]). To distinguish direct from indirect targets, we searched for the presence of the regulatory motif (i.e. the binding site of a regulatory protein) in the promoter region of the identified PhoPQ-dependent genes. To this end a motif model corresponding to a small conserved DNA-sequence (T/G)GTTTA was used [2,3]. Based on this analysis that combines microarray- and sequence (motif) data, the direct PhoPQ-dependent regulon could be reconstructed in both species. Subsequent comparison of the regulons pointed towards a limited overlap of PhoPQ-dependent genes between *S. typhimurium* and *E. coli* and suggests a specialized function of this two- component system in both species (e.g. pathogenesis in *S. typhimurium*).

References

1. M. W. Bader, W. W. Navarre, W. Shiau, H. Nikaido, J. G. Frye, M. McClelland, F. C. Fang, and S. I. Miller, Regulation of *Salmonella typhimurium* virulence gene expression by cationic antimicrobial peptides. *Mol. Microbiol.*, 50:219-230, 2003
2. S. Lejona, A. Aguirre, M. L. Cabeza, V. E. Garcia, and F. C. Soncini, Molecular characterization of the Mg²⁺-responsive PhoP-PhoQ regulon in *Salmonella enterica*. *J. Bacteriol.*, 185:6287-6294, 2003
3. S. Minagawa, H. Ogasawara, A. Kato, K. Yamamoto, Y. Eguchi, T. Oshima, H. Mori, A. Ishihama, and R. Utsumi, Identification and molecular characterization of the Mg²⁺ stimulon of *Escherichia coli*. *J. Bacteriol.*, 185:3696-3702, 2003

37. Protein Prenyltransferases: Bedside-to-Bench & Genome-to-Bedside Issues

Sebastian Maurer-Stroh, Stefan Washietl and Frank Eisenhaber

Three different protein prenyltransferases (farnesyltransferase, geranylgeranyltransferases I and II) catalyze the attachment of prenyl lipid anchors to the C-termini of a variety of eukaryotic cellular proteins to facilitate protein-

membrane but also specific protein-protein interactions. Cross-specificity between farnesyl- and geranylgeranyltransferases and resulting alternative prenylation account for the tolerance of farnesyltransferase inhibitors in normal mammals while, under pathological conditions, lack of farnesylation can hinder the function of, e.g., oncogenic forms of Ras in transforming cells. However, up to now too many open questions as well as contrasting results for farnesyltransferase inhibition in experiments and clinical trials remain. We discuss possible altered functions of substrates in correlation with different states of sequential posttranslational processing or changes in anchor length as a consequence of modification by an alternative prenyltransferase. Furthermore, we analyzed protein prenyltransferase genes in complete genomes and discuss the existence and distribution of corresponding pseudogenes. Interestingly, while the number of exons remains comparable, the length of introns and consequently the gene sizes differ characteristically between type I and II prenyltransferases. Isoforms of the enzymes resulting from alternative splicing might contribute to unpredictable effects in clinical applications. Additionally, we discuss evolutionary relationships to the cholesterol biosynthetic pathway exemplified by structural homology of protein prenyltransferase β subunits with squalene-hopene cyclases. Besides being developed as anti-cancer agents, farnesyltransferase inhibitors show efficacy against an increasing number of parasitic diseases. A specially developed BLAST search strategy for protein prenyltransferases in genomic data of fungal and protozoan pathogens unveils a series of new pharmaceutical targets for prenyltransferase inhibition.

38. Predicting the binding specificity of miRNA to the 3' untranslated region of target genes

Slabbinck Bram, Van Criekinghe Wim

Since the discovery of microRNAs lin-4 and let-7 hundreds of microRNAs have been identified in a plethora of living organisms. Despite the initial focus on detecting new microRNAs a lot of questions regarding its function remain unresolved. It has been shown that the small RNA molecules have regulatory potential. Indeed, in animals, microRNAs function as translational regulators by the imperfect binding of the 3' untranslated regions (3'UTR) of so-called target genes. The main challenge is to identify the target gene, thereby proposing mechanism of action and elucidating the function of these microRNAs. To address that question we have designed a variation of the Smith-Waterman (SW) dynamic programming algorithm that accounts for the imperfect multiple binding complementarity characterized by bulges and mismatches. A "mask" is a matrix of scores which are used to extend the gap creation and the gap extension penalties in the conventional SW enabling the possibility to create gaps of more than one nucleotide on one sequence or on both sequences thereby accommodating the bulges. To detect a miRNA-target gene (3'UTR) binding pairs we have designed a graph to detect the multiple binding capability and to determine from which arm of the precursor hairpin the mature microRNA is derived. The algorithm was benchmarked on the well-known and defined microRNA-target gene pairs: lin-4, let-7, lin-14, lin-28 and lin-41. We could detect all the known binding duplexes. In summary, our algorithm can detect target genes of a given microRNA and can be used to search miRNA on a predefined target

39. SNPbox: high throughput automated primer design from gene to genome

Stefan Weckx, Peter De Rijk, Christine Van Broeckhoven, Jurgen Del-Favero

Single Nucleotide Polymorphisms (SNPs) are currently the markers of choice in genetic studies of complex diseases. In order to use SNPs as genetic markers, a SNP map must be constructed. The validation of SNPs from public databases and the discovery of novel SNPs by resequencing require the design of primers for PCR and sequencing. Most of the existing programs for primer design can design only one primerset at a time and are therefore time consuming for large-scale primer design. We developed a modular software package SNPbox that automates and standardizes the generation of primers suitable for usage in high throughput platforms. SNPbox relies on Primer3 for the primer design and combines this program together with other publicly available software tools such as BLAST, Spidey and RepeatMasker with newly developed algorithms. Of the 2500 primersets we designed using SNPbox so far, 95% successfully amplified on genomic DNA using uniform PCR conditions. SNPbox also can be used for the design of primersets for mutation analysis, STR marker genotyping and microarray oligo design. We present the SNPbox web server at <http://www.SNPbox.org>, which hosts the SNPbox web service as well as the data of SNPbox analysis of all Ensembl exons1. The data of this genome-wide SNPbox application can be visualized in Ensembl's ContigView through a DAS (Distributed Annotation System) annotation server and detailed primer information can be retrieved from our server.

1 Weckx, S., De Rijk, P., Van Broeckhoven, C., Del-Favero, J. SNPbox: web based high throughput primer design from gene to genome. (2004) Nucl. Acids. Res., Web server issue (in press)

40. Promoter analysis of def and glo-like mads-box genes through phylogenetic footprinting in eudicot plants.

Stefanie De Bodt, Gunter Theissen and Yves Van de Peer

Comparing the promoter regions of genes that are expressed in similar patterns both within the same species and across related taxa can greatly help in identifying cis-regulatory elements that confer conserved expression patterns. In particular the analysis of orthologous regulatory regions in multiple species can enhance current attempts to decipher the cis-regulatory code. Conversely, species-specific alterations in expression patterns are often thought to play an important role in generating interspecific variation. DEF and GLO-like MADS-box genes act as floral organ identity genes, specifying the development of petals and stamens in the 2nd and the 3rd floral whorl, respectively. These genes have been the subjects of extensive studies that try to unravel the structural complexity of extant flowering plants. How the regulation of these genes has evolved in diverse plant species has long been a mystery to plant developmental biologists. Different potential regulators have been suggested and expression patterns of some of these genes have been described. However, a comprehensive assessment of these data and a link to structural features in the upstream region of these genes have been lacking. Phylogenetic footprinting provides us with a technique to analyse promoters in an evolutionary context. However, until now, the application of this technique has been mainly limited to human-mouse comparisons or to closely related plant species all belonging, for example, to the Brassicaceae lineage. These studies do not provide a basis for selecting an adequate number of informative plant species for comparative analysis in more distantly related

plant species, neither do they describe the rates and modes of regulatory evolution in these species. The comparative promoter analysis of DEF and GLO-like MADS-box genes has revealed that phylogenetic footprinting can be successfully applied in the wide range of eudicot plant evolution, and that evolutionary events such as divergence after duplication and speciation are reflected in the cis-regulatory element composition of the upstream region of the genes involved. Moreover, this analysis provides us with candidate cis-regulatory elements responsible for organ- or developmental stage-specific expression.

41. Annotation of the green alga *Ostreococcus tauri*

Steven Robbens, Stephane Rombauts, Sven Degroeve, Pierre Rouze and Yves Van de Peer

In collaboration with the Laboratoire Arago, Banyuls, France, we are performing the full genome annotation of the unicellular green alga *Ostreococcus tauri*. This alga is the smallest eukaryotic organism described until now (its size is comparable to that of a bacterium) and has a nuclear genome of about 11.5 Mb, divided over 19 chromosomes ranging in size from 120 to 1500 Kb. *Ostreococcus tauri* was discovered in the Mediterranean Thau lagoon (France) in 1994. Its cellular organisation is rather simple: *O. tauri* has a relatively large nucleus with only one nuclear pore, a single chloroplast, one mitochondrion, one Golgi body and a very reduced cytoplasmic compartment. The presence of only one chloroplast and mitochondrion makes it interesting to use not only for evolutionary studies, but also for experimental studies. Phylogenetic analysis placed *Ostreococcus tauri* within the Prasinophyceae, an early branch of the Chlorophyta (green algae). Morphologically, the absence of flagella is the most typical characteristic compared to other green algae. The sequencing of the genome took place in the Laboratoire Arago, Banyuls, France, while the gene prediction and annotation will be done in our team. Once the annotation will be completed, a comparative analysis of *Ostreococcus tauri* and *Chlamydomonas reinhardtii*, another green algae, will be performed. This will provide an insight in the genome organization and gene content (different gene families) among the green algae. Afterwards, other members of the green lineage (*Physcomitrella patens* [moss], *Arabidopsis thaliana* and *Oryza Sativa*) will be used in comparative analyses. The diversity of species within the green lineage will give us the chance to see how plant genomes, and different gene families, have evolved over evolutionary time.

42. Text-Based Gene Profiling with Domain-Specific Views

Steven Van Vooren, Bert Coessens, Patrick Glenisson, Yves Moreau, and Bart De Moor

There is a trend in life sciences towards ever growing amounts of high-throughput assays. Data interpretation and formulation of hypotheses tend to lag behind compared to data acquisition, and although the first generation of statistical algorithms designed to scrutinize single, large-scale data sets have found their way into the biological community, an important challenge remains: to connect their results to existing knowledge. Despite the fairly large number of biological databases currently

available, a lot of relevant information is available in free-text format, and is hence unstructured and not directly amenable to large scale and automated analysis. Examples are textual annotations, scientific abstracts and full publications. Moreover, many of the public interfaces do not allow queries with a broaderscope than a single biological entity (gene or protein). Conversely, the process of successfully gaining insight into complex genetic mechanisms will increasingly depend on a complementary use of a variety of resources, including biological databases, ontologies, and specialized literature on one hand, and expert knowledge on the other. We therefore consider the knowledge discovery process to be of a cyclic nature, requiring several iterations between various information sources to extract a reliable hypothesis. It is at this point that we introduce TXTGate, a web-based framework that combines different literature sources from selected public biological resources in a flexible text mining system designed towards the analysis of groups of genes. Gene groups submitted to TXTGate are profiled by means of a summary of top scoring terms and can be further analyzed via subclustering according to their textual profiles. The textual profiles are constructed in correspondenceto a set of tailored vocabularies from which the user can make his/her choice, yielding term- as well as gene-centric views on the cluster of genes at hand. Through the use of these carefully crafted domain vocabularies (composed fromstandardized gene or term nomenclatures) we demonstrate TXTGate summaries are suited for cluster profiling as well as further exploration via a variety ofexternal databases and ontologies, hereby closing the analysis cycle. In this talk we discuss application architecture and principles, and explore relevant biological cases to illustrate how TXTGate provides added value in anactual research context. TXTGate is available online. A paper introducing our application has been accepted for publication in Genome Biology.

43. Abundance, distribution and composition of simple sequence repeats in the genomes of epsilon-Proteobacteria

Tom Coenye and Peter Vandamme

We determined the abundance, distribution and composition of simple sequence repeats (SSRs) in the genomes of epsilon-Proteobacteria and evaluated whether the presence of these repeats could explain the remarkable diversity observed in *Helicobacter pylori*. Our data show that members of the epsilon-Proteobacteria contain a large number of SSRs, which are evenly distributed over the genome. Small mononucleotide SSRs (3 - 8 bp), dinucleotide SSRs \geq 6bp and trinucleotide SSRs \geq 9 bp are evenly distributed over coding and non-coding regions; mononucleotide SSRs \geq 9 bp and tetranucleotide SSRs \geq 12 bp were more likely to be localised in non-coding regions. Differences in nucleotide composition between the genome and SSRs were noted for mono- and dinucleotide SSRs; most notably was the overrepresentation of poly(A) and poly(T) mononucleotide SSRs. SSRs in both *H. pylori* genomes have a very high upper length limit (up to 182 bp). The genomes of both *H. pylori* strains investigated seem to share some characteristics that suggest that the presence of repetitive DNA contributes to the generation of excessive intergenomic diversity. These properties include the presence of many more SSRs than expected by chance alone, the presence of much longer SSRs than seen so far, and the fact that they contain many SSRs in coding regions.

44. Validation of a gene network inference procedure that incorporates biological prior information, using in silico generated microarray datasets

Van Leemput K, Naudts B, Marchal K, Engelen K, De Moor B and Verschoren A

The problem of gene network inference, i.e., reconstructing transcriptional networks that underlie the cell responses from microarray data, is extremely complex. Due to the complexity of the interactions and the restricted number of experiments available, the inference problem is under determined. Therefore, it is still an open question what the exact data and experiment requirements are to produce reliable reconstructions of gene networks. In this study we address this question by the combined use of an in silico network generator that produces biologically realistic bacterial gene networks and corresponding microarray datasets, and a novel inference procedure that makes use of biological prior knowledge. The generator produces realistic networks based on the known part of the *E. coli* network. Simulated microarray datasets are generated from these networks, and prior knowledge, representing known data from literature and other experiments, is inserted. Aside from the network that will be probed by the inference procedure, a separate network structure is mixed with the data to simulate the noise present in a realistic experiment, where a lot of links are not exercised in the actual samples and specific conditions tested. As a first result of this study we observed that the quality of the inferred network not only depends on the structure of the network, but also on the specificities of the interactions between the genes and the sensitivity of the expression profiling experiments. Secondly, prior knowledge acting as a guide in the search for an optimal network has a major influence on outcome of

the inference procedure. Finally, applying an appropriate experiment design may significantly decrease the number of experiments needed.

45. EMBER a new tutorial on sequence analysis and bio computing

Viorica Ghita, Valérie Ledent, Robert Herzog, Ioannis Selimas, Marc Brugman, Terry Attwood

EMBER is a new tutorial on sequence analysis and bio computing developed by a consortium of teams*, including the Bioinformatics group led by Professor Robert Herzog. The project coordinator is Professor Terri Attwood from the University of Manchester: the principal authors include Ioannis Selimas, from the Manchester group and Marc Brugman from the Expert Centre for Taxonomic Identification. EMBER can be used as a practical component of traditionally taught classes, but can also be used by independent users, the online material being structured such that the chapters gradually increase in difficulty, from a basic introductory tutorial, to more advanced exercises and case studies. Each chapter is divided into several sections, providing: the AIM, background INFO (offering theoretical aspects of the subjects tackled), INSTRUCTIONS (presenting practical exercises on line), and References. The tutorial gives an overview of primary, composite and secondary protein databases, focusing on basic and more advanced areas through the use of Web tools, including: similarity searches (BLAST, PSI- BLAST); protein family analysis (PROSITE, eMOTIF, BLOCKS, PRINTS, Pfam); multiple alignment (Clustal, DIALIGN, T-COFFEE, CINEMA, Jalview); physicochemical parameters and profile prediction (ProtParam and ProtScale); transmembrane helix prediction (MEMSAT, TMpred); secondary structure prediction (Jpredet, NNPREPREDICT); 3D prediction, comparison and visualisation (RasMol, QuickPDB, Cn-3D); homology modelling (Swiss Model, Geno-3D); fold recognition (GenThreader, 3D-PSSM); phylogenetic analysis; SRS (sequence retrieval); mining the human genome; and so on.

*University of Manchester (United Kingdom), Swiss Institute of Bioinformatics (Switzerland), University of Nijmegen (The Netherlands), University of the Western Cape (South Africa), European Bioinformatics Institute (United Kingdom), Instituto Gulbenkian de Ciencia (Portugal), ULB University of Bruxelles (Belgium), Canada Institute for Marine Biosciences (Canada), Research Institute for Genetic engineering and Biotechnology (Turkey), Expert Center for Taxonomic Identification (The Netherlands).

References

- Attwood, T.K. on behalf of the EMBER consortium (2002) EMBER - A European MultimediaBioinformatics Educational Resource. EMBnet.news, 8(1), 15-19
- Mabey, J.E. & Attwood, T.K. (2001) EMBER: a European Multimedia Bioinformatics EducationalResource. CAL-laborate. A collaborative publication on the use of Computer Aided Learning fortertiary level physical sciences, 6, 13-16.

46. The Belgian Biodiversity Information Facility

Wautelet F., Duflost J., Mergen P., Herzog R.

BeBIF (<http://www.be.gbif.net>) is the Belgian National Participant Node of the worldwide biodiversity information network Global Biodiversity Information Facility (GBIF, <http://www.gbif.org/>). GBIF's main goal is to integrate and make worldwide biodiversity related data freely available to all. BeBIF fulfils the role and the tasks assigned to a national node by GBIF, is member of the GBIF NODES Committee, the GBIF Data Access and Data Interoperability sub-committee (DADI) and the Multilingual Working group. BeBIF plays an active role in related international organisations and projects like the European Network for Biodiversity Information (ENBI) and the Taxonomical Data Working Group (TDWG)

At the national level, there is an endeavour to build a bioinformatics infrastructure to integrate "Belgian" biodiversity resources. We play the role of national data viewing unit and gateway of biodiversity data within the Belgian Federal Science Policy Office initiative "Biodiversity.be". The central portal is combining interrogations of distributed and centralized databases. BeBIF plays the role of a Information Technology Center at national level having its IT specialist sharing their expertise with the other partners and recommend technical solutions according to their needs as well as guarantee the necessary interoperability between national initiatives to expose and exchange Belgian Biological data according to international standards (DarwinCore, ABCD) using IT tools (DIGIR, BioCASE data providers) developed in collaboration with GBIF and TDWG.

The main goal of BeBIF is that metadata and data listed in the definition of a node according to the Memorandum of Understanding (MOU) of GBIF can be found and provided in the appropriate standards to expose "Belgian" biodiversity information to its end-user and transmit them in the appropriate standard to the GBIF central portal in a distributed network. BeBIF goes also further in its investments by developing software and data validation tools in accordance with our user needs. Our mission is also to encourage and assist Belgian data nodes to share data with GBIF under common standards. We can also play the role of IT advisors and to some extent fulfill helpdesk activities.

Participants list

Stein Aerts

stein.aerts@esat.kuleuven.ac.be
KULeuven ESAT-SCD
Kasteelpark Arenberg 10
3001 Leuven

Joke Allemeersch

joke.allemeersch@esat.kuleuven.ac.be
e
ESAT-SCD K.U.Leuven
Kasteelpark Arenberg 10
3001 Heverlee (Leuven)

Hassan Anerhour

hassan@scmbb.ulb.ac.be
ULB-SCMBB

Erick Antezana

erick@amaze.ulb.ac.be
Service de Conformation de
Macromolécules Biologiques et de
Bioinformatique (SCMBB)
Université Libre de Bruxelles
Boulevard du Triomphe - CP 263
1050 Bruxelles

Torik Ayoubi

torik.ayoubi@gen.unimaas.nl
University of Maastricht
Genome Center, Department of
Population
Genetics
Universiteitssingel 50
P.O.Box 616, # 16
6200 MD Maastricht
The Netherlands

Christophe Biot

cbiot@ulb.ac.be
Bioinformatique Génomique et
Structurale,
CP 165/61
Université Libre de Bruxelles
50, Av. F. Roosevelt
B-1050 Bruxelles
Belgique

Eric Bonnet

erbon@psb.ugent.be
Plant Systems Biology
technologiepark 927
9052 gent

Guy Bottu

gbottu@dbm.ulb.ac.be
ULB - Bioinformatique (BEN)
12, Rue des Professeurs Jeener &
Brachet
6041 Gosselies

Sylvain Brohée

sylvain@scmbb.ac.be
ULB - SCMBB
Blvd du Triomphe
CP 263

Qianhua Cai

caiqh@163.net
Master program of Bioinformatics in
Leuven
Av. Vital Riethuisen 78
1083 Brussels

Hugo Ceulemans

Hugo.Ceulemans@med.kuleuven.ac.be
e
EMBL/Katholieke Universiteit Leuven
Faculteit geneeskunde, afdeling
biochemie
Campus Gasthuisberg, O&N
Herestraat 49
B-3000 LEUVEN

Tom Coenye

Tom.Coenye@UGent.be
Laboratorium voor Microbiologie,
Universiteit Gent
K.L. Ledeganckstraat 35
9000 Gent

Marc Colet

Marc.Colet@ulb.ac.be
ULB - Bioinformatique
12, Rue des Professeurs Jeener &
Brachet
6041 Gosselies

David Coornaert

dcoorna@dbm.ulb.ac.be
ULB - Bioinformatique (BEN)
12, Rue des Professeurs Jeener &
Brachet
6041 Gosselies

Fabian Couche

fcouche@scmbb.ulb.ac.be
SCMBB - ULB
Université Libre de Bruxelles
Service de Conformation de
Macromolécules Biologiques et de
Bioinformatique (SCMBB)
Boulevard du Triomphe - CP 263
1050 Bruxelles

Matias Cserhati

VIB - Ghent University

Pieter De Bleser

pieterdb@dmbr.UGent.be
Bioinformatics Core, Dept. of
Molecular
Biomedical Research, VIB & UGent
Fiers-Schell-Van Montagu Research
Building
Technologiepark 927 , B-9052
Zwijnaarde ,
Belgium

Stefanie De Bodt

stefanie.debodt@psb.ugent.be
Departement of Plant Systems
Biology,
Ghent
Technologiepark 927
9052 Ghent

Simon De Keyzer

simon@scmbb.ulb.ac.be

SCMBB - ULB

Université Libre de Bruxelles
Service de Conformation de
Macromolécules Biologiques et de
Bioinformatique (SCMBB)
Boulevard du Triomphe - CP 263
1050 Bruxelles

Marc De Maeyer

marc.demaeyer@fys.kuleuven.ac.be
celestijnenlaan 200 D
3001 Leuven
kuleuven

Peter De Rijk

Peter.DeRijk@ua.ac.be

Yves Dehouck

ydehouck@ulb.ac.be
Université Libre de Bruxelles
Av. Fr. Roosevelt 50, CP 165/64
1050 Bruxelles

Olivier Delgrange

Olivier.Delgrange@umh.ac.be
Université de Mons-Hainaut
Le Pentagone
Avenue du champ de Mars, 6
B-7000 MONS

Solange Demeure

sdemeure@ulb.ac.be
ULB - IBMM - Bioinformatique
12, Rue des Professeurs Jeener &
Brachet
6041 Gosselies

Benoit Dessailly

benoit@scmbb.ulb.ac.be
ULB - SCMBB
CP 263
Campus La Plaine,
Bld du Triomphe,
1050 IXL

Yves Deville

yde@info.ucl.ac.be
Université catholique de Louvain
Département d'informatique
Place Ste Barbe 2
1348 Louvain-la-Neuve

Croes Didier

didier@scmbb.ulb.ac.be
SCMBB - ULB
Service de Conformation des
Macromolécules Biologiques et de
Bioinformatique
Université Libre de Bruxelles
(Campus Plaine, building BC, C6, 6st
level)
bd du Triomphe - CP 263
B-1050 Bruxelles
Belgium

Grégoire Dooms

dooms@info.ucl.ac.be
Université catholique de Louvain
Département d'Ingénierie
informatique,
Batiment Réaumur,
place St Barbe 2,
1348 Louvain-la-Neuve

Luc Ducazu

luc@biolinux.org
BioLinux
Hoorndriesstraat 17
9820 Merelbeke

Luc Duchateau

Luc.Duchateau@UGent.be
Ghent University
Salisburylaan 133
9080 Merelbeke

Johan Duflost

jduflost@ben.vub.ac.be
Belgian Biodiversity Information
Facility
ULB, Campus Plaine; CP 257
blv du Triomphe
B-1050 Bruxelles

Olivier Dugas

odugas@dbm.ulb.ac.be
IBMM - ULB
Institut de biologie et de médecine
moléculaires
ULB CP300, rue des Professeurs
Jeener et
Brachet
12, 6041 Charleroi

Wim Dumon

wim.dumon@chello.be
KULeuven
Savoyestraat 12/3
3000 Leuven

Pierre Dupont

pdupont@info.ucl.ac.be
UCL - Comput. Science & Engin. Dept.
2, Place Sainte-Barbe
B-1348 Louvain-la-Neuve (Belgium)

Geneviève Dupont

gdupont@ulb.ac.be
ULB

Steffen Durinck

steffen.durinck@esat.kuleuven.ac.be
ESAT-SCD KULeuven
kasteelpark arenberg 10, 3001
Heverlee

TALBI EI-Ghazali

talbi@lfl.fr
University of Lille
Bat.M3 59655 villeneuve d'ascq

Slama Farsi

slfarsi@ulb.ac.be
ULB
Bd General Jacques 162
1050 Brussels

Frédéric Fays

frederic@scmbb.ulb.ac.be
SCMBB - ULB
Université Libre de Bruxelles
Service de Conformation de
Macromolécules Biologiques et de
Bioinformatique
Boulevard du Triomphe - CP 263
1050 Bruxelles

Kobe Florquin

koflo@psb.ugent.be

Bioinformatics & Evolutionary
Genomics
DEPARTMENT OF PLANT SYSTEMS
BIOLOGY GHENT UNIVERSITY/VIB
Technologiepark 927
B-9052 Gent, Belgiu
Laurent Gatto
lgatto@ulb.ac.be
Free University of Brussels - IBMM
Unit of Evolutionary Genetics
rue Jeener & Brachet
B-6041 Gosselies

Christophe Gengler
christophe.gengler@fundp.ac.be
Facultés Universitaires Notre-Dame de
la
Paix de Namur
Rue de Bruxelles 61
B-5000 Namur

Dirk Gevers
dirk.gevers@UGent.be
Bioinformatics & Evolutionary
Genomics
UGent/VIB
Technologiepark 927
B-9052 Gent

Peter Ghazal
p.ghazal@ed.ac.uk
Edinburgh National School
Scottish Centre for Genomics
Technology &
Informatics
Edinburgh - Scotland

Viorica Ghita
vghita@dbm.ulb.ac.be
ULB - Bioinformatique (EMBER)
Bd. du Triomphe
1050 Bruxelles

Patrice Godard
pgodard@dbm.ulb.ac.be
Université Libre de Bruxelles
rue des prof. Jeener et Brachet, 12
6041 Gosselies
Belgium

Bassem Hassan
bassem.hassan@med.kuleuven.ac.be
Laboratory of Neurogenetics
Department of Human Genetics

Flanders Interuniversity Institute for
Biotechnology (VIB)
University of Leuven School of
Medicine
O&N 06.547, Herestraat 49
3000 Leuven
Belgium

Robert Herzog
rherzog@dbm.ulb.ac.be
ULB – IBMM - Bioinformatique
12, Rue des Professeurs Jeener &
Brachet
6041 Gosselies

Oluwasegun Ijyemi
abayomi_ijiyemi@yahoo.com
Vrije Universiteit Brussel
Diabetes research centre(MEBO)
Laarbeeklaan 103, 1090 Jette
Brussels

Paul Janssen
pjanssen@sckcen.be
Belgian Nuclear Research Centre
Boeretang 200, B-2400-MOL, Belgium

Koen Janssens
*Koen.Janssens3@student.kuleuven.a
c.be*
KULeuven
Collegestraat 37
2220 Heist op den Berg

Frizo Janssens
frizo.janssens@esat.kuleuven.ac.be
Kasteelpark Arenberg 10
3001 Heverlee (Leuven)

Richard Kamuzinzi
rkamuz@dbm.ulb.ac.be
ULB - IBMM - Bioinformatique
12, Rue des Professeurs Jeener et
Brachet
6041 Gosselies
Luc Krols
Luc.Krols@vib.be
VIB / Peakadilly
Rijvisschestraat 120
9052 Zwijnaarde

Jean Marc Kwasigroch
Jean.Marc.Kwasigroch@ulb.ac.be
Bioinformatique et Génomique
Structurale
Université Libre de Bruxelles
(CP 165/61)
50, Av. F. Roosevelt B-1050 Bruxelles
Belgique
Pieter Laeremans
pieter@laeremans.org
K.U. Leuven (student)
Schapenstraat 48
3000 Leuven

Leila Lahlimi
leila.lahlimi@fundp.ac.be
FUNDP

Christophe Lambert
christophe.lambert@bioxpr.be
BioXpr s.a.
Rue du Séminaire, 22
5000 NAMUR

Jana Leban
janaleban@yahoo.ca
4 rue de la Fourragère
1180, Bruxelles

Valérie Ledent
valerie.ledent@ulb.ac.be
U.L.B. - I.B.M.M. - Bioinformatique
12, Rue des Professeurs Jeener &
Brachet.
6041 Gosselies

Christian Lemer
chris@scmbb.ulb.ac.be
SCMBB - ULB

Raphael Leplae
raphael@scmbb.ulb.ac.be
ULB
CP 263
Bld du Triomphe
1050 Bruxelles

Gispi LIMA
gipsi@scmbb.ulb.ac.be
ULB - SCMBB

Bd. du Triomphe
1050 Bruxelles

Marc Logghe
marc.logghe@devgen.com
Devgen
Technologiepark 30
9052 Gent-Zwijnaarde

Ari Loytynoja
ari@ebi.ac.uk
EBI

Jesintha Mary Maniraja
jesintha@scmbb.ulb.ac.be
Université Libre de Bruxelles
CP263
SCMBB Department
Boulevard du Triomphe
Bruxelles,
Belgium - 1050

Amin Mantrach
amantrac@ulb.ac.be
ULB
87, Avenue Adolphe Buyl
Batiment C, IRIDIA
1050 Bruxelles

Kathleen Marchal
kathleen.marchal@esat.kuleuven.ac.be
e
KUL/ESAT

Geert Martens
Geert.Martens@vub.ac.be
Vrije Universiteit Brussel
Diabetes Research Center,
Faculteit Geneeskunde en Farmacie
Laarbeeklaan 103
1090 Brussel

Sebastian Maurer-Stroh
stroh@imp.univie.ac.at
IMP - Institute of Molecular Pathology
Bioinformatics - Group Eisenhaber
Dr.Bohrg. 7 1030 Vienna, Austria

Joseph Mavor
jmavor@dbm.ulb.ac.be
IBMM - ULB
Bioinformatics Dept.

Jeroen Meeus
jeroen.meeus@ugent.be
Universiteit Gent
Stropstraat 13
9000 Gent

Gerben Menschaert
gerbenm@devgen.com
deVGen
Technologiepark 30
9000 Gent

Bjorn Menten
bjorn.menten@ugent.be
UGent

Michel Milinkovitch
mcmilink@ulb.ac.be
IBMM, Free University of Brussels
(ULB)
12 rue Jeener & Brachet, 6041
Gosselies,
Belgium.

Pieter Monsieurs
Pieter.Monsieurs@esat.kuleuven.ac.be
e
KULeuven - ESAT

Alberic Ndimubandi
andimuba@ulb.ac.be
ULB - Bioinformatique (BCCM)

Bd. du Triomphe
1050 Bruxelles

Irene Nooren
irenen@devgen.com
Devgen NV
Technologiepark 30
9052 Gent-Zwijnaarde

Michal Okoniewski
micah.okoniewski@ua.ac.be

Fred Opperdoes
opperdoes@bchm.ucl.ac.be
UCL

Filip Pattyn
Filip.Pattyn@UGent.be
Universiteit Gent
Centrum voor Medische Genetica,
UZ Gent, 1K5,
De Pintelaan 185,
9000 Gent

Nathalie Pochet
nathalie.pochet@esat.kuleuven.ac.be
KULeuven, ESAT-SCD, Biol
Kasteelpark Arenberg 10
3001 Heverlee (Leuven)
Belgium

Steven Robbens
steven.robbens@psb.ugent.be
Ugent
Technologiepark 927
9052 Gent

Stephane Rombauts
stephane.rombauts@psb.ugent.be
BIOLOGY, GHENT UNIVERSITY, VIB
DEPARTMENT OF PLANT SYSTEMS
Technologie Park 927, B-9052 Gent,
Belgium

Pierre Rouzé
pierre.rouze@psb.ugent.be
INRA / VIB
Department of Plant Systems Biology
Ghent University, Technologie park
927, B-

9052 GENT

Jean-Louis Ruelle

jean-louis.ruelle@gskbio.com

GSK Biologicals
Rue de l'Institut, 89
1330 Rixensart
Belgium

Skhiri Gabouje Sabri

sskhirid@ulb.ac.be

ULB
UB4.131, CP 165/15
Université Libre de Bruxelles
Avenue F. D. Roosevelt 50
B-1050 Bruxelles
Belgique

Olivier Sand

oly@amaze.ulb.ac.be

SCMBB - ULB
Campus Plaine, bd du Triomphe -
CP263
B-1050 Bruxelles

Xavier Santolaria

xsa@scmbb.ulb.ac.be

Martin Sarachu

mad@biol.unlp.edu.ar

AR.EMBnet
115 entre 49 y 50, 1900, La Plata,
Argentina

Eddie Schrevens

eddie.schrevens@agr.kuleuven.ac.be

KULeuven

Gert Sclep

Gert.Sclep@psb.ugent.be

VIB - Plant Systems Biology
Technology Park 927
B-9052 Gent
Belgium

Qizheng SHENG

Qizheng.Sheng@esat.kuleuven.ac.be

ESAT-SCD-SISTA

Kasteelpark Arenberg 10
B- 3001 Leuven - Heverlee
Belgium

Cedric Simillion

cedric.simillion@psb.ugent.be

UGent

Fernanda Sirota Leite

fernanda@scmbb.ulb.ac.be

SCMBB - Service de Conformation de
Macromolécules Biologiques et de
Bioinformatique
Université Libre de Bruxelles
Boulevard du Triomphe - CP 263
1050 Bruxelles

Jean-Pierre Szikora

szikora@icp.ucl.ac.be UCL 7459

UCL

74, av Hippocrate
B - 1200 Bruxelles

Denis Thieffry

thieffry@ibdm.univ-mrs.fr

Université de Marseille LGPD-IBDM
FRANCE

Gert Thijs

gert.thijs@esat.kuleuven.ac.be

K.U.Leuven
ESAT-SCD
Kasteelpark Arenberg 10
3001 Leuven

Gregoire Thomas

gregoire.thomas@ugent.be

UGent - VIB09

A. Baertsoenkaai 3, 9000 Gent,
Belgium

Morgane THOMAS - CHOLLIER

morgane.thomas-chollier@wanadoo.fr

V.U.B. / U.L.B.
Pleinlaan 2 - 1050 Bruxelles

Joseph TRAN

jtran@scmbb.ulb.ac.be

Université Libre de Bruxelles
Service de Conformation de
Macromolécules Biologiques et de
Bioinformatique (SCMBB)
Boulevard du Triomphe - CP 263
1050 Bruxelles

Przemko Tylzanowski

przemko@med.kuleuven.ac.be
University of Leuven

Alfonso Valencia

valencia@cnb.uam.es
Univ. Autonoma de Madrid
Centro de Biotecnologia
Madrid - ESPANA

Wim Van Criekinge

wim.vancriekinge@ugent.be
UGent

Yves Van de Peer

yvdp@psb.ugent.be

Raf Van de Plas

raf.vandeplas@esat.kuleuven.ac.be
Katholieke Universiteit Leuven
Kasteelpark Arenberg 10
3001 Heverlee - Leuven

Tim Van den Bulcke

timvandenbulcke@pandora.be
Tibotec BVBA
Gen De Wittelaan L 11B 3
2800 Mechelen

Ruth Van Hellemont

ruth.vanhellemont@esat.kuleuven.ac.be
ESAT-SCD
Kasteelpark Arenberg 10 3001
Leuven (Heverlee)

Koen Van Leemput

koen.vanleemput@ua.ac.be

Steven Van Vooren

Steven.VanVooren@esat.kuleuven.ac.be
Departement Elektrotechniek
Katholieke Universiteit Leuven
Kasteelpark Arenberg 10
3001 Leuven (Heverlee)

Ivo Van Walle

ivwalle@vub.ac.be
Vrije Universiteit Brussel
Pleinlaan 2
1050 Brussel

Stefan Van Yper

Stefan@biomath.UGent.be
University Ghent
Faculteit Landbouwkundige en
Toegepaste
Biologische Wetenschappen
Coupure Links 653
9000 Ghent
Belgium

Klaas Vandepoele

Klaas.Vandepoele@psb.ugent.be
PSB - VIB Ugent
Technologiepark 927
9052-Gent

Steven Vercauysse

stcru@psb.ugent.be
UGent / VIB
Technologiepark 927
9000 Gent

Frank Vernailen

frank.vernailen@pandora.be
Kolveniersgang 115
9000 Gent

Ghita Viorica

vghita@dbm.ulb.ac.be
ULB
Brussels, Belgium

Dominique Vlieghe

dominique.vlieghe@dmbr.ugent.be
DMBR/VIB1 UGhent

Technologiepark 927
9052 Gent-Zwijnaarde

ULB CP 165/15

Guojun Wang

gwang@ulb.ac.be

ULB

Campus Plaine, bd du Triomphe -
CP263
B-1050 Bruxelles

Frédéric Wautelet

fwautele@ben.vub.ac.be

Belgian Biodiversity Information
Facility
(BeBIF)

ULB, Campus Plaine CP 257
blv du Triomphe
B-1050 Bruxelles

Stefan Weckx

stefan.weckx@vib.be

VIB MicroArray Facility (MAF)
UZ Gasthuisberg - Onderwijs en
Navorsing
Herestraat 49
3000 Leuven

Didier Willame

didier.willame@ieee.org

ULB - Sc. Appliquees - SLN
4 rue de la Fourragere
1180, Bruxelles

Jan Wuyts

jan.wuyts@psb.ugent.be

Universiteit Gent
Bioinformatics research group
Department of Plant Systems Biology
University of Gent / VIB
Technologiepark 927
B-9052 Gent

Stéphane Zampelli

2, Place Sainte-Barbe
B-1348 Louvain-la-Neuve (Belgium)
Comput. Science & Engin. Dept.

Esteban Zimanyi

ezimanyi@ulb.ac.be

Salle Dupréel Campus du Solbosh Université libre de Bruxelles

