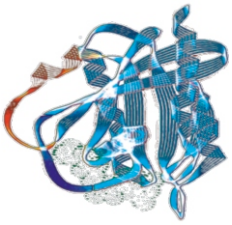
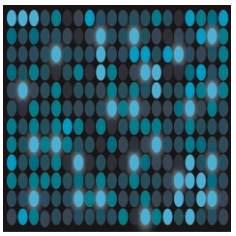
The First
BENELUX BIOINFORMATICS CONFERENCE
April 14 & 15, 2005





Ghent, Belgium

UNIVERSITEIT GENT

VIB
VLAAMS INTERUNIVERSITAIR INSTITUUT VOOR BIOTECHNOLOGIE

Acknowledgements

Organizing Committee

- Jeroen Raes
- Stephane Rombauts
- Steven Maere
- Francis Dierick
- Pierre Rouzé
- Yves Van de Peer

Scientific Committee

- Martijn Huynen (University of Nijmegen)
- Jack Leunissen (University of Wageningen)
- Kathleen Marchal (University of Leuven)
- Jan-Peter Nap (Hanze University & Rijksuniversiteit Groningen)
- Pierre Rouzé and Yves Van de Peer (Ghent University)
- Antoine van Kampen (Universiteit Amsterdam)
- Jacques van Helden (Université libre de Bruxelles)
- Eric Depiereux(University of Namur)

Sponsors

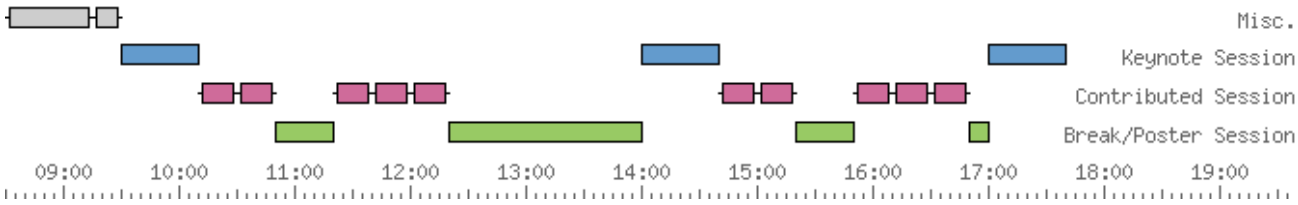
We gratefully acknowledge the support of our sponsors:

Sun Microsystems
Bayer Crop Science
Oracle Life Sciences
Keygene
Network Appliance
FWO

Abstract book also available as pdf:

http://bioinformatics.psb.ugent.be/bbc2005/

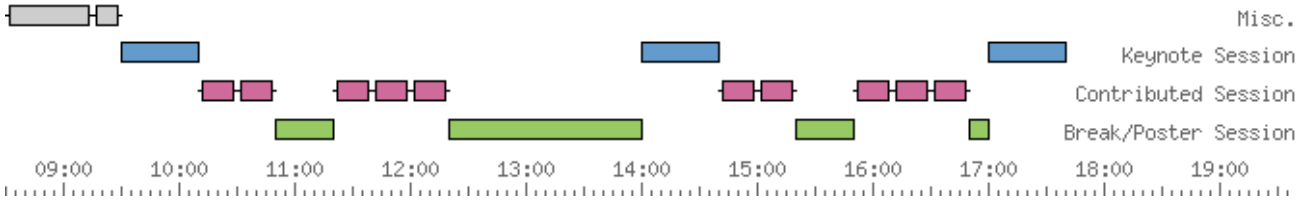
Day one (April 14) : Networks and Data Mining



Schedule details

- **08:30 - 09:15** : *Posters preparation.*
- **09:15 - 09:30** : *Welcome talk.* Yves Van de Peer
- **09:30 - 10:10** : *Keynote.* Berend Snel. Comparative transcriptomics: the evolution of co-regulation.
- **10:10 - 10:30** : Karen Lemmens. Discovering regulatory modules from heterogeneous information sources.
- **10:30 - 10:50** : Peter Van Loo. Parameterless in silico identification of novel cis-regulatory modules.
- **10:50 - 11:20** : *Coffee Break / Poster Session*
- **11:20 - 11:40** : Patrick Ogao. Effective Exploration of Biological Network Information in Immersive Virtual Environments.
- **11:40 - 12:00** : Christian Lemer. aMAZE: a multi-layer data model for biological processes.
- **12:00 - 12:20** : Qizheng Sheng. Query-driven biclustering for microarray data analysis by Gibbs sampling.
- **12:20 - 14:00** : *Lunch Break / Poster Session*
- **14:00 - 14:40** : *Keynote* Laurent Duret. Relationships between genome organization and gene expression in mammals.
- **14:40 - 15:00** : Antoine van Kampen. A comparison of the human and mouse transcriptome map: are RIDGES conserved?
- **15:00 - 15:20** : Xin-Ying Ren. Local co-expression domains in the genome of Arabidopsis thaliana.
- **15:20 - 15:50** : *Coffee Break / Poster Session*
- **15:50 - 16:10** : Nathalie Pochet. M@CBETH: optimizing clinical microarray classification.
- **16:10 - 16:30** : Sacha van Hijum. SIMAGE: Simulation of DNA MicroArray Gene Expression data.
- **16:30 - 16:50** : Kristof Engelen. Normalization of cDNA microarrays using external control spikes.
- **16:50 - 17:00** : *Short Break*
- **17:00 - 17:40** : *Keynote.* Robert Gentleman. Using Categories to analyse genomic data.

Day two (April 15) : Genomics and Proteomics



Schedule details

- **08:30 - 09:15** : *Posters preparation.*
- **09:15 - 09:30** : *Welcome talk.* Yves Van de Peer
- **09:30 - 10:10** : *Keynote.* Steven Maere. Modeling the birth and death of genes in Arabidopsis thaliana to explain plant evolution and complexity.
- **10:10 - 10:30** : Rekin's Janky. A taxonomy-traversing approach to discover cis-acting elements in prokaryotes.
- **10:30 - 10:50** : Ruth Van Hellemont. A novel approach to identify regulatory motifs in distantly related genomes.
- **10:50 - 11:20** : *Coffee Break / Poster Session*
- **11:20 - 11:40** : Tom Michoel. Physical stability of coding and non-coding DNA regions.
- **11:40 - 12:00** : Steven Van Vooren. CGHGate: Array-CGH and Human Genome Annotation.
- **12:00 - 12:20** : Jeroen Raes. Functional divergence of proteins through frameshift mutations.
- **12:20 - 14:00** : *Lunch Break / Poster Session*
- **14:00 - 14:40** : *Keynote.* Anders Krogh. A plant has hundreds of non-conserved microRNAs.
- **14:40 - 15:00** : Henk-Jan Joosten. A Novel Database System For Easy Correlation Of Heterogeneous Data Of One Protein Superfamily Using a Structure Based Multiple Sequence Alignment.
- **15:00 - 15:20** : Ivo Van Walle. Align-m 2: multiple alignment with focus on specificity rather than sensitivity.
- **15:20 - 15:50** : *Coffee Break / Poster Session*
- **15:50 - 16:10** : Benoit H. Dessailly. Prediction of Functional Sites in Proteins using the Relationship between Stability and Function.
- **16:10 - 16:30** : Alexandre M.J.J. Bonvin. Data-driven docking for the study of biomolecular complexes: combining WHISCY with HADDOCK.
- **16:30 - 16:50** : Pierre Geurts. Proteomic and genomic data classification using decision tree based ensemble methods - Application to the diagnosis of inflammatory diseases.
- **16:50 - 17:00** : *Short Break*
- **17:00 - 17:40** : *Keynote.* Benno Schwikowski. Computational methods in high-throughput mass spectrometry.

Oral Presentations

Discovering regulatory modules from heterogeneous information sources

K. Lemmens [1], T. De Bie [1], P. Monsieurs [1], K. Engelen [1], B. De Moor [1], N. Cristianini [2] and K. Marchal [3]
[1] K.U.Leuven, ESAT-SCD, Kasteelpark Arenberg 10, 3001 Leuven-Heverlee, Belgium; [2] U.C.Davis Dept. of Statistics, 360 Kerr Hall, One Shields Ave, CA 95616, US; [3] K.U. Leuven, Centre of Microbial and Plant Genetics and ESAT-SCD, Kasteelpark Arenberg 20, 3001 Leuven-Heverlee, Belgium

Nowadays, data representative of different cellular processes are being generated at large scale. Based on these "omics? data, the action of the regulatory network that underlies the organism's behavior can be observed. Whereas until recently bioinformatics research was driven by the development of methods that deal with the analysis of each of these data sources separately, the focus is now shifting towards integrative approaches treating several data sources simultaneously.
Indeed, by analyzing each of these "omics? data separately, different aspects of the cellular adaptation are studied independent of each other. However, combining data allows gaining a much more holistic insight into the network studied.
Besides an added biological value, the simultaneous analysis of coupled datasets also has a technical advantage. High-throughput experiments are characterized by a low signal/noise ratio. Moreover, because of technical and biological limitations, only a restricted number of independent experiments are available (technical replica's). The amount of information of a high-throughput experiment thus is limited. Although sometimes from a different angle, coupled datasets provide information on the same biological system. Combining such datasets will, therefore, increase the confidence in the final analysis result (higher specificity).

We present in this study a combinatorial method for inference of transcriptional modules from 3 independently obtained heterogeneous data sources: ChIP-chip data (chromatin immunoprecipitation on arrays) provide information on the direct physical interaction between a regulator and the upstream regions of its target genes; motif information as obtained by phylogenetic shadowing describes the DNA recognition sites of these regulators in the promoter regions of the target genes; and microarray experiments identify the expression behavior of the target genes in the conditions tested. Combining these 3 types of "omics? data allows reconstructing the structural composition of the basic building blocks of transcriptional networks i.e. transcriptional modules.

Our approach distinguishes itself from previous work (Bar-Joseph et al., 2003; Lapidot and Pilpel, 2003), in that most existing approaches exploit the availability of heterogeneous data sources in a sequential or an iterative way. Our method takes the different data sources into account in a highly concurrent way. By doing so, our method allows correlating a set of regulators with their corresponding regulatory motifs and elicited profiles in a very natural and direct way.
Using our method on publicly available yeast data allowed demonstrating the biological relevance of the inference.

References

Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, Gifford DK. (2003). Computational discovery of gene modules and regulatory networks. Nat Biotechnol, 21:1337-1342

Lapidot M and Pilpel Y. (2003). Comprehensive quantitative analyses of the effects of promoter sequence elements on mRNA transcription. Nucleic Acids Res, 31(13): 3824-3828.

karen.lemmens@esat.kuleuven.ac.be Karen Lemmens

Parameterless in silico identification of novel cis-regulatory modules

Peter Van Loo, Stein Aerts, Jan Cools, Yves Moreau, Bart De Moor and Peter Marynen
Department of Human Genetics, Flanders Interuniversity Institute for Biotechnology and University of Leuven, Bioinformatics group, Department of Electrical Engineering, University of Leuven, Belgium

Metazoan transcription regulation occurs through the concerted action of multiple transcription factors that bind cooperatively to cis-regulatory modules (CRMs). Existing computational methods to detect these CRMs often assume or parameterize many inherent characteristics (e.g. number of transcription factor binding sites, spatial extension of the CRMs). We investigated possibilities to eliminate these parameters. We developed ModuleMiner, a novel CRM-detection algorithm that does not require any prior knowledge to be given by the user.
Cis-regulatory modules in a set of co-regulated genes are modelled as a combination of motif models (position weight matrices), supplemented by a number of parameters representing inherent spatial characteristics. The ModuleMiner algorithm finds the optimal CRM model, optimising the number of motif model building blocks, the spatial characteristics and the combination of motif models, using a genetic algorithm. The resulting CRM models can be used in a genome-wide search to identify new target genes, predicted to be co-regulated with the original set of genes. We show that these CRM models indeed have predictive power by both in silico and in vitro validation.
Often, multiple similar CRM models score highly against a set of co-regulated genes. Using only the best scoring CRM model neglects this intrinsic variability. Similar high scoring CRM models can be clustered and used to construct large CRM supermodels. These supermodels are expected to be less noise sensitive than the individual CRM models.

For the computational validation, we performed a leave-one-out cross-validation on a tightly co-regulated set of smooth muscle genes. We used the area under the curve (AUC) of the resulting “rank ROC curves? as a measure of performance. This way, we showed that the ModuleMiner algorithm (AUC = 0.80) performs equally well as the ModuleSearcher algorithm (AUC = 0.81) with an optimal choice of parameters (note that the optimal choice of parameters is problem dependent), and clearly outperforms the ModuleSearcher algorithm when suboptimal parameters are chosen (AUC = 0.71).
When we used CRM model clustering to construct CRM supermodels, performance was significantly improved (AUC = 0.85).

For the experimental validation, the ModuleMiner algorithm was used to construct a CRM model responsible for the up-regulation of a set of genes during TPA-induced differentiation of the myeloid cell line HL-60 (this set was derived from microarray data). Using this model, we identified novel target genes in the human genome, hypothesised to be co-regulated with the original set. By PCR, we could show that 80 % of these novel target genes were indeed expressed in differentiated HL-60 cells. Up-regulation was validated using real-time quantitative PCR. 42 % of the assayed genes were found to be significantly up-regulated. These results indicate that the constructed CRM model is loosely (but significantly) correlated with up-regulation during differentiation and very highly correlated with expression in differentiated HL-60 cells.

In summary, the ModuleMiner algorithm performs comparably to the existing CRM detection algorithms, provided that the latter are used with an optimal set of parameters, which is problem dependent. Our in vitro validation showed that the predicted CRM models correlate well with gene expression, but correlation drops steeply from the general (i.e. expression) to the more specific level (i.e. significant up-regulation).
We have also shown in silico that extending the method to create large CRM supermodels raises performance significantly.

Peter.VanLoo@med.kuleuven.ac.be Peter Van Loo

Effective Exploration of Biological Network Information in Immersive Virtual Environments

Patrick Job Ogao, Oscar Kuipers, Paul Welinga, Bram Stolk, Jorrit Adriaanse, Menno Visser and Jos Roerdink
University of Groningen, SARA Computing and Netwroking Services

Motivation: Data generated by high-throughput measure-ment technologies does embed a considerable amount of hidden fundamental data relations. In order to exploit their contents, a variety of visual and computational approaches are employed. These approaches utilize, (i) heterogeneous data types, (ii) software tools which run on different hard-ware and software configurations and, (iii) detached and incoherent visual displays thus inhibiting ones exploratory experience with the data sets.
Results: To overcome these limitations, we present a us-ability experience with a virtual reality framework for genomic visualization, SARAgene. By using results obtained from a usability evaluation of one of SARAgene's components that visualizes online protein interaction data sets as a graph, we describe an extension SARAgene-DBTBS that enables an integrated presentation and exploration of transcriptional regulation network data Key features described include: vis-ual aesthetics, multiple graph object selection and manipula-tion interface, and interactive techniques that enable one to dynamically link, a graph view with pathway and chromo-some maps from online biological databases.

Note: SARAgene is freely available for academic purposes.

ogao@cs.rug.nl Patrick Ogao

aMAZE: a multi-layer data model for biological processes.

Christian LEMER, Hassan ANHEROUR, Erick ANTEZANA, Fabian COUCHE, Simon DE KEYZER, Yves DEVILLE, Frederic FAYS, Olivier HUBAUT, Olivier SAND, Jacques VAN HELDEN, Shoshana WODAK, Jean RICHELLE.
1 Service de Conformation des Macromolecules Biologiques et de Bioinformatique - ULB

The goal of the aMAZE project is to develop a workbench for the representation and analysis of heterogenous networks of molecular interactions and cellular processes that occur in the living cell, such as gene expression and regulation, enzymatic transformations and regulation, metabolic and signal transduction, pathways, transport processes and so on.

The aMAZE data model has been designed to represent these complex and heterogeneous networks in an integrated way. One of our design constraints has been to reflect the biologist view as closely as possible. We therefore chose to design our model at a high conceptual level, adopting the Generalized Entity-Relationship (GER) modeling paradigm. The aMAZE workbench automatically generates a relational model from the conceptual model and implements it in a relational DBMS of one's choice. The workbench also enables querying the database and manipulating the retrieved data at the conceptual level.

The PhysicalEntities are the usual building blocks of this kind of data model: coumpounds, polypeptides, genes. But we also consider the reactions, expressions and transcriptional regulations as building blocks; they are defined as InteractionEvidences. All these building blocks constitute the biological/biochemical knowledge layer of the aMAZE model.

The heart of the aMAZE data model is the BioEvent that describes a simple biological event. A BioEvent involves BioParticipants through some roles: input, output, director (catalyst) and modulator (effector); it is documented by one or more InteractionEvidences. A BioParticipant represents a PhysicalEntity with an optional state and cellular location. This model is very general and allows to fit very different types of events - genetic, metabolic, signaling. Since BioParticipants are linked to their corresponding physical entity components, it is easy to determine the different BioEvents in which a specific physical entity is involved. A BioParticipant may be shared by different BioEvents that are not directly related; in such a case, the BioParticipant represents a pool of the corresponding PhysicalEntity at the specific location. The BioEvent/BioParticipant graph is taken to represent the corresponding biological networks and represents the systemic knowledge layer of our model.

The last layer of our model corresponds to the Processes. A Process is a subgraph of the BioEvent/BioParticipant graph. We use an intermediate entity, ProcessStep, to represent each BioEvent in the context of the described Process. While the data in the BioEntity/BioParticipant graph are tightly integrated, in contrast, Processes are less constrained. The model allows to annotate different versions of the same biological pathway, stores partially overlapping pathways, … This functional layer corresponds to the process or pathway knowledge.

The aMAZE data model is implemented in the aMAZE database (http://www.amaze.ulb.ac.be). The database offers several query interfaces: the LightBench allows to address simple queries via a web browser. The WorkBench provides a flexible interface to design complex queries by traversing relationships between objects. A java API is also available as a library, that can be integrated into custom applications. Finally, queries can also be addressed via the low-level SQL interface.

Several features of our model have been implemented in the BioPAX standard, and we are now working closely with the BioPAX team in order to extend this standard.

The power of the data model will be illustrated by examples covering the different types of cellular networks.

chris@amaze.ulb.ac.be Christian Lemer

Query-driven biclustering for microarray data analysis by Gibbs sampling

Qizheng Sheng, Karen Lemmens, Kathleen Marchal, Bart De Moor, and Yves Moreau
Department of Electrical Engineering (SCD), Katholieke Universiteit Leuven

Probabilistic graphical models (e.g., clustering by a mixture of Gaussians) have become a popular choice for modeling microarray data because of their capacity of handling the high level of noise of microarrays in a principled way. However, the the probability distribution provided by the graphical model usually contains many modes, because of the complexity of the underlying biological process. In clustering, these modes correspond to the different clusters that can be identified in the data. The dissection of the probability distribution into clusters is a difficult and often arbitrary process. Moreover, the largest clusters, which are the easiest to identify, are often not to the most interesting to the biologist because they correspond to well-known generic biological functions - where few novel findings are to be expected. This lack of sharpness of clustering algorithms has kept them into a vague exploratory role. Because for biologists, one of the main questions is always “what are the genes that are related to a particular function (or in a specific pathway) of interest to me??

Bayesian probabilistic models have shown promises in providing answers to this type of question, by transforming the existing knowledge of biologists into the prior probabilities that are incorporated into the model. Because in Bayesian models, the likelihood of the data is multiplied by the prior to deliver the posterior probabilistic model, a proper prior can substantially raise the mode that is most relevant to the biological question in the posterior model.

Coming back to the clustering problem for microarray data, and supposing that the biologists have at hand a specific set of genes (called the “seed genes? hereafter) they know to be related to some common biological function, the question for their query to the microarray data is “which other genes in this data set share similar expression profiles as the seed genes and thus might be involved in the same function??

We generalize this question further by considering which data set could be considered to answer such a question. Until recently, clustering would be performed on the array data from a single study addressing a limited biological situation. Yet, in the last few years, large data sets (called microarray compendia) consisting of multiple biological conditions or data from multiple studies have demonstrated their effectiveness as the basis of guilt-by-association studies. So in a complex compendium, where it may not be clear which microarray conditions are truly most relevant to the biological question at hand, the question becomes “which genes are functionally related to the seed genes, and in the meantime, in which experimental conditions is this biological function involved?? Otherwise stated, given the seed genes, we want to recruit genes (presented in the microarray data set) that share similar expression profiles under a subset of conditions. This is what we call the “query-driven biclustering? problem.

Another similar problem also exists for the other orientation of microarray data. For example, given a set of patients who share a certain pathological similarity (i.e., given the seed experiments), the question is to recruit other patients of the same type, and in the meantime, to identify the genes that provide a footprint that characterize those patients. Hereafter, we will refer to the former problem as the query-driven biclustering of genes, and the latter one as the query-driven biclustering of experiments.

Biclustering techniques for microarray data are receiving increasing attention in bioinformatics. Unlike conventional clustering algorithms, the biclustering algorithms aim at finding genes that show consistent behavior only under a subset of experiments. The discovery of such relationship between the genes and the conditions provides crucial information for unveiling genetic pathways. However, most of the existing biclustering algorithms focus on revealing the global pattern of the data instead of being motivated by a specific biological query.

In this talk, we will illustrate the ability of Bayesian models (for microarray data) in addressing the query-driven biclustering problems. Gibbs sampling is used for the parameterizations of the Bayesian model of the microarray data. The query-driven biclustering problem is tackled by applying a strong prior, which is constructed based on the expression profiles of the seed genes (or the expression data under the seed experiments). We describe two Gibbs sampling algorithms for query-driven biclustering: a version for discretized data using a multinomial likelihood and a Dirichlet prior, and a version for continuous data using a Gaussian likelihood and Gaussian-Wishart prior.

qizheng.sheng@esat.kuleuven.ac.be Qizheng Sheng

A comparison of the human and mouse transcriptome map: are RIDGES conserved?

Ramin Monajemi (1), Marcel van Batenburg (1), Sander van Hooff (1), Jan Koster (2), Harmen Bussemaker (3), Rogier Versteeg (2), Antoine van Kampen (1)
(1) Bioinformatics Laboratory, Academic Medical Center, Amsterdam, (2) Department of Human Genetics, AMC, Amsterdam, (3) Bioinformatics, Columbia University, New York

Transcriptome maps link gene expression profiles to the genomic DNA sequence and thereby reveal expression activity along the genome. As part of ongoing research we have established transcriptome maps for human (Caron et al, 2001; Versteeg et al, 2003), mouse (Monajemi et al, in prep). Other groups have established transcriptome maps for, for example, arabidopsis, yeast and drosophila. Our human and mouse maps revealed pronounced gene expression domains with significant increased or reduced expression as compared to the average genome. These domains were named RIDGES (Regions of Increased Gene Expression) and anti-RIDGES respectively.
Experimental research on the role of nuclear compartmentalization in the regulation of gene expression increasingly includes the RIDGE concept. For example, Tam and coworkers (2005) studied the organization and human telomeres and specifically focussed on 4q35 which is associated with the myopathy FSHD and 17q. Their research associates RIDGES with the spatial organization of these two chromosomes. Other work (Osborne et al, 2004) hypothesized that genes from different RIDGES could be (functionally) associated on the level of nuclear transcription sites.
We have shown that correlating domains occur for gene density, intron length, GC content and SINE/LINE repeats. Furthermore, first results from a comparison of the mouse and human transcriptome maps via a synteny map and a set of orthologous genes indicate the evolutionary conservation of RIDGES (Monajemi et al, in prep). However, it remains to be shown whether the gene content of RIDGES is (fully) conserved and/or if the actual gene content is less important and RIDGES merely exist as a result of their localization in the cell nucleus and thus as a result of their position on the genome. Nevertheless, all this bioinformatics research clearly revealed a higher order structure of eukaryotic genomes that plays an important role in gene regulation.
.

References

Caron HN, van Schaik BDC, van der Mee M, van Sluis P, Hermus M-C, van Asperen R, Riggins G, Heisterkamp SH, Baas F, Boon K, Voûte PA, van Kampen AHC and Versteeg R(2001) The Human Transcriptome Map reveals a clustering of highly expressed genes in chromosomal domains, Science, 291, 1289-1292

Osborne CS, Chakalova L, Brown KE, Carter D, Horton A, Debrand E, Goyenechea,B, Mitchell JA, Lopes S, Reik W, Fraser P. (2004) Active genes dynamically colocalize to shared sites of ongoing transcription. Nature Genetics, 36(10), 1065.

Tam R, Smith KP, Lawrence JB (2004) The 4q subtelomere harboring the FSHD locus is specifically anchored with peripheral heterochromatin unlike most human telomeres. The Journal of Cell Biology, 167(2), 269.

Versteeg R, van Schaik BDC, van Batenburg MF, Roos M, Monajemi R, Caron HN, Bussemaker HJ, van Kampen AHC (2003) The Human Transcriptome Map reveals extremes in gene density, intron length, GC content and repeat pattern for domains of highly and weakly expressed genes. Genome Research, 13, 1998-2004.

a.h.vankampen@amc.uva.nl Antoine van Kampen

Local co-expression domains in the genome of Arabidopsis thaliana

Xin-Ying Ren, Mark W.E.J. Fiers, Willem J. Stiekema, Jan-Peter Nap
Laboratory of Molecular Biology, Plant Sciences Group, Wageningen University and Research Center (WUR) & Applied Bioinformatics, Plant Research International, WUR

ABSTRACT

Expression of genes in eukaryotic genomes is known to cluster, but cluster size is generally loosely defined and highly variable. We have here taken a very strict definition of 'cluster' as sets of physically adjacent genes that are highly co-expressed and form so-called local co-expression domains. The Arabidopsis thaliana genome (TIGR5) was analyzed for the presence of such local co-expression domains to elucidate its functional characteristics. We used expression data sets that cover different experimental conditions, organs, tissues and cells from the MPSS repository and micro array data (Affymetrix) from a detailed root analysis. With these expression data, we identified 828 (MPSS) and 1598 (micro array) local co-expression domains consisting of 2 to 4 genes with a pair-wise Pearson's correlation coefficient (R) larger than 0.7. This number is about 1 to 5-fold higher than the numbers expected by chance. A small (5-10%) yet significant fraction of genes in the Arabidopsis genome is therefore organized into local co-expression domains. These local co-expression domains were distributed over the genome. Genes in such local domains were for the major part not categorized in the same functional category (GOslim). Neither tandemly duplicated genes, nor shared promoter sequence, or gene distance explained the occurrence of co-expression of genes in such chromosomal domains. This indicates that other parameters in genes or gene positions are important to establish co-expression in local domains of Arabidopsis chromosomes.

xinying.ren@wur.nl Xin-Ying Ren

M@CBETH: optimizing clinical microarray classification

Nathalie Pochet, Frizo Janssens, Frank De Smet, Kathleen Marchal, Johan Suykens and Bart De Moor
K.U.Leuven, ESAT-SCD (BioI), Kasteelpark Arenberg 10, B-3001 Leuven (Heverlee), Belgium

Microarray classification can be useful to support clinical management decisions for individual patients in for example oncology. However, comparing classifiers and selecting the best for each microarray dataset can be a tedious and nonstraightforward task. The M@CBETH (a MicroArray Classification BEnchmarking Tool on a Host server) web service offers the microarray community a simple tool for making optimal two-class predictions in a clinical setting [1]. M@CBETH aims at finding the best prediction among different classification methods by using randomizations of the benchmarking dataset. The web service is available at http://www.esat.kuleuven.ac.be/MACBETH/.

Using microarray data allows making predictions on for example therapy response, prognosis and metastatic phenotype of an individual patient. Microarray technology has shown to be useful in supporting clinical management decisions for individual patients in combination with classification methods. Finding the best classifier for each dataset can be a tedious and nonstraightforward task for users not familiar with these classification techniques. We present a web service that compares, for each microarray dataset introduced to this service, different classifiers and selects the best according in terms of randomized independent test set performances.

Systematic benchmarking of microarray data classification revealed that either regularization or dimensionality reduction is required to obtain good independent test set performances [2]. Regularization - as is performed in Support Vector Machines (SVM) - already led to the Gist web service, which offers SVM classification on the web. Our web service allows comparing different classification methods. By exploring different combinations of nonlinearity and dimensionality reduction, our benchmarking study showed that the most optimal classifier can differ for each dataset. Also important, but often underestimated in the model building process, is the fine-tuning of all hyperparameters (e.g. regularization parameter, kernel parameter, number of principal components). Exploring all combinations to find the most optimal classifier for each dataset can be complicated.

The M@CBETH website offers two services: benchmarking and prediction. After registration and logging on to the web service, users can request benchmarking or prediction analyses. Users are notified by email about the status of their analyses running on the host server. They can also check this on the analysis results page, which gives an overview of all analyses and contains links to corresponding results pages.

Benchmarking, the main service on the M@CBETH website, involves selection and training of an optimal model based on the submitted benchmarking dataset and their corresponding class labels. This model is then stored for immediate or later use onto prospective data. Benchmarking results in a table showing summary statistics for all selected classification methods, highlighting the best method. Prospective data can be submitted and evaluated immediately during the same benchmarking analysis. Via the prediction service, the M@CBETH website offers a way for later evaluation of prospective data by reusing an existing optimal prediction model. For both services, if the corresponding prospective labels are submitted, the prospective accuracy is calculated. Otherwise, labels are predicted for all prospective samples. This latter application is useful for classifying new unseen patients in clinical practice. Users can select the classification methods that will be compared, change the number of randomizations and switch off normalization. Issues concerning data format are discussed in the tutorial on the website.

The algorithm behind this web service uses different classification methods - based on Least Squares SVM (LS-SVM) (based on linear and Radial Basis Function (RBF) kernels), Fisher Discriminant Analysis (FDA), Principal Component Analysis (PCA) and kernel PCA (based on linear and RBF kernels) - are considered. More detailed information on these methods can be found in Pochet et al. [2]. The benchmarking dataset is reshuffled until the number of requested randomizations is reached. All randomizations are split (2/3 of the samples for training, the rest as test set) in a stratified way (class labels are equally distributed over the training-test split). Iteratively, all selected classification methods are applied to all randomizations. In each iteration, selection of the hyperparameters is first performed by means of leave-one-out cross-validation (LOO-CV), then the model is trained based on the training set and finally this model is then applied onto the test set resulting in a test set accuracy (ACC). The mean randomized test set ACC is calculated for each classification method. The best generalizing method - with best test set ACC - is then used for building the optimal classifier onto the complete benchmarking dataset, which is stored for application onto prospective datasets.

[1] Pochet,N.L.M.M., Janssens,F.A.L., De Smet,F., Marchal,K., Suykens,J.A.K. and De Moor,B.L.R. (2005) M@CBETH: a microarray classification benchmarking tool, Applications note submitted to Bioinformatics.
[2] Pochet,N., De Smet,F., Suykens,J.A.K. and De Moor,B.L.R. (2004) Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction, Bioinformatics, 20,3185-3195.

nathalie.pochet@esat.kuleuven.ac.be Nathalie Pochet

SIMAGE: Simulation of DNA MicroArray Gene Expression data

Casper J. Albers2, Ritsert C. Jansen2, Jan Kok1, Oscar P. Kuipers1, and Sacha A.F.T. van Hijum1
1. Department of Molecular Genetics; 2. Groningen Bioinformatics Centre; University of Groningen, 9751 NN Haren, The Netherlands

Simulation of DNA microarray data allows a researcher to avail of as many datasets as required to compare e.g. statistical methods for differential expression. In experimentally obtained DNA microarray data the actual (differential) gene expressions are either unknown or only estimated. Simulated DNA microarray data contain genes with (differential) expressions known beforehand. The latter makes simulated DNA microarray data a valuable addition to experimentally obtained DNA microarray data for the validation of new DNA microarray data analysis methods. We introduce a sophisticated and realistic model implemented in a free to use web-based computer program to simulate large amounts of DNA microarray data.

s.a.f.t.van.hijum@rug.nl Sacha van Hijum

Normalization of cDNA microarrays using external control spikes

Kristof Engelen, Bart Naudts, Koen Van Leemput, Bart De Moor and Kathleen Marchal
BIO@SCD, ESAT, KULeuven

Normalization of microarray measurements is the first step in a microarray analysis flow. It aims at removing consistent sources of variations to make measurements mutually comparable. Reliable normalization is essential since the results of all subsequent analyses (such as e.g. clustering) might largely be influenced by the normalization procedure. For normalization of cDNA different methods have been described. Although some approaches inherently work with absolute intensities (e.g. ANOVA[1,2]), in general, preprocessing, of cDNA microarrays largely depends on the calculation of the log-ratios of the measured intensities. A common normalization step consists of the linearization of the Cy3 vs. Cy5 intensities (e.g. loess[3]). It assumes the distribution of gene expression is balanced and shows little change between the biological samples tested (Global Normalization Assumption). Global mRNA changes that result in an uneven distribution of expression changes however, have been shown to occur more frequently than what is currently believed[4,5], and could have a significant impact on the interpretation of data normalized according to the Global Normalization Assumption.
Therefore, in this study we describe a different way of normalizing cDNA microarray data. In contrast to previous approaches, our methodology is based on a physically motivated model, consisting of two major components. We explicitly model the hybridization of mRNA transcripts to their corresponding cDNA probes and the relation between the measured fluorescence and the amount of hybridized, labeled mRNA. The parameters of this model and the incorporated error distributions are estimated from external control spikes: mRNA transcripts that are added to the hybridization solution in known concentrations. Using a publicly available data set, we show that our procedure, due to the inherent nonlinearity of the model, is capable of adequately linearizing the data, without making any assumptions on the distribution of gene expression (as opposed to the Global Normalization Assumption). More importantly, since our model links mRNA concentration to measured intensity, we are also able to estimate the absolute concentrations of mRNA transcripts in the hybridization solution with high accuracy.

1. Kerr MK, Martin M, en Churchill GA 2000. Analysis of variance for gene expression microarray data. J. Comput. Biol. 7: 819-837.
2. Jin W, Riley RM, Wolfinger RD, White KP, Passador Gurgel G en Gibson G 2001. The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. Nat Genet 29: 389-395.
3. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J en Speed TP. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 30: e15.
4. van de Peppel J, Kemmeren P, van Bakel H, Radonjic M, van Leenen D, Holstege FC. 2003 Monitoring global messenger RNA changes in externally controlled microarray experiments. EMBO Rep. 4(4):387-393.
5. van Bakel H, Holstege FC. 2004 In control: systematic assessment of microarray performance. EMBO Rep. 5(10):964-969.

kristof.engelen@esat.kuleuven.ac.be Kristof Engelen

A taxonomy-traversing approach to discover cis-acting elements in prokaryotes

Rekin's Janky and Jacques van Helden
Service de Conformation de Macromolécules Biologiques et de Bioinformatique (SCMBB) - Université Libre de Bruxelles (ULB) - Boulevard du Triomphe - CP 263 - 1050 Bruxelles (BELGIUM)

BACKGROUND
One of the major challenges of current genomics is to decipher the mechanisms of regulation of gene expression. Transcriptional regulation is mediated by interactions between transcription factors (TF) and specific cis-acting elements. Several pattern discovery algorithms have been developed to discover putative regulatory motifs in sets of co-regulated genes from the same genome. But it can be a hard and long process to get those groups of genes as it depends most of the time on experimental analysis (e.g. microarrays).
A complementary approach is the phylogenetic footprinting which takes advantage of the increasing number of sequenced genomes to detect phylogenetically conserved cis-acting elements in upstream sequences of clusters of orthologous genes (COGs). This approach is based on the hypothesis that, due to selective pressure, regulatory elements tend to evolve at a slower rate than surrounding non-coding sequences. Blanchette and Tompa (2000-2003) developed a dedicated algorithm, Footprinter, which takes a set of upstream sequences as input and a corresponding binary tree, and searches for conserved elements in each branch of the tree. The program gave impressive results for some examples from higher organisms. However, motifs are scored according to parsimony criteria, and implicitly relies on an assumption of equiprobable and independent residues, which is not optimal for pattern discovery in microbial genomes as determined by our tests on yeast and bacteria.
Our strategy is to use the phylogenetic footprinting in order to detect such phylogenetic footprints, by looking for conserved elements in fully sequenced prokaryote genomes.
MATERIAL AND METHODS
Building upon the Footprinter principle (traversing the whole taxonomic tree), we applied to each branch a pattern-discovery method based on the statistical detection of over-represented dyads, using dyad-analysis (van Helden, 2000 and 2003). This algorithm uses a simple background probabilistic model estimating the dyad probabilities, as the product of frequencies of the two corresponding trinucleotides. We have tested two background models : “MONAD?, by calculating the expected frequencies from the input sequences, and “EXPFREQ?, from a taxon-specific reference set.
In addition, we combined our prediction of transcription units in all bacteria (not published), according to Salgado-Moreno’s method (Salgado, 2000 ; Moreno-Hagelsieb and Collado-Vides, 2002), in order to get the upstream sequence of the operon leader gene when the gene of interest is predicted to be within the operon. This would optimise the pattern discovery as we don’t expect regulatory motifs to be in such short intergenic distances (<50bp) and within a transcription unit.
A new algorithm footprint-analysis was implemented to manage these tasks at each level of a taxonomic tree.
For illustration of the potential of this algorithm, we focused on the gene lexA, a well-characterized bacterial gene. This gene, coding for a repressor protein, LexA, involved in the SOS response of DNA-damage, is highly conserved across bacteria. In the upstream region of this gene, we can identify the binding site of the TF LexA as its gene is self-regulated. In our results, the LexA binding site, also called SOS box (CTGTn8ACAG) (Erill, 2004; Salgado, 2004), is detected at all levels of the taxonomy from Enterobacteria and above. The maximal significance is obtained at the level of Gammaproteobacteria, with the model EXPFREQ. As a general rule, the MONAD model is more stringent, with the advantage of being less noisy, but this is at the cost of sensitivity for detecting less conserved motifs. The analysis of the results for the whole tree of lexA orthologs can reveal divergences of this SOS box. Indeed, a distinct motif consensus (AACATATGTTCT) is detected for the Firmicutes, which is compatible with the well known consensus (CGAACRNRYGTTYC) for LexA binding site specific for Gram positive bacteria, for which Firmicutes are the best representants. Additionally, we also found similarities for another known target of LexA, the gene recA, as we can deduce a consensus CGAACANNNNTTCG for the Firmicutes from the high scored motifs.
A systematic analysis with other genes is under progress to extract general trends. We are currently performing a systematic evaluation of the algorithm by applying it to all bacterial genes, and comparing motifs discovered in E.coli ancestor taxons with those annotated in RegulonDB (Huerta, 1998; Salgado, 2004). This quantitative evaluation will allow us to assess the overall performance of the program, and identify the strengths and limitations of the approach.
CONCLUSIONS
The application of pattern discovery to predict cis-acting elements at different taxonomical levels will hopefully allow us to infer evolutionary events related to transcriptional regulation, such as motif divergence, conservation, appearance,… In the future, this approach will permit us to identify groups of co-regulated genes to open the way for a better understanding of the evolution of transcriptional regulatory networks.

rekins@scmbb.ulb.ac.be Rekin's Janky

A novel approach to identify regulatory motifs in distantly related genomes

Ruth Van Hellemont [1], Pieter Monsieurs [1], Gert Thijs [1], Bart De Moor [1], Yves Van de Peer [3] and Kathleen Marchal [2]
[1] ESAT-SCD, K.U. Leuven, Kasteelpark Arenberg 10, 3001 Leuven-Heverlee, Belgium; [2] Centre of Microbial and Plant Genetics, K.U. Leuven, Kasteelpark Arenberg 20, 3001 Leuven-Heverlee, Belgium; [3] Plant Systems Biology. Bioinformatics and Evolutionary Genomics, VIB / Ghent University, Technologiepark 927, 9052 Gent, Belgium

Phylogenetic footprinting, where regulatory motifs are identified by comparing the intergenics of different genome sequences, has proven successful in the identification of regulatory motifs in higher eukaryotes, in particular yeasts and vertebrates. Nevertheless, there are some important shortcomings to the current methodologies. For instance, global sequence alignment programs used to uncovering conserved non-coding sequences generally have difficulties aligning sequences of heterogeneous length, in particular when the overall similarity between the sequences is low. Motif detection methods and local alignment algorithms provide a solution to this problem but are sensitive to low signal-to-noise ratio's, i.e. the presence of short motifs (5-8 bp) in long intergenic sequences. To assess these difficulties, most apparent when applying phylogenetic footprinting to distantly related organisms, we developed a two-step procedure that combines both approaches. First, only the most conserved parts of genomes of closely related organisms are selected. Subsequently, these are combined with the intergenic region of a more distantly related genome sequence and subjected to the second step where conserved elements or 'blocks' are identified using a novel Gibbs sampling based algorithm, BlockSampler. To validate our newly developed strategy we analyzed four well-studied benchmark data sets: cfos, hoxb2, pax6 and scl. In contrast to two other algorithms, commonly used for phylogenetic footprinting our method detected most of the previously described motifs, and in addition detected many additional conserved elements.

ruth.vanhellemont@esat.kuleuven.ac.be Ruth Van Hellemont

Physical stability of coding and non-coding DNA regions

Tom Michoel, Yves van de Peer
Bioinformatics and Evolutionary Genomics, Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

The computation of the thermal stability and statistical physics of nucleic acids is a classical problem going back to the 1960's. Originally, the development of theoretical models to describe the untwisting and separation of both strands of the DNA double-helix was motivated by the question to understand physical experiments on DNA denaturation, but in recent years it has become clear that these models can also be used to analyse the genomic content of DNA sequences. In several instances (Benham 1993, 1996; Yeramian 2000), genes can be identified as physically stable regions, and promotors as physically unstable, although the correlation between the genetic and the physical stability map varies within a DNA sequence. To explain this variation, Yeramian (2000) has suggested that the role of physics in delimiting coding regions is in the evolutionary process of being erased. To push these ideas further, we have developed a new method to compute DNA thermal stability properties which is both very simple to implement and very fast to execute, allowing the analysis of whole genome sequences. In this talk we will review previous results on the physics of DNA, discuss computational methods for whole genome sequences, and present a tentative outlook on future bioinformatics applications.

tom.michoel@psb.ugent.be Tom Michoel

CGHGate: Array-CGH and Human Genome Annotation

Steven Van Vooren (1), Nicole Maas (2), Joris Robert Vermeesch (2), Yves Moreau (1), Bart De Moor(1)
(1) ESAT/SISTA/BIOI, Faculty of Engineering, Katholieke Universiteit Leuven, Leuven, Belgium; (2) Center for Human Genetics, Faculty of Medicine, Katholieke Universiteit Leuven, Leuven, Belgium

Background
As Microarray-CGH is introduced into the clinical practice for the identification of submicroscopic deletions and amplifications in a genome, means and methods to handle the growing amount of related data in a structured way become essential for clinical geneticists and are of interest to researchers implied in developmental biology. We present CGHGate, a web based application that combines a constitutional cytogenetics database supporting informative and complete case descriptions and tools for search, visualisation, genome annotation, and data/text-mining geared at microarray-CGH related data.

Methods
First of all, CGHGate features standardized filing of CGH-microarray case reports in a constitutional cytogenetics database. In order to ensure high quality of clinical reports on microdeletions and duplications, both clinical and cytogenetic data need to be sufficiently informative and correct. This requires a high resolution delineation of the chromosomal aberration and a detailed clinical description. To this end, we define a reporting standard that supports detailed clinical descriptions including biometry, growth and development, and malformations. It can be linked to medical ontologies and controlled vocabularies such as LDDB, MeSH and SNOMED, incorporating free-text descriptions and keywords as well.
Secondly, CGHGate allows researchers to visualise patient and annotation data on a genomic region of interest within the tool, but also via a DAS Server (Distributed Annotation Server), which allows the cytogenetic database to be linked the Ensembl genome browser or a DAS compliant viewer.
Thridly, the tool features text- and data mining methods. These allow retrieval of case reports describing "similar? patients with a corresponding "phenotype profile? or "genotype profile?. It constructs phenotypic profiles from free text linked to patients, using ontologies. Additionally, it automatically annotates patient reports with biomedical literature (linking them to PubMed abstracts). All text mining features are based on the vector space model and IDF (Inverse Document Frequency) indices of case reports and medline abstracts, using tailored controlled vocabularies that are domain specific, ie. gene, disease and dysmorphology centric. On top of this model, techniques such as Latent Semantic Indexing and Random Indexing are used to increase retrieval performance.
CGHGate can identify candidate genes for the phenotypic features linked to a group of cases of interest. The web application performs regular search tasks as well, such as querying for specific phenotypic features, free-text and ontology keywords, genes, clones and genomic regions.

Results
We present a cytogenetic database for constitutional cytogenetics that features web based data mining and visualisation tools. Although CGHGate can be deployed as a local installation at a cytogenetics center, it is aimed at making array CGH case reports and clinical features openly available to the community to advance research in the field in an effort to share case information among multiple research groups.

Conclusion
CGHGate aims not only at prognosis and care but also contributes to the annotation of the human genome. The generation of a phenotypic genome map combined with text mining features will enable the identification of genes involved in developmental processes and will delineate novel clinically recognisable entities. The database will be publicly available and linked to Ensembl via DAS, combining microdeletion and duplication case reports with a search and mining tool aimed at identifying clinical and phenotypic features.

Acknowledgements
This research was supported by grants from the Research Council K.U. Leuven (GOA-Mefisto-666, GOA-Ambiorics, IDO), the Fonds voor Wetenschappelijk Onderzoek - Vlaanderen (G.0115.01, G.0240.99, G.0407.02, G.0413.03, G.0388.03, G.0229.03, G.0241.04), the Instituut voor de aanmoediging van Innovatie door Wetenschap en Technologie Vlaanderen (STWW-Genprom, GBOU-McKnow, GBOU-SQUAD, GBOU-ANA), the Belgian Federal Science Policy Office (IUAP V-22), and the European Union (FP5 CAGE, ERNSI, FP6 NoE Biopattern, NoE E-tumours).

Steven.VanVooren@esat.kuleuven.ac.be Steven Van Vooren

Functional divergence of proteins through frameshift mutations

Jeroen Raes and Yves Van de Peer
Bioinformatics and Evolutionary Genomics, Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

Frameshift mutations are generally considered to be deleterious and of little importance for the evolution of novel gene functions. However, by screening an exhaustive set of vertebrate gene families, we found that when a second transcript coding for the original gene product can compensate for this mutation, frame shift mutations can be retained for tens to even hundreds of millions of years, and allow for new gene functions to be acquired. Furthermore, our results indicate that in particular transcription factors and transmembrane proteins are prone to this type of mutations, probably because of their specific structural organization of functional domains.

jerae@psb.ugent.be Jeroen Raes

A Novel Database System For Easy Correlation Of Heterogeneous Data Of One Protein Superfamily Using a Structure Based Multiple Sequence Alignment

Henk-Jan Joosten, Simon Folkertsma, Frank van Zimmeren, Remko Kuipers, Erik Ittmann, Peter Schaap and Gert Vriend
WUR (Wageningen University and Research center)

A major difficulty for scientists that want to gain new biological insights about a protein (family) is correlating the sheer amount of data available on the Internet and in the literature. The data is often heterogeneous and divided over multiple databases, making it difficult and time consuming to make correlations. Moreover, protein sequences from different databases or homologues sequences from different organisms have inconsistent amino acid numbering. To cope with these problems, molecular class-specific databases are needed, that organize data around a single class of molecules using a common numbering scheme for structurally equivalent amino acids. Here we describe a system that can automatically build databases in which every amino acid is individually stored and connected via a structure based multiple sequence alignment (3D-MSA). The system starts with building a large accurate 3D-MSA of a class of proteins (superfamily). Such alignments can be built by making separate subfamily alignments starting from structures belonging to different subfamilies. These separate subfamily alignments can be aligned to each other by determining the structurally equivalent amino acid positions using a superpositioning of the starting structures. In the first step structure files are automatically superimposed and a common core of structurally conserved amino acid positions is defined to which a 3D numbering scheme is applied. Then profiles derived from the starting structures, with high gap-penalty scores at structurally conserved positions, are used to build accurate subfamily multiple sequence alignments. Every amino acid from the resulting 3D-MSA is stored in the database together with its corresponding 3D number. Storing separate amino acids with their 3D number has a number of advantages. Firstly, it solves the problem of inconsistent numbering. The numbering used in the database of origin is also stored, enabling easy navigation between the different numbering schemes. Secondly, it enables storage of amino acid associated data, like mutational information, ligand- and substrate contacts, protein interaction sites, phosphorylation sites, hydrogen-bonding status, solvent accessibility, etc. in relation to the amino acids. Thirdly, evolutionary information, such as correlated mutations or conservation of amino acid positions, intrinsically present in the alignment is directly coupled to all other amino acid related information stored in the database. In addition, all this information can easily be mapped in a 3D environment because the system automatically makes homology models of all sequences in the alignment and it also incorporates the 3D numbering scheme in all models as well as in all starting structures. The alignment and all associated data are visualized and can be retrieved via interactive HTML pages. By clicking on the amino acid in the alignment the user can retrieve all data related to that specific amino acid.

henk-jan.joosten@wur.nl Henk-Jan Joosten

Align-m 2: multiple alignment with focus on specificity rather than sensitivity

Ivo Van Walle, Ignace Lasters, Lode Wyns
AlgoNomics NV, Technologiepark 4, 9052 Gent; Vrije Universiteit Brussel, Department of Ultrastructure, Oefenplein 1, 1050 Brussel

Alignments of sequences with low similarity generally have low accuracy as well. This accuracy can be measured in terms of correctly aligned residues (sensitivity), but also, and frequently forgotten, in terms of incorrectly aligned residues or "over-alignment" (specificity).We present here an extension of our Align-m program, called P2M, which generates highly specific multiple alignments from pairwise data, using a truly multiple approach.

Align-m was compared to ClustalW, T-Coffee, ProbCons, DiAlign, Muscle, Mafft and Poa on the SABmark reference alignment database, which covers the entire SCOP/PDB. With respect to sensitivity, Align-m performed similarly to the other programs. Specificity however was much higher, with 64% of the aligned residues being correct, compared to only 36% on average for the other programs. A further improvement to 83%, though with a corresponding drop in sensitivity, was possible by providing entirely independently generated pairwise alignments to P2M. The potential advantage of multiple over pairwise alignment therefore seems to lie primarily in a drastic reduction of incorrect residue pairs, up to a point where specifity becomes reliable, rather than in increased sensitivity.

Download: http://bioinformatics.vub.ac.be

Contact: ivo.van.walle@algonomics.com

ivo.van.walle@algonomics.com Ivo Van Walle

Prediction of Functional Sites in Proteins using the Relationship between Stability and Function.

Benoit H Dessailly, Marc F Lensink, Shoshana J Wodak
Service de Conformation des Macromolécules Biologiques et de Bioinformatique, Université Libre de Bruxelles, 1050-Ixelles

Being able to predict active sites, and therefore key functional residues in proteins of known structure, has outstanding interest in such applications as drug-design. This is especially relevant in the context of structural genomics projects which aim at solving structures of a great number of proteins, including many for which no functional information is known beforehand.

It was shown experimentally that functionally important residues often have an unfavorable contribution to the stability of proteins. For example, mutation of functionally important residues in the active-site of barnase almost systematically yields more stable but inactive variants of the enzyme. Comparable analyses have been conducted on a wide variety of biological systems leading to similar observations. A precursory computational analysis previously indicated that this property might be useful for predicting functionally important residues in proteins of known three-dimensional structure.

Yet, however important the extent of this stability-function relationship is, it is expected that many non-functionally related residues will also have a destabilizing effect. To overcome this problem, we developed an original method to predict functional sites in proteins, based on the detection of groups of closeby destabilizing residues, as such clusters are less likely to occur by chance than singly located destabilizing residues. The destabilizing effect of all residues is computed using an all-atom forcefield taking into account non-bonded interactions and an empirical solvation term. Groups of spatially contiguous destabilizing residues are detected in the three-dimensional structure using a hierarchical clustering approach.

The method has been parameterized on a set of very high quality structures, and validation is carried out in a systematic manner on a large set of good quality structures containing no close homologues. An in-depth analysis of the results is presented for a set of proteins representing a quite diverse panel of functions. We show that correct predictions are obtained for most systems analysed. Additionally, we describe automatic ways to detect probable false positives.

benoit@scmbb.ulb.ac.be Benoit H. Dessailly

Data-driven docking for the study of biomolecular complexes: combining WHISCY with HADDOCK.

A.M.J.J. Bonvin, A.D.J. van Dijk, S.J. de Vries
Bijvoet Center for Biomolecular Research, Utrecht University, the Netherlands

Protein-protein interactions play a key role in biological processes. Identifying the interacting residues is a first step toward understanding these interactions at a structural level. We have developed an interface prediction program called WHISCY. It combines surface conservation and structural information to predict protein-protein interfaces. The accuracy of the predictions is more than three times higher than a random prediction. These predictions have been combined with another interface prediction program, ProMate [1], resulting in an even more accurate predictor. The usefulness of the predictions was tested using our data-driven docking program HADDOCK [2,3] (http://www.nmr.chem.uu.nl/haddock). HADDOCK distinguishes itself from ab-initio docking methods in the fact that it encodes information from identified or predicted protein interfaces in ambiguous interaction restraints (AIRs) to drive the docking process. Flexibility is accounted for in different ways during the docking which allows to model (small) conformational changes taking place during complex formation.
In my talk I will present the combination of WHISCY with HADDOCK together with results from our participation to the blind docking experiment CAPRI (Critical Assessment of PRedicted Interactions) (http://capri.ebi.ac.uk).

References:

1. Neuvirth H., Raz R. & Schreiber G. (2004). J Mol Biol 338, 181-199.
2. Dominguez C., Boelens R & Bonvin A.M.J.J. (2003). J Am Chem Soc 125 1731-1737.
3. van Dijk A.D.J., Boelens R. & Bonvin A.M.J.J. (2005). FEBS Journal 272 293-312.

a.m.j.j.bonvin@chem.uu.nl Alexandre M.J.J. Bonvin

Proteomic and genomic data classification using decision tree based ensemble methods - Application to the diagnosis of inflammatory diseases

Pierre Geurts (1,3), Marianne Fillet (2,3), Dominique de Seny (2,3), Marie-Alice Meuwis (2,3), Marie-Paule Merville (2,3), Louis Wehenkel (1,3)
(1) Service de Méthodes Stochastiques, Department of Electrical Engineering and Computer Science, University of Liège (2) Laboratoire de chimie médicale, Department of Life Sciences, University of Liège (3) CBIG - Centre of Biomedical Integrative Genoproteomics, University of Liège

Modern medical instrumentation and acquisition technologies (mass spectrometry, microarray, sequencing tools…) generate large datasets describing for example patients, animals, tissues, or cells. The analysis of such amount of data is impossible without the help of efficient computer based tools. Data Mining refers to the application of machine learning and visualization techniques in order to help a human expert to extract potentially interesting and synthetic knowledge from these large volumes of raw data. Potential medical applications are the automatic design of diagnostic or prognostic tools for a given disease or the identification of potential biomarkers for this disease.

Recent developments in machine learning allow one to exploit datasets characterized by small numbers of very high-dimensional samples, without prior feature selection or extraction. We propose a systematic approach for mining proteomic and genomic data based on decision tree ensemble methods. The overall objective of the proposed software is to use experimental datasets in order to identify one or several biomarkers specific of a given disease, or able to discriminate among a certain class of diseases, or indicative of treatment response, and to construct predictive models exploiting the biomarker intensities to help physicians in the context of medical diagnosis or prognosis. It includes clearly defined pre- and post-processing steps as well as the invocation of a toolbox of generic decision tree based methods (Bagging, Random Forests, Extra-Trees, Boosting). We choose decision tree methods because these methods are computationally very efficient and can be easily exploited to assist physicians in identifying among a large number of candidate biomarkers those that are best suited for a particular discrimination task.

We provide results of our approach on datasets of two experimental studies based on surface-enhanced laser desorption/ionization time of flight mass spectrometry (SELDI-TOF-MS). They concern the diagnosis of patients suffering from inflammatory diseases, namely rheumatoid arthritis and inflammatory bowel diseases. The motivation of the first study is to complement existing diagnostic tests for rheumatoid arthritis in order to allow an earlier diagnosis of this disease. The goal of the second study is to provide a robust and easy-to-use diagnostic tool much less invasive than current diagnostic tests including endoscopy and histology.

Each database collects mass spectra obtained from sera of healthy and disease patients using different chip arrays including strong anion-exchange (Q10), weak cation-exchange (CM10), or hydrophobic (H4) surface. The same framework was applied to the two problems and has given very promising results both for the induction of predictive models and for the identification of biomarkers. In both cases, we obtained predictive models which are more than 80 % specific and sensitive[1]. Sensitivity and specificity were further increased to more than 90% by combining several MS replicas per patient. These results are superior to those obtained with standard pre- and post-processing techniques used in these applications (peak-detection and p-values based biomarker selection) and are also competitive with existing practice for the diagnosis of these diseases.

Moreover, the methodology revealed a small number of variables which are deemed sufficient by the tree based models to discriminate among classes of patients and thus constitute potential biomarkers specific to the studied diseases. The interest of these biomarkers is twofold. First, the downstream purification and identification of the proteins associated with these biomarkers could help physicians to better understand such diseases and to highlight new therapeutic targets. Second, the application of machine learning algorithms on these reduced sets of biomarkers could possibly provide more reliable models than those using the whole set of variables.

The approach being generic, it could be applied to other problems and other proteomic and/or genomic data acquisition schemes. Current work concerns its application to microarray data for the classification of brain tumors.

[1] Proteomic mass spectra classification using decision tree based ensemble methods, P. Geurts, M. Fillet, D. de Seny, M.-A. Meuwis, M.-P. Merville, L. Wehenkel. Submitted.

p.geurts@ulg.ac.be Pierre Geurts

Poster Presentations

A biologically plausible synthetic gene network generator for analyzing inference algorithms

Koen Van Leemput, Tim Van den Bulcke, Bart Naudts, Kathleen Marchal, Alain Verschoren, Bart De Moor
ESAT-SCD, K.U.Leuven, Kasteelpark Arenberg 10, B-3001 Heverlee; ISLab, Dept. Math. and Comp. Sc., University of Antwerp, Groenenborgerlaan 171, B-2020 Antwerpen; CMPG, Fac. of Applied Bioscience and Eng., K.U.Leuven, Kasteelpark Arenberg 10, B-3001 Heverlee

Ongoing technological advances have made the application of high throughput assays, such as microarrays, common practice for the biologist. One of the main challenges is the development of high-quality methods to shed light on the complex network of regulatory interactions between the various constituents of a living system.

Assessing the quality of clustering algorithms, correlation analysis or gene network inference algorithms is difficult without the ability to compare the reconstructed results with the real underlying regulatory network. Validation strategies of inference algorithms for example, are often limited to confirming previously known interactions in the reconstructed network. In such an approach, false positive interactions/edges are not penalized and the algorithm can usually only be applied to data of a single, well-known network.

There is a clear need for ways to thoroughly test learning algorithms in a fast and reproducible manner. The goal of this work is a synthetic network generator that addresses the limitations of existing network simulators. We propose a generator that is based on networks with a realistic biological topology and capable of generating biologically realistic sampled expression data.

In this work, we describe a simulator of transcriptional regulatory networks (TRNs) that generates simulated microarray data. The topology of this network is obtained by selecting a random subnetwork from a well-characterized TRN, in this application either E. coli or S. cerevisiae. In a second step, transition functions based on Michaelis-Menten and Hill kinetics are assigned to the edges in the network and biologically plausible parameter values are assigned. In a third step, simulated data is sampled from the resulting network. Both biological noise and measurement noise are included in the system. Several parameters can be adapted in order to generate various types of datasets.

Our analyses showed that biological TRNs have topological properties that are different from several types of random networks. Data simulated by our network generator exhibits a full range of expression values similar to those obtained by real microarray experiments.

koen.vanleemput@ua.ac.be Koen Van Leemput

A stepwise procedure for gene classification on large genome wide expression datasets

Eijssen LMT, Lindsey PJ, Peeters R, Westra R, Eijsden van RGE, Bolotin-Fukuhara M, Smeets HJM, Vlietinck RFM
Dept of Genetics and Cell Biology, Universiteit Maastricht

To extract functional information on genes and processes from data sets containing expression values for large numbers of genes, analysis methods are required that can both computationally deal with these amounts of data and focus on specific research questions. Therefore, a stepwise procedure that combines principal component analysis with the results of several discriminant functions, was developed in order to specifically find genes involved in processes of interest. In a data set of global expression profiles of 300 gene knock-outs in Saccharomyces cerevisiae, we selected five subgroups, based on the cellular compartment of the proteins involved. Each subgroup could be classified by the expression pattern of a limited number of genes. The genes discriminating "mitochondrion" from the other subgroups were evaluated in more detail and the thiamine pathway turned out to be one of the processes involved. Logistic models using thiamine related genes succeeded in predicting whether specific yeast knock-outs were mitochondrial or not. Our approach is effective in finding genes or pathways related to a certain biological process or function and is efficient in extracting meaningful information from large microarray experiments.

lars.eijssen@gen.unimaas.nl L.M.T. Eijssen

Ab-initio protein/DNA docking using HADDOCK

Marc van Dijk, Aalt-Jan van Dijk, Alexandre M.J.J Bonvin
Department of NMR Spectroscopy, Bijvoet Center for Biomolecular Research, Utrecht University, Bloembergen gebouw, Padualaan 8, 3584 CH Utrecht , The Netherlands

Research has progressed to the post genome age. The long and often difficult task of elucidating structure and function of the many putative proteins just started. The biological function of these proteins is determined in large extend by their interaction with other proteins and the DNA. Biochemical research methods can provide us with a wealth of information regarding these interactions. None of these methods however provide us with information on a atomic level.
3D structure determination is mainly performed by X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy. Single protein structure determination is performed readily with both techniques. Solving protein-protein or protein-DNA complexes however has its limitations. The dynamic nature of protein-protein and protein-DNA interactions hampers the production of good quality crystals and the current size limit of NMR of about 100 kDa renders the study of large complexes difficult. To compensate for the limitations of these experimental approaches computational methods have been developed. These programs use various methods for docking the individual proteins in a "ab initio" way. For protein-protein interactions the field of computational docking matured. Very few docking algorithms however have been applied to predict protein-DNA interactions.

Here we present HADDOCK (www.nmr.chem.uu.nl/haddock) [1] as a tool for protein/DNA docking. HADDOCK makes use of biochemical or biophysical information to drive the docking. Mutagenesis and NMR chemical shift data for example can be used. To limit the search trough conformational space an ensemble of different DNA conformations is used as starting material. Flexibility can be introduced during several stages of the docking.
We have tested the protocol on two well defined systems: the bacteriophage 434 Cro protein and the E. Coli Lac repressor headpiece. The best solutions for both cases were within 2.5 Å backbone RMSD of the published complexes. The predicted protein / DNA interactions show a high degree of similarity to the published interaction data.

HADDOCK allows for a rapid and accurate docking of protein / DNA complexes based on the use of biochemical or biophysical information. As such HADDOCK can be a valuable tool for the study of protein / DNA interactions in a ab-initio way.

Reference:
[1] Cyril Dominguez, Rolf Boelens and Alexandre M.J.J. Bonvin (2003).HADDOCK: a protein-protein docking approach based on biochemical and/or biophysical information. J. Am. Chem. Soc. 125, 1731-1737.

mvdijk@nmr.chem.uu.nl Marc van Dijk

Activation pathways of Rat-{Delta}{alpha}-Chymotrypsin revealed by MD and TMD methods

J. Mátrai1, M. De Maeyer1, A. Jonckheer1, E. Joris1, G. Verheyden2, P. Krüger3 and Y. Engelborghs1
PhD student

The activation of chymotypsinogen into chymotrypsin happens via the proteolytic cleavage of the R15-I16 bond and the subsequent rotation of residue I16 from the solvent into the interior of the protein [1]. As a result, a stabilizing salt bridge between the amino terminus of I16 and the side chain of D194 is formed. The transition from the inactive form with the solvent exposed I16 to the active form with the buried I16 residue can be induced in vitro by a neutral to basic pH change [2, 3]. The kinetics of the activation process can be followed by Stopped Flow Fluorescence (SFF) experiments while the structural features of the transition can be explored by in silico Molecular Dynamics (MD) and Targeted Molecular Dynamics (TMD) [4] simulations.
To get further insight into the activation process, mutants were constructed and studied by SFF measurements and MD/TMD simulations. Our results indicate the existence of parallel activation pathways. They show correlation with the experimental results in terms of activation enthalpy and reveal the advantages and disadvantages of the TMD method.

1) Wang, D.C., Bode, W., Huber, R. 1985. Bovine Chymotrypsinogen-A X-Ray Crystal-Structure Analysis and Refinement of A New Crystal Form at 1.8 A Resolution. J. Mol. Biol. 185:595-624.
2) Fersht, A.R., Requena, Y. 1971. Equilibrium and rate constants for the interconversion of two conformations of -chymotrypsin. The existence of a catalytically inactive conformation at neutral p H. J.Mol. Biol. 60:279-290.
3) Stoesz, J.D., Lumry, R.W. 1978. Refolding Transition of Alpha-Chymotrypsin - Ph and Salt Dependence. Biochemistry 17:3693-3699.
4) Wroblowski, B., Diaz, F., Schlitter, J., Engelborghs, Y. 1997. Protein Engineering 10/10, 1163-1174

janka.matrai@fys.kuleuven.ac.be Janka Mátrai

Analysis of large-scale gene duplications in fishes as model systems for vertebrate genome evolution

Tine Blomme, Klaas Vandepoele, Cedric Simillion, Yves Van de Peer
Bioinformatics and Evolutionary Genomics, Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

It has long been suggested that gene and genome duplications play important roles in the evolution of organismal complexity. For example, some evidence for whole genome duplications during the early evolution of the vertebrates has been uncovered (referred to as the 2R hypothesis) and it has been hypothesized that the extra genes created by these duplications have permitted an increase in physiological and anatomical complexity. In our group we found evidence that the ray-finned fishes underwent an additional genome duplication about 350 million years ago. Current research focuses on investigating the consequences of this duplication event in fishes by comparing the genomes of Tetraodon nigroviridis (green spotted pufferfish), Takifugu rubripes (tiger pufferfish) and Danio rerio (zebrafish). In particular, we will look for different patterns in gene loss after duplication in these species. Moreover, we will try to correlate gene retention / gene loss of gene duplicates with the biological function of genes. In addition, we will use comparative techniques (such as phylogenetic footprinting) to study functional divergence of the promoter sequences of the duplicated genes. These results will be linked to expression data. Furthermore, comparing the divergence patterns of the fish promoter sequences will bring about relevant information regarding gene regulation in other vertebrates (e.g. human).

tiblo@psb.ugent.be Tine Blomme

Analysis of shared proteins: a promising method to resolve the eukaryotic Tree of Life

Eiko E. Kuramae1, Vincent Robert1, Berend Snel2, Michael Weiß3 & Teun Boekhout1,4
1Centraalbureau voor Schimmelcultures, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands;2 Nijmegen Center for Molecular Life Sciences, University Medical Center St. Radboud, pa CMBI, Toernooiveld 1, 6526 ED, Nijmegen, The Netherlands;3 Spezielle Botanik und Mykologie, Universität Tübingen, Auf der Morgenstelle 1, D-72076 Tübingen, Germany; 4 University Medical Center, Department of Medicine, Div. Acute Medicine & Infectious Diseases, Utrecht, The Netherlands

Our understanding of the Tree of Life (TOL) is still fragmentary. Until recently, molecular phylogeneticists built trees based on ribosomal RNAs and selected protein sequences which, however, usually suffered from lack of support for the deeper branches. Now, phylogenetic hypotheses can be based on the analysis of full genomes. Here we present results from a phylogenetic analysis of concatenated sequences of orthologous genes present in the genomes of 21 fungal, 3 animal and one plant species. A total of 531 proteins occurring in all the genomes studied, were analyzed using four different phylogenetic methods. Our results agree well with current hypotheses on the phylogeny of higher fungi. However, the single tree that we inferred from our dataset shows for the first time excellent nodal support for each branch, which suggests that it reflects the true phylogenetic relationships of the species involved. Our results demonstrate that eukaryotic phylogeny may strongly benefit from the use of full genome comparisons, as we demonstrate here for the fungal kingdom.

kuramae@cbs.knaw.nl Eiko E. Kuramae

Annotation of the genome of the ectomycorrhizal basidiomycete Laccaria bicolor.

Jan Wuyts 1,2; Stephane Rombauts 1; Pierre Rouzé 1; Yves Van de Peer 1; Francis Martin 2
1) Bioinformatics and Evolutionary Genomics, Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium; 2) Unité "Interactions Arbres/Micro-Organismes", INRA-Nancy / Université Henri Poincaré, 54280 Champenoux, France

Soil microorganisms are an extremely diverse group consisting of bacteria, fungi and other eukaryotic microbes. Together they fulfill essential functions enabling forest ecosystems to sustain themselves on relatively nutrient-poor soils. This requires a highly tuned recycling of biomass as well as nitrogen fixation and mineral weathering.

One group of soil microorganisms, the ectomycorrhizal fungi, ensures the availability of minerals to trees. These species have developed a highly effective system to mobilize and absorb water and minerals: the ectomycorrhizal symbiosis. Consequently this system contributes to the biogeochemical processes that are essential to maintain sustainable forest ecosystems.

As part of the “Populus Community Sequencing” project, the Joint Genome Institute (JGI) has sequenced the genome of Laccaria bicolor by way of shotgun sequencing. We will be responsible for the structural annotation of this genome. This will be accomplished with the help of the Eugène program, which we have previously used in the annotation of the Arabidopsis, Populus and Ostreococcus genomes. Eugène is a unique platform for genome annotation because it allows to take into account an arbitrary number of information sources (sequence similarity, splice site predictions, coding potential supplied by an Interpolated Markov Model (IMM), start site prediction, …) that will all be taken into account to provide a high quality gene annotation.

Currently, different projects are underway to sequence the genomes of other saprotrophic, pathogenic and mycorrhizal fungi. By way of a comparative analysis of all these genomes we will be able to study the structural evolution of these genomes. Furthermore, by combining these comparative analyses with expression data we will be able to identify genes that are responsible for the initiation of the symbiosis between the fungi and the trees, but also for the association between fungi and bacteria.

jan.wuyts@psb.ugent.be Jan Wuyts

Application of Lempel-Ziv complexity to alignment-free sequence comparison of protein families

Sofiène Bacha (1) and Denis Baurain (2)
(1) Department of Electrical Engineering and Computer Science, Montefiore Institute B28, University of Liège, B-4000 Liège ; (2) Department of Life Sciences, Institute of Botany B22, University of Liège, B-4000 Liège

Based on the fundamental concept of evolution applied to informational macromolecules, comparison of genetic sequences is one of the cornerstones of modern biology. From a computational point of view, a genetic sequence can be considered as a string of characters from one of the two alphabets of Life: DNA for genes (4 nucleotides) and amino-acids for proteins (20 residues). Sequence comparisons are usually performed to estimate some sort of evolutionary distance between a pair of sequences. In a phylogenetic context, pairwise distances are used to infer a tree that should represent relationships between sequences, i.e. which ones are closely related and which ones are distantly related. Most sequence comparison methods require a first step, which is aligning characters thought to be homologous and thus supposed to be derived from the same position in the ancestral sequence. Hence, comparison may proceed along sequences on a pairwise basis to score differences at each position by making use of a transition matrix associated to an evolutionary model. Although efficient algorithms are readily available, sequence alignment remains difficult due to gene (or protein domain) rearrangements, inversion, transposition and translocation at the substring level, as well as to unequal length of sequences following insertion and deletion events. Moreover, since the alignment often requires human intervention, this step is prone to subjectivity and can lead to biased results.

Two main categories of alignment-free methods have been proposed to overcome the limitations of the alignment-based sequence comparisons. The first category is founded on the statistics of word frequency, on the distances defined in a Cartesian space defined by the frequency vectors, and on the information content of frequency distribution, whereas the second category includes methods that do not require resolving the sequence with fixed word length segments. Among the latter, methods based on information theory and coding are particularly elegant. In such methods, the distance metric is the algorithmic complexity as defined in Kolmogorov complexity theory. While there are presently no absolute measures of algorithmic complexity, it can be estimated through the use of compression algorithms that are assumed to be efficient. Along the same lines, Otu and Sayood (2003) have recently proposed to rely on the well-known Lempel-Zip (LZ) complexity to estimate the distance between two DNA sequences.

Because it is based on exact repeats, the LZ complexity works well with the small DNA alphabet. However, when applied to protein sequences, such an approach is expected to miss many significant similarities, which should lead to worse performance. This would be due to the greater complexity of the amino-acid alphabet and especially to the subtle and partly overlapping relationships characterizing the 20 residues.

In this work, we present several variants of a simple strategy in which protein sequences are encoded to a new alphabet prior to computation of the LZ complexity. The key idea is to capture as much information as possible in order to enhance the phylogenetic performance of the method when applied to proteins. This includes (i) back-translating sequences to variably degenerate binary codons and (ii) mapping sequences to strings of binary-coded sets in an attempt to account for the biochemically meaningful grouping of amino-acids. We then evaluate the usefulness of our proposals by comparing their performance against both word statistics methods and alignment-based similarity measures in the context of the recognition of SCOP/ASTRAL relationships as described by Vinga et al. (2004). Furthermore, we examine their ability to infer evolutionary history by applying them to the phylogeny of several metal transport protein families for which robust trees based on aligned sequences have recently been made available.

denis.baurain@ulg.ac.be Denis Baurain

Application of The Vector Space Model for Information Retrieval to (I) the Classification of the Promoters of Human Genes and (II) Association Analysis of Group-specific Transcription Factor Binding Sites

Pieter J. De Bleser, Mohamed Lamkanfi, Michael Kalai, Dominique Vlieghe, Bram De Craene, Geert Berx, Frans Van Roy and Peter Vandenabeele
Bioinformatics Core, Department of Molecular Biomedical Research VIB- Ghent University, B-9052 Zwijnaarde, Belgium

Traditionally, co-expressed genes are identified by cluster analysis of their expression data, whereas searching their genomic sequences for motifs that are statistically over-represented leads to the actual detection of potential transcription factor binding sites. Here, we consider the reverse problem: given a set of genes, can we predict which genes are likely to be co- regulated based on the sequence features of their nearby non-coding DNA sequences? Acknowledging the analogy to the problem of automatic text classification, we have adapted the vector model for information retrieval in an attempt to predict which genes in a given set are potentially co-regulated. In addition, this vector model allows the easy identification of the associations between the sequence features considered within the predicted sets of potentially co-regulated genes.

We have analyzed this method by assessing its ability to segregate two groups of 10 artificial promoter sequences, each containing a set of transcription factor binding sites (TFBS) specific for its group. We demonstrate that the two groups of promoters are reliably segregated when using either oligonucleotide sequences or known TFBS as textual features. Furthermore, our method is capable of identifying the associations that exist between the implanted TFBS in the complete set of 20 promoters. We then used this approach to classify the promoters of the human caspase genes, the promoters of human direct target genes of p53 and to identify the associated sequence elements within the groups of promoters. We show that the predicted groups and their biological functions are highly correlated and retrieve in a single step the group characteristic TFBS associations.

pieterdb@dmbr.ugent.be Pieter De Bleser

Applying well-established software engineering practices to data-centric biological applications: two case studies.

Richard Kamuzinzi, Morgane Thomas-Chollier, Albert Herzog, Valérie Ledent
Université Libre de Bruxelles

Richard Kamuzinzi1, *, Morgane Thomas-Chollier1,2,*, Albert Herzog1 and Valérie Ledent1

1 Université Libre de Bruxelles, Institut de Biologie et de Médecine Moléculaires, Charleroi-Gosselies, Belgium
2 Vrije Universiteit Brussel, Laboratory for Cell Genetics, Brussels, Belgium
* Equal contributions

High throughout projects have led to the exponential growth of data. The use of current standards from the software engineering field provides for more reusable, scalable, adaptable and maintainable applications that are able to cope with new trends coming from distributed bio-applications. We present two case studies, namely Associoport and Kermit, which illustrate a way to enforce these design practices in a bioinformatics context. Currently still in progress, these two biological projects are respectively dedicated to store and present experimental data on protein-protein interactions and on annotated genomic sequences. Both systems are based upon the standard three-tiers architecture. The overall design of Kermit and Associoport is organised in layers containing loosely-coupled components with clearly-defined responsibilities. The design of our databases is exclusively based on well-established relational concepts like normalization and is modelled by Entity-Relation conceptual schemes. Conversely, the programmatic access is driven by OO modelling. In this multi-model approach, the domain objects and tables of the databases were sometimes dissimilar and demanding to connect (Object/Relational (O/R) mismatch). One particular aspect of our implementation is the use, in a biological context, of the O/R mapping framework Hibernate to tackle this mismatch. This approach has considerably reduced the development time and helped us to focus on meeting scientific business requirements. In the context of Hibernate usage, we propose a decomposition of the Domain Object Model into fine-grain and coarse-grain objects, respectively produced by the database-driven and object-driven modelling of the Associoport and Kermit projects.

valerie.ledent@ulb.ac.be Valérie Ledent

BAGEL: a web-based BActeriocin GEnome mining tooL

A. de Jong*, S.A.F.T. van Hijum, J. Kok, O. P. Kuipers
University of Groningen

BAGEL automatically annotates proteins sharing complex combinations of properties. Bacteriocins were used as a test-case for this novel concept. Bacteriocins are bacterial antimicrobial peptides which are of great interest scientifically, as food-preservatives and, importantly, for their potential as future antibiotics. These small peptides can be subdivided into 5 separate classes which are defined by a combination of features: (i) peptides with specific iso-electric points, (ii) the proteins have distinctive secondary structures, (iii) the peptides bear amino acid motifs, and (iv) their genomic context. As these anti-microbial peptides are small and share no strong sequence homology with other, known, bacteriocins, they are often missed in genome sequence annotations. To tackle this problem a web-based service is offered to determine (small) ORFs in a genome sequence. A database of known and experimentally validated bacteriocins and their classification was constructed. Using this database and empirically determined classification rules, the known bacteriocins were successfully on-the-fly annotated from genome sequences. In addition, a number of novel bacteriocins were detected in the genome sequence of food and pathogenic bacteria.

Anne.de.Jong@rug.nl Anne de Jong

BIGRE: An Ontology Driven Bioinformatics Service Integration Environment

Olivier DUGAS, Joseph MAVOR, Pierre BUYLE,Quentin DALLONS, Quentin DALLONS, Amin MANTRACH, Utku SALIHOGLU, Hugues BERSINI, Vincent ENGLEBERT, Marc COLET
Service de Bioinformatique, IBMM, Universite Libre de Bruxelles

This paper describes the BIGRE project in which an ontology-based serviceintegration framework is used to support in-silico experiments in Biology.The growing quantity and distribution of Bioinformatics databases (719databases in 2005 and associated services, require an ever growing userexpertise and knowledge, especially as many resources need to be tied togetherinto a workflow to accomplish a useful goal.The BIGRE ontology includes a data model and a service model intended tosupport the BIGRE distributed software framework. The data model allows usto reconstruct heterogeneous Bioinformatics service data output, databaseformats, Biological concepts, etc. The result is the ability of BIGRE to buildcomplex workflows that link legacy databases and services into a singleBioinformatics environment.The BIGRE architecture is designed to be decoupled from the ontology so as toallow other domains to utilize the framework in other contexts.

odugas@dbm.ulb.ac.be Olivier DUGAS

Characterization of the universal mitochondrial cohort of genes

Cedric Simillion, Martin Embley, Yves Van de Peer
Department of Plant Systems Biology - Ghent University/VIB and School of Biology - University of Newcastle upon Tyne, UK

It has long been tought that certain unicellular, mainly parasitic, eukaryotes such as Giardia and the microsporidia had diverged from the eukaryote taxa before the mitochondrial endomsymbiosis event as they seemed to lack a mitochondrion. Recently however, it has recently become clear that all these organisms do contain a highly reduced form of mitochondrial organelle, called the mitosome. Next to this, other organisms such as the sexually transmitted parasite Trichomonas vaginalis contain an organelle called the hydrogenosome which is also believed to be homologous to the mitochondrion. At present, it is unknown what the exact function of these organelles, let alone if there is a common function for all mitochondria-derived organelles.

The goal of the project presented here is to use computational biology to identify candidate functions for the microsporidian mitosome, and to determine if we can identify a cohort of conserved 'mitochondrial' genes in diverse aerobic, parasitic and anaerobic eukaryotes. This cohort would provide the best candidates for a common essential function (if it exists) for the 'mitochondrial' organelle under diverse living conditions. To achieve this, we use a diverse range of sequence comparison and data mining tools to identify proteins that are associated with the mitochondrial organelle in a wide range of eukaryotic genomes.

cedric.simillion@psb.ugent.be Cedric Simillion

CoGenT++: An extensive and extensible data environment for computational genomics

Leon Goldovsky1, Paul Janssen2, Dag Ahrén1,3, Benjamin Audit4, Ildefonso Cases5, Nikos Darzentas1, Anton J. Enright6, Victor Kunin1, Núria López-Bigas1, José M. Peregrin-Alvarez7, Mike Smith1, Sophia Tsoka1 & Christos A. Ouzounis1

1 Computational Genomics Group, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, UK 2 Laboratory for Microbiology, Belgian Nuclear Research Center, SCK/CEN, Boeretang 200, B-2400-MOL, Belgium 3 Institute of Agrobiotechnology, National Center for Research and Technology, PO Box 361, Thermi, Thessaloniki, GR-57001 Greece 4 Laboratoire de Physique, Ecole Normale Supérieure, 46 Allée d' Italie, Lyon CEDEX 07, F-69364 France 5 Transcription Networks Group, National Center for Biotechnology, CNB-CSIC, Cantoblanco Madrid 28049, Spain 6 Sanger Institute, Wellcome Trust Campus, Cambridge CB10 1SA, UK 7 Hospital for Sick Children, 555 University Avenue, Toronto ON M5G 1X, Canada

CoGenT++ is a new working environment for computational research in comparative and functional genomics and amalgates a number of services based on existing databases and precomputed data. The core of this platform is the CoGenT database [1] currently holding data of 221 completely sequenced (and published) genomes, including 22 eukaryota, 20 archeae, and 179 bacteria. The 822,114 protein sequences in this version of CoGenT (as of March 13, 2005), together with all entries of Swiss-Prot (release 42.11) [2], were used to establish a very large pairwise similarity database, ProXSim, forming the basis for secondary information resources that are highly useful for metabolic reconstructions (MeRSy), phylogenetic tree analyses (GPS), phylogenetic profiling (ProfUse), protein clustering (TRIBES), ortholog extraction and –validation (OFam), and protein fusion analysis (AllFuse). In addition, the CoGenT++ environment is now fully cross-linked to major molecular biology databases, namely UniProt, EMBL, GenBank, RefSeq, and PDB, using the MagicMatch algorithm [3]. The central CoGenT database is constantly being updated, entailing an incremental update of ProxSim. This is ensured by using the BlastP bit-score as an estimate of sequence similarity, avoiding time consuming and costly recalculation of the entire database [4]. Thus, CoGenT++ provides a scalable, customisable, and automatic update mechanism for the pairwise comparison of all (sequenced and published) genomes, allowing clustering, ortholog searches, pathway predictions, and more, for an ever growing body of genome data. In its current form, this novel platform provides access to more than 40 Gigabytes (GB) of data. An elaborate interactive schema and download of components are available at http://cgg.ebi.ac.uk/cgg/.

[1] Janssen P.J., Enright A.J., Audit B., Cases I., Goldovsky L., Harte N., Kunin V., Ouzounis C.A. (2003) COGENT: a flexible data environment for computational genomics Bioinformatics 19(11): 1451-1452
[2] other protein databases can be used and linked to the system as well
[3] Smith et al. (submitted); accessible at: http://cgg.ebi.ac.uk/services/magicmatch/
[4] All CoGenT data are fully available as flat-file and MySQL dumps

pjanssen@sckcen.be Paul Janssen

Comparative Annotation: exploiting a hidden wealth

Rombauts S., Sterck L., Robbens S., Degroeve S., Schiex S., Rouze P., Van de Peer Y.
Bioinformatics and Evolutionary Genomics, Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

With the completion of the human genome, resources became available in terms of hardware, time and people to sequence genomes from other organisms. This implies that there is a wealth of genome sequence data in the pipeline that is going to be at the disposal of the scientific community.

Although raw data will be generated at a relative high pace, genome annotation, being the first step towards exploring the genomes, needs to adapt its strategies to analyze the available data. Therefore, we have setup a gene prediction platform that allows us to exploit and analyze the available genomes in an automatic way. This gene calling platform called Eugene was jointly developed at the University of Ghent and INRA Toulouse and has successfully been applied to different plant genomes, from algae to trees.

For instance, we have been involved in the annotation of the green alga Ostreococcus tauri that, due to many peculiarities of its genome, still needed a high degree of expert driven annotation. The poplar genome structural annotation on the other hand, could be achieved in roughly 8 months - amongst other things because of the availability of the Arabidopsis genome - , including the building of training set, training of the whole software on poplar, and the gene prediction itself.

strom@psb.ugent.be Stephane Rombauts

Comparative evolutionary analysis of diatoms and other protists based on complete genome sequences

Cindy Martens(1), Klaas Vandepoele(1), Koen Sabbe(2), Wim Vyverman(2), Yves Van de Peer(1)
(1)Bioinformatics and Evolutionary Genomics, Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium (2)Laboratory of Protistology & Aquatic Ecology, Department of Biology, Ghent University, Krijgslaan 281, S8, B-9000 Ghent, Belgium

Diatoms are unicellular, photosynthetic eukaryotes that belong to the monophyletic group of the Chromalveolates. In spite of their economic importance and their ecological success in the oceans, very little is known about their molecular biology.

In the context of the International Diatomics Project, in which our research group will take part, there will soon be two completely sequenced diatom genomes available. Together with other genomic data of several Chromalveolata, we will study the evolution of genes and gene families within this group of protists by means of phylogenetic profiles. This gives us the opportunity to characterise the genome dynamics and the conservation of gene organisation between closely and more distantly related protists.

Since the current phylogeny of the diatoms is mainly based on morphological characteristics, a second objective is to clear up the molecular phylogeny within the Bacillariophyceae. Based on conserved low copy gene families within the pennate and centric diatom genome, we will try to identify candidate phylogenetic markers. In cooperation with the lab Protistology and Aquatic Ecology, these phylogenetic markers will be experimentally validated.

To study the evolution within the diatoms and Chromalveolates, we will also examine the complexity of transcriptional regulation. Thanks to the availability of two diatom genomes, we can identify the regulatory elements by means of comparative methods like phylogenetic footprinting. The conservation of these cis-regulatory elements can also be studied in other Chromista and/or Alveolata to examine if the transcriptional regulation is generally conserved within these protists. Finally, we hope to compare the genomic complexity of these morphologically simple eukaryotes with more complex eukaryotes at both the gene and transcriptional level.

cindy.martens@psb.ugent.be Cindy Martens

Critical Assessement of PRedicted Interactions (CAPRI): Current Status of Protein-Protein Docking Methods.

Raúl Méndez, Marc F. Lensink and Shoshana J. Wodak
Service de Conformation des Macromolécules Biologiques et de Bioinformatique (SCMBB). Université Libre de Bruxelles, Campus La Plaine, Bvd du Triomphe - CP263, B-1050 Bruxelles. Belgium

The current status of docking procedures for predicting protein-protein interactions starting from the tree-dimensionalstructure of the individual components is assessed from a major evaluation of blind predictions. This evaluation is performed as part of a community-wide experiment on Critical Assessement of PredictedInteractions (CAPRI). Sixteen newly determined structures of protein-protein complexes were available as targets, comprising 3 complexes involved in cellular signalling, an enzyme inhibitorcomplex, 7 antigen antibody complexes, 2 homo-oligomers, a T-cell receptor beta-chain with a superantigen, and 2 components of thebacterial cellulosome. The structure of the complex was revealed only at the time of the evaluation. A total of 2331 predictions submitted by 34 groups over 5 rounds up to day have been so far evaluated and their results deposited at http://capri.ebi.ac.uk/. These groups used a wide range of algorithms and scoring functions, some of them were completely novel or being improved during successive rounds of CAPRI.

The quality of the predicted interactions was evaluated by comparing residue-residue contacts and interface residues to those in the X-ray structures, fitting the predictions onto the target complex. Twenty-six groups, which included a fully automatic web server, produced predictions ranking from acceptable to highly accurate for all the targets, including those where the structure of the bound and unbound forms differ substantially.

These results, presented here together with a brief survey of the methods used by participants of CAPRI Rounds 1-5, suggest that genuine progress in the performance of docking methods is being achieved, with CAPRI acting as the catalyst. The CAPRI experiment has started in September 2001, going now by its Round 6, which its evaluation is still in progress, and it is open to everybody willing to accept the challenge.

raul@scmbb.ulb.ac.be Raúl Méndez

CRMD and systems biology: towards a general computational framework

Wouter Van Delm, Yves Moreau and Bart De Moor
Div. ESAT SCD : SISTA, KULeuven, Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium

The detection and location in DNA of transcription factor binding sites is a problem within functional genomics which has not yet been solved. The first generation of prediction-algorithms was searching for one motif at a time. Being unsuccessful for higher eukaryotes - where more factors work together and the signal-to-noise ratio (SNR) is lower - a second generation is searching for specific combinations of motifs (modules) which influence together the expression of a gene. This research project - which started half a year ago - wants to model the modules with Hidden Markov Models, based on their successful application in other sequence-modeling domains. Existing training algorithms suffer too much from local optima. Therefore more appropriate Markov Chain Monte Carlo methods are developed for HMMs in a Bayesian framework. They have proven their value in similar applications. To benefit from research in neighboring fields, the problem will be placed within the context of systems biology. The final aim of the project is to create within the next 3.5 years a general theoretical probabilistic framework and practical software platform for Cis-Regulatory Module Discovery (CRMD) which will be validated on a case study concerning heart failure.

wvandelm@esat.kuleuven.ac.be Wouter Van Delm

Data driven protein-protein docking: HADDOCK's adventures in CAPRI

A.D.J. van Dijk, C. Dominguez, S.J. de Vries and A.M.J.J. Bonvin
Utrecht University

We have shown previously that, given high-resolution structures of the free molecules, determination of structures of protein complexes is possible by combining biochemical and/or biophysical data like mutagenesis data or NMR derived chemical shift perturbation data with docking approaches. Here we apply this method, implemented in the HADDOCK (High Ambiguity Driven DOCKing) package, to the targets in the fourth and fifth round of CAPRI (Critical Assessment of Predicted Interactions), using all kind of available biochemical and/or biophysical data, in combination with sequence conservation data.
Analysis of our results proves our method to be robust for most targets, although it is clearly depending on the availability and quality of data for the complexes under study. Furthermore, we show that it is worth to pay the computational price for adding flexibility in docking, as this improves the docking results.

a.j.vandijk@chem.uu.nl Aalt-Jan van Dijk

Development of a high throughput annotation pipeline: cyrille2

Mark Fiers, Joost de Groot and Roeland van Ham
Plant Research International

High throughput sequencing must be matched by high throughput annotation. Given the large number of annotation tools available, many thousands of interdependent analyses are required for an in-depth annotation of even a single BAC sequence. In order to perform this we are in the process of developing and implementing an automated annotation pipeline, cyrille2. The software will be able to take an incoming dataset (i.e. a set of BAC sequences) and in a fully automated manner annotate the data on a linux cluster and store the annotation results in a database. The annotation results will then be dynamically visualized using publically available software such as the Generic Genome Browser and Ensembl.
Important features of the software will be:

* High Throughput. It will be possible to process very large pipelines or datasets.
* Flexible. It will be easy to add other annotation tools.
* Incremental. New data or databases should be incorporated efficiently, without having to rerun the complete pipeline.
* Database independent. Use of another database or multiple databases is easy.
* Interchangable components. By using a standard language (bioMoby XML) for communication between the pipeline components, introducing new components will be easy. Also implementing new components is simplified this way.
* Ease of use. The software will have a clear GUI and will be operable by a non expert annotator.

The current development status of the pipeline will be presented.

Mark.Fiers@wur.nl Mark Fiers

Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences

Eric Bonnet, Jan Wuyts, Pierre Rouzé and Yves Van de Peer
University of Ghent / VIB

Most non-coding RNAs are characterized by a specific secondary and tertiary structure that determines their function. Here, we investigate the folding energy of the secondary structure of non-coding RNA sequences, such as microRNA precursors, transfer RNAs and ribosomal RNAs in several eukaryotic taxa. Statistical biases are assessed by a randomization test, in which the predicted minimum free energy of folding is compared with values obtained for structures inferred from randomly shuffling the original sequences.

In contrast with transfer RNAs and ribosomal RNAs, the majority of the microRNA sequences clearly exhibit a folding free energy that is considerably lower than that for shuffled sequences, indicating a high tendency in the sequence towards a stable secondary structure. This statistical test can be used in the framework of the detection of genuine miRNA sequences, both in animals and plants.

The dataset, software and additional data files are freely available as supplementary information on our Website.

eric.bonnet@psb.ugent.be Eric Bonnet

Exploring the plant transcriptome through phylogenetic profiling

Klaas Vandepoele and Yves Van de Peer
Bioinformatics and Evolutionary Genomics, Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

Publicly available protein sequences represent only a small fraction of the full catalogue of genes encoded by the genomes of different plants, such as green algae, mosses, gymnosperms and angiosperms. In contrast, an enormous amount of expressed sequence tags exists for a wide variety of plant species, representing a substantial part of all transcribed plant genes. Integrating protein and EST sequences in comparative and evolutionary analyses is not straightforward because of the heterogeneous nature of both types of sequence data. By combining information from publicly available EST and protein sequences for 32 different plant species, we identified more than 250,000 plant proteins organized in over 12,000 gene families. Approximately 60% of the proteins are absent from current sequence databases, but provide important new information about plant gene families. Analysis of the distribution of gene families over different plant species through phylogenetic profiling reveals interesting insights into plant gene evolution, and identifies species and lineage specific gene families, orphan genes, and conserved core genes across the green plant lineage. We counted a similar number of approximately 9,500 gene families in monocotyledonous and eudicotyledonous plants and found strong evidence for the existence of at least 33,700 genes in rice. Interestingly, the larger number of genes in rice compared to Arabidopsis can partially be explained by a larger amount of species-specific single copy genes and species specific gene families. In addition, a majority of large gene families, typically containing more than 50 genes, is bigger in rice than Arabidopsis, whereas the opposite seems true for small gene families.

klaas.vandepoele@psb.ugent.be Klaas Vandepoele

FIVA: FUNCTIONAL INFORMATION VIEWER and ANALYZER for functional profiling of bacterial transcriptome data

E.J.Blom, D.W.J. Bosman, S.A.F.T. van Hijum, J.B.T.M.Roerdink, O.P.Kuipers
Molecular Genetics, University of Groningen, Institute for Mathematics and Computing Science, University of Groningen

Transcriptional profiling experiments allow determining global gene transcript levels of an organism. For these experiments clustering techniques can be used to group genes based on their (temporal) expression patterns. Genes involved in biologically related processes should exhibit similar expression patterns and thus appear in the same cluster. Analysis and interpretation of these clusters derived from transcriptome analyses is time consuming.
Therefore an application that determines functional categories which are significantly overrepresented would greatly assist in interpreting transcriptome data.

Results
We developed FIVA (Functional Information Viewer and Analyzer) which is capable of processing clusters of genes exhibiting similar gene expression patterns. Currently four different modules containing functional information have been implemented: (i) gene regulatory interactions, (ii) metabolic pathways, (iii) cluster of orthologous groups (COG) of proteins - and (iv) gene ontologies. To validate FIVA, two publicly available transcriptome data sets were analyzed. The first dataset contained a transcriptional profile of a knock-out experiment. We demonstrate that FIVA is able to identify a large number of known targets. The analysis of the second dataset reveals a biological process which was not discussed in the original study.

Conclusions

FIVA was developed to aid users in quickly identifying relevant biological processes affected in a DNA microarray experiment. We demonstrate that our program is able to assist in functional profiling of large sets of genes and generate a comprehensive overview.

e.j.blom@rug.nl Evert Jan Blom

Gene expression analysis and classification of Leigh and MELAS skeletal muscle and skin fibroblasts.

R. van Eijsden, L. Eijssen, R. Mineri, P. Lindsey, C. v.d. Burg, W. Sluiter, K. Schoonderwoerd, I. de Coo, H. Scholte, M. Rubio, T. Ayoubi, V. Tiranti,J. Geraedts, H. Smeets.
Maastricht University, Dep. of Genetics & Cell Biology

Mitochondrial encephalomyopathies are genetically and clinically heterogeneous. Disorders, like Leigh syndrome, can be caused by defects in different genes, whereas single mutations, like the mtDNA A3243G mutation, have a variable clinical manifestation. In addition, the gene defects can be located in either the nuclear or mitochondrial DNA. To improve the diagnosis and explain the pathology observed, we applied global gene expression profiling in samples from patients and controls. Fibroblasts from patients with Leigh syndrome due to a Surf1 mutation could be separated from controls by unsupervised hierarchical clustering using the expression values from only 12 (t-test), 13 (SAM: Significance Analysis of Microarrays) or 14 genes (LIMMA: Linear Models for Microarrays). Combining these three gene lists yielded 21 genes, which we are currently investigating for involvement in the pathology and for their ability to classify. In a second study we determined gene expression profiles in muscle from affected and unaffected carriers of the MELAS-A3243G mtDNA mutation and controls. Unsupervised hierarchical clustering could separate carriers from controls with the 2 top genes from a t-test. The most significantly altered pathways involved protein degradation and synthesis and ROS damage with a clear difference between symptomatic and asymptomatic carriers. Our data indicate that gene expression profiling is promising for classifying patients with mitochondrial encephalomyopathies and in elucidating new pathophysiological concepts.

Rudy.vanEijsden@gen.unimaas.nl R.G.E. van Eijsden

Genome analysis of the world's smallest free-living eukaryote Ostrecococcus tauri unveils unique genome heterogeneity

Steven Robbens, Stephane Rombauts, Pierre Rouzé, Jan Wuyts, Sven Degroeve, Hervé Moreau, and Yves Van de Peer
Bioinformatics and Evolutionary Genomics, Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

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 12.65 Mb, divided over 19 chromosomes. 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 experimental studies, but also for evolutionary 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.

Regarding the genome itself, there are a number of very unexpected and unique findings that have never been observed in any of the eukaryotic genomes sequenced to date. The most striking is the genome heterogeneity, where the chromosomes 2 and 18 are surprisingly different from all other 17 chromosomes. Genes on chromosome 2 have unique intron features, unlike the genes found on the other chromosomes. Also, most of the transposable elements in Ostreococcus are found on both aberrant chromosomes. Another unique and important aspect of this genome is the extreme simplification of the gene complement. For many pathways and cellular processes (one example being the cell cycle where only one member of each cell cycle gene family was detected), the number of genes is kept to an absolute minimum. The genome of Ostreococcus is also exceptionally compact. The result is that gene transcripts, and sometimes even the CDS themselves, are overlapping, a feature rarely observed in eukaryotes. Furthermore, we observed the clustering of functionally related genes comparable to bacterial organization.

In addition to the nuclear genome, the chloroplast and mitochondrial genome were also sequenced and annotated. The circular chloroplast genome is about 60,600 bp long with an overall GC content of 43% and contains 93 genes. The mitochondrial genome with its 66 genes is about 42,500 bp long and has an almost identical gene repertoire compared to Nephroselmis olivaea, another green alga. In order to classify Ostreococcus tauri within the tree of life, an extensive phylogenetic analysis was performed. Phylogenies were inferred on the basis of concatenated chloroplast genes, concatenated mitochondrial genes, and a combination of chloroplastic, mitochondrial and nuclear genes. All datasets and methods confirmed Ostreococcus being a basal green alga, closely related to Nephroselmis.

steven.robbens@psb.ugent.be Steven Robbens

Genome-wide Analysis of the Sequence Conservation of Gene Regulatory Sites

Dominique Vlieghe, Pieter De Bleser and Frans van Roy
Bioinformatics Core, Department for Molecular Biomedical Research, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

In recent years, phylogenetic footprinting has emerged as a powerful tool for predicting gene regulatory sites. Although this methodology is based on solid biological principles, a systematic validation of its strengths and weaknesses has not been published yet. Here we report on the analysis of the sequence conservation of 1,700 regulatory sites extracted from experimental studies.
A tool, denoted genomeTF, was developed to map TRANSFAC regulatory sites onto the human genome. Human genes and their upstream regions were then compared with orthologous counterparts from a large set of organisms: chimpanzee, mouse, rat, chicken, pufferfish, zebrafish, fruit-fly and nematode. The results show that phylogenetic footprinting is indeed very suitable for gene regulatory site prediction: the average sequence identity of aligned regulatory sites is 99, 72 and 54% between man and chimpanzee, rodents and chicken, respectively. Chicken represents the practical evolutionary limit for valuable phylogenetic comparisons to man. Going further back in evolutionary time reduces the functionality of the alignments for regulatory site predictions. The main reason for error is the non-alignment of regulatory sites: this effect attributes entirely to the reduction in predictive power observed for regions far away from the transcription start site or for distant orthologues. Use of several organisms in a multiple sequence alignment, together with further improvements in genome sequencing and gene and orthology annotation, could improve footprinting approaches for these difficult cases.

Dominique.vlieghe@dmbr.ugent.be Dominique Vlieghe

Genome-wide detection and analysis of cell-wall bound proteins with LPxTG-like sorting motifs

Jos Boekhorst(1), Mark de Been(1,2), Michiel Kleerebezem(2,3), Roland J. Siezen(1,2,3)
(1) Center for Molecular and Biomolecular Informatics, Radboud University Nijmegen, the Netherlands; (2) 2Wageningen Centre for Food Sciences, Wageningen, the Netherlands; (3) NIZO food research, Ede, the Netherlands

Surface proteins of gram-positive bacteria often play a role in adherence of the bacteria to host tissue and are frequently required for virulence. A specific sub-group of extracellular proteins contain the cell-wall sorting motif LPxTG, which is the target for cleavage and covalent coupling to the peptidoglycan by enzymes called sortases.
A comprehensive set of putative sortase substrates was identified by in silico analysis of 199 completely sequenced prokaryote genomes. A combination of detection methods was used, including secondary structure prediction, pattern recognition, sequence homology, and genome context information. With the hframe algorithm, putative substrates were identified that remained undetected by other methods due to errors in ORF calling, frameshifts or sequencing errors. In total, 732 putative sortase substrates encoded in 49 prokaryote genomes were identified. We found striking species-specific variation for the LPxTG motif.
A Hidden Markov Model based on putative sortase substrates was created, which was subsequently used for the automatic detection of sortase substrates in recently completed genomes. This model detects putative sortase substrates with a false positive rate of 5%.
A database was constructed, LPxTG-DB http://bamics3.cmbi.kun.nl/sortase_substrates), containing for each genome a list of putative sortase substrates, sequence information of these substrates, the organism-specific HMMs based on the consensus sortase recognition motif, and a graphical representation of this consensus.

J.Boekhorst@cmbi.ru.nl Jos Boekhorst

In Silico Prediction Of Cis-regulatory elements By Use Of Phylogenetic Footprinting.

M. Wels1,2,3, R. Kerkhoven 2, W.M. de Vos 1,4, M. Kleerebezem 1,3, R.J. Siezen 1,2,3
1. Wageningen Centre for Food Sciences, 2. Radboud University Nijmegen, 3. NIZO food research, 4. Wageningen University.

Cis-regulatory elements of Lactobacillus plantarum were predicted by comparative analysis of the upstream regions of conserved genes and operons in different bacteria. For two species sets, with different evolutionary distance to L. plantarum, operons were predicted based on intergenic distance, gene orientation and predicted rho-independent termination signals. Next, the genes in the operons were compared to each other. Operons in which >50% of the proteins were orthologous, were considered to be part of one "cluster of orthologous operons" (COOP). For each COOP, the upstream regions of the operons were compared using the motif prediction tool MEME. Comparisons were made between the two different species sets using the profile comparison program COMPASS. Subsequently, the upstream regions of all predicted operons were scanned for the presence of the MEME-generated motif. The species in the Lactobacillaceae set, which have a smaller evolutionary distance to L. plantarum than the Bacilli set, are predicted to have more conserved cis-regulatory elements, from which several could be validated in literature. On the other hand, using species from the order of Bacilli showed to be better suited for the prediction of regulons: the predicted motifs occur more often in front of different (literature validated) co-regulated operons, in comparison to the motifs from the Lactobacillaceae set.

mwels@cmbi.ru.nl Michiel Wels

Initial analysis of heterosis expression data with machine learning methods

Jeroen Meeus, Elena Tsiporkova
Computational Biology Division, Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

Heterosis refers to an improved performance of F1 hybrids with respect to their parents. It has been observed that a cross between quasi-homozygous parents in some cases leads to an offspring (F1) that is better in terms of yield, stress resistance or speed of development compared to both the parents. Heterosis is particularly important for commercial agricultural crops as for instance corn, sugar beet and sunflower, as well as for vegetables and for the commercial forestry. Besides a commercial interest there is a more fundamental scientific interest associated with the biological phenomenon of heterosis performance, as an excellent example of what complex genetic interactions can lead to.

Although the phenomenon of heterosis has already been studied for many years, no complete genetic explanation has been found. Heterosis is in some way correlated with the genetic distance between the parents and their level of homozygosity, which in turn causes a certain level of heterozygosity in the first generation of the offspring. However the latter explains only at most half of the heterosis encounters. The F1 hybrids are nowadays tested for heterosis features mostly via a "trial-and-error" method which, considering the great number of parental lines that have to be tested, turns out to be an enormous and also rather expensive task. The choice of parental lines for an F1 program could be performed in a more selective fashion if sufficient information concerning the combining abilities of the parental lines would be available in advance to the breeder.

In our search for the genetic basis of heterosis, we use Arabidopsis thaliana as a model system. Around fifty F1 crosses are being made between different Arabidopsis ecotypes chosen to induce a broad spectrum of heterosis observations. These F1s are phenotypically characterized for several properties as for instance biomass and speed of development. Moreover all the ecotypes considered as parental lines are being genotyped using AFLP-marker technology. A complete transcription profiling of both parents and hybrids using micro-arrays is also about to be completed. The resulting data is expected to provide information about the difference in gene activity between different ecotypes and between the hybrids that show heterotic effects and those that don't.

The analysis of micro-array data entails very specific problems. Due to the cost of such an experiment, one is often forced to perform analysis on a small number of data points in a very high dimensional space. The method should also be able to cope with very high noise levels. In the past years it has been shown that a relatively simple learning paradigm, the support vector machine (SVM), outperforms even the most elaborately tuned expert systems and neural networks in learning to recognize patterns from sparse training examples. Underlying its success are the mathematical foundations of statistical learning theory.

The analysis of the partially available micro-array data has already been launched, both with standard biometrics techniques and soft-computing algorithms. The ultimate goal is to build a computational model on the basis of Arabidopsis gene expression data that enables prediction of hybrid performance with higher efficiency than the usual genetic markers. The results obtained with micro-arrays and AFLP markers will of course be analysed and compared. The classification markers will subsequently be used for identification of key genes involved in the biological processes underlying heterosis and it will further be investigated in which genetic networks these genes play a role. Finding a set of genes that show a high correlation with biomass production can be a first step in looking for the genetic basis of heterosis. Methods under current investigation include regression trees and support vector machines. We will present our initial findings in the heterosis data with these techniques.

jeroen.meeus@ugent.be Jeroen Meeus

Insights into the role of specific residues in the formation of the domain swapped dimer of GB1 by computational protein design.

Fernanda L Sirota, Sebastian Maurer-Stroh, Shoshana J Wodak
ULB Universtite Libre de Bruxelles, Service de Conformation de Macromolecules Biologiques et de Bioinformatique, Bruxelles – Belgium and IMP Institute of Molecular Pathology, Vienna – Austria

The B1 domain of the immunoglobulin G binding protein (protein G) from group G of Streptococcus has been widely used as model for studying and understanding the mechanisms of protein folding and unfolding. Several attributes of this protein contributed to its success and popularity as a model. Among these are its small size (56 residue length), the absence of disulphide bridges and its thermodynamic properties that resemble those of larger proteins in addition to its resistance to denaturation by heat or urea. Consequently, the B1 domain is an extremely well characterized model system, both theoretically and experimentally. However, it has been recently experimentally determined that the quadruple mutant of GB1 (L5V / F30V / Y33F / A34F) forms a domain swapped dimer (Byeon et al., 2003). Further in the same work of Byeon et al., it was demonstrated that the mutation of a single residue is pivotal for the protein to adopt the dimeric fold. This finding was critical to point out how a single mutation can affect the stability of the monomeric fold triggering the formation of a swapped dimer. By default, proteins that differ in one amino acid only are thought to adopt the same fold. Here, it is shown that even for a well known and studied protein, as the B1 domain of protein G, this is not the case. This has far reaching consequences for protein structure prediction, protein function assignment and protein aggregation with all its importance in biomedical and industrial applications. Over the past year, we have investigated by means of computational protein design the contributions of the above four mutated residues to the triggering of the dimerization mechanism in the light of atomic detailed models. We have found out that steric hindrance and van der Waals interactions are the main contributors to the events leading to the domain swapping, suggesting that A34F unfavorably interacts with W43 and V54, instead of the initial view of a clash involving A34F and V39.

Reference: Byeon,I.J., Louis,J.M., and Gronenborn,A.M. (2003). A protein contortionist: core mutations of GB1 that induce dimerization and domain swapping. J. Mol. Biol. 333, 141-152.

fernanda@scmbb.ulb.ac.be Fernanda L Sirota

Isolating metabolic modules from gene expression data using a graph-based clustering approach.

Joseph Tran, Jacques van Helden
SCMBB - ULB

Relationship between structure, function and regulation in complex cellular networks is still a major challenge. Integrating all these information in order to predict cellular networks description is usually quite complex. In dynamic mathematical modeling, combining experimental and theoretical approaches usually meets difficulties because of the lack of experimental data.
In contrast, graph-based analyses only rely on network topology. Previous approaches of this type revealed that metabolic network could be decomposed into subunits using clustering coefficient (Ravasz et al., 2002) or using metabolites degree of network connectivity (Schuster et al., 2002). Moreover, using this last criterion most of these subunits referred to accepted metabolic pathways (Gagneur et al., 2003). Although graph analysis can lead to isolate biologically meaningful sub-networks, incorporation of additional constraints such as regulation, thermodynamics and so on, can help to cope with metabolic pathway characterization. Indeed, there is no rigorous definition yielding pathways of disparate content and size. Some attempt was previously done, by coupling transcriptional regulation with network-based analyses (Covert and Palsson, 2003) to infer metabolic pathways. However, this approach has not yet been scaled-up to genome-scale biological systems.
Here, we propose a graph-based clustering approach to infer pathways using gene regulation information. In the current study, this approach is applied to budding yeast dna chips covering transcriptional response to a large number of environmental conditions. For each condition, synexpression groups are selected. In order to understand the cellular function of these groups, 2 approaches are applied. First, these groups are compared to functional catalogs (MIPS, GO) and annotated pathways (KEGG, aMAZE) to detect significant associations. However this restricts the answer to previously characterized functional group or annotated pathway. Besides, our knowledge about metabolism has been established on the basis of a small number of model organisms, and, even for those, it is largely incomplete. Many alternative or novel pathways remain to be characterized. Thus, the pathways in which some co-expressed enzymes participate may be unknown. In addition to this knowledge-based approach, we apply graph-based clustering in order to build pathways from a set of co-expressed enzymes.
The metabolic graph is composed of all known reactions and compounds, and therefore we dont restrict the analysis to the specific known subset of enzymes from one organism. There are two major reasons: (i) incomplete genome annotation, unidentified enzymes still remain, (ii) spontaneous reactions. Previous studies already applied graph-based approaches on un-weighted metabolic graph to infer pathways (van Helden et al., 2001, 2002). But, pool metabolites (H2O, ATP, NADP+, etc.) were arbitrarily excluded from this graph. To avoid any exclusion, metabolite degree of network connectivity was used to assign weight on compounds of the metabolic graph. And recently, it was shown that this only weighting improves sensibly the relevance of path-finding procedure (Croes et al., submitted). Starting from this point, we develop our methodology. In addition to compounds, reactions weight was assigned according to the enzymes expression level in the different conditions. Indeed, this enables us to take into account the environmental (genomic and experimental) context in the whole metabolic graph. Then, the shortest path distance is computed for every pair of co-expressed enzymes using path-finding procedure. And finally, we applied single linkage clustering to merge path-finding computed solution paths between closest co-expressed enzymes pair. Indeed, short path distance characterizes enzymes pairs within most annotated metabolic pathways (Croes et al., submitted). We obtain a resulting solution graph that enables us to detect metabolic modules (activated/repressed in each condition), i.e. set of enzymes that are connected in the resulting graph.

jtran@scmbb.ulb.ac.be Joseph Tran

Large-scale structural analysis of the core promoter in mammalian and plant genomes.

Kobe Florquin, Yvan Saeys, Sven Degroeve, Pierre Rouzé, and Yves van de Peer
Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

DNA encodes at least two independent levels of functional information. The first level is for encoding proteins and sequence targets for DNA-binding factors, while the second one is contained in the physical and structural properties of the DNA-molecule itself. Although the physical and structural properties are ultimately determined by the nucleotide sequence itself, the cell exploits these properties in a way in which the sequence itself plays no role other than to support or facilitate certain spatial structures.
In this work, we focus on these structural properties, comparing them between different organisms, and assessing their ability to describe the core-promoter. We prove the existence of distinct types of core-promoters, based on a clustering of their structural profiles. These results show that the structural profiles are very conserved within plants (Arabidopsis and rice) and animals (Human and Mouse), but differ considerably between plants and animals.
Furthermore, we show that these structural profiles can be an alternative way of describing the core promoter, in addition to motif or IUPAC-based approaches. Using the structural profiles as discriminatory elements to separate promoter regions from non-promoter regions, we show that reliable models can be built to identify core-promoter regions using a strictly computational approach. A major contribution to this field stems from our result that there is not a single class of core-promoters, but various distinct types exist. As a result, the existence of different classes of core-promoters should be taken into account when designing mathematical models to identify them.

Kobe.Florquin@psb.ugent.be Kobe Florquin

MicroArray Service at MAF

Paul Van Hummelen, Tom Bogaert, Kathleen Coddens, Kirsten DeSchouwer, Binita Dutta, Ruth Maes, Kurt Seeuws, Stefan Weckx
MicroArray Facility, VIB (www.microarrays.be)

VIB has an established core facility focused on the use of DNA chips (www.microarrays.be). A dedicated team of 8 people is devoted to conduct MicroArray experiments as a service to academia and industry. The Facility prides itself in giving personalized and complete services with follow-up help and advice. Since its inception in 1999, the MicroArray Facility ran over 2,000 arrays resulting in more than 27 million data-points and can hence count on a significant experience. The VIB MicroArray facility encourages the researchers to publish their MicroArray data in MIAME compliant databases. We support therefore submission of MicroArray experiments to Array Express (EMBL-EBI) and GEO (NCBI).
The MicroArray Facility has the ability to work with in-house produced DNA chips based on cDNAs and oligomeres. These arrays cover the full genomes of Arabidopsis, yeast and Salmonella; or partial genomes by 20,000 genes for human and mouse and smaller customized arrays for tobacco, Crohn's disease etc.
Next to the in-house arrays, the MicroArray Facility is a fully licensed and recognized service provider for 3 different commercial MicroArray platforms: Affymetrix (GeneChip®), Agilent Technologies (Agilent SurePrint), and Amersham Biosciences (CodeLink™).
In addition to the MicroArray service, the MicroArray Facility is also involved in several technology developments. Two major projects "Compendium of Arabidopsis Gene Expression (CAGE)" and "New strategy for the development of functional and performant starter cultures for foods in function of Food Qualitomics", were funded by the EU-FP5 and the Belgian IWT-Vlaanderen, respectively. In addition to these projects the MicroArray Facility also developed an own in-house CGH MicroArray to study duplication and deletions in the human genome (Graux et al., 2004; Van Buggenhout et al., 2004) and new systems to increase the hybridization sensitivity and hybridization speed by using shear-driven forces (Vanderhoeven et al., 2004).

paul.vanhummelen@vib.be Paul Van Hummelen

MODELING OF THE INTEGRASE BINDING

Arnout R.D. Voet1, Abel Jonckheer2, Frauke Christ3, Zeger Debyser3
1Laboratory for Bio-Molecular Modelling and 2Laboratory for Bio-Molecular Dynamics,

HIV-1 integrase catalyzes the integration of viral DNA into the host genome, a necessary step in the HIV-1 replication cycle. It is generally accepted that the virus employs cellular proteins to accomplish replication. Recently LEDGF/p75 was identified as a cellular cofactor of HIV-1 integration. We have modeled the structure of the integrase binding domain (IBD) of LEDGF/p75 by using threading, a comparative protein modeling method. The models showed an arrangement of 4 long helices burying the hydrophobic residues in the core of the structure. One model contained a fifth small helix in the loop connecting the second and the third helix. The topology of the IBD models is common to transcriptional proteins. These IBD models were docked to a tetrameric full length HIV-1 integrase model and an HIV-1 integrase X-ray structure lacking the C-terminal domain. Analysis of the docked results revealed an important HIV-1 IN interaction area on the IBD surface (residues 363-370 and 404-410). Validation of the interface residues and the putative models by site-directed mutagenesis in HIV-1 integrase was initiated.

arnout.voet@fys.kuleuven.ac.be Arnout Voet

Modelling clinical, ultrasound, microarray and proteomic data with Bayesian networks to study ovarian masses

Olivier Gevaert, Frank De Smet, Yves Moreau, Bart De Moor, Dirk Timmerman
ESAT-Sista, Katholieke universiteit Leuven, Kasteelpark Arenberg 10, 3000 Leuven

Ovarian cancer does not occur very frequently (only 4% of cancers among women) but is mostly diagnosed at a late stage because of the lack of symptoms at an early stage. Therefore ovarian cancer is often called "the silent killer" and clinical management of this disease is complicated by several factors. Firstly, it is difficult to distinguish benign and malignant ovarian masses before surgery based solely on clinical and ultrasound data. A correct diagnosis still largely depends on the experience of the clinician. This distinction is important because there is a favourable effect on the prognosis if the patient is directed immediately to an ovarian cancer specialist when diagnosed with an ovarian malignancy. Secondly, in the case of early stage disease it is difficult to foresee if the tumor will recur or not after surgery. An option in early stage disease is to give adjuvant therapy after surgery (e.g., chemotherapy) although it is known that some patients would not benefit from it (since they will never have a relapse without adjuvant therapy and are cured whatsoever) and would only be subjected to unnecessary toxicity. At this moment however, no reliable clinical parameters are available that can predict recurrence in early stage disease. A third issue concerns treatment in advanced stage disease (FIGO stage III or IV). Some tumors will prove to be resistant to platin based chemotherapy (platin resistance). Correct prediction of platin resistance would enable medical doctors to supply patients suffering from this disease with realistic information about their prognosis and enable optimal management strategies.
The goal of this research project is to construct models that are able to discriminate benign from malignant masses before surgery, to predict recurrence after surgery in early stage disease and to predict the response to platin based chemotherapy in advanced stage disease. To achieve this, we will use Bayesian networks to model and combine different types of data: clinical, ultrasound, proteomic and microarray data. Proteomic data can result in a better pre-operative distinction between benign and malignant ovarian masses while both microarray and proteomic data can enhance the performance when predicting recurrence or the response to therapy. Note that we already applied Bayesian networks to distinguish pre-operatively between benign and malignant ovarian masses using clinical and ultrasound data from a multicenter study: the International Ovarian Tumor Analysis consortium (IOTA).
Bayesian networks are a marriage between probability theory and graph theory. A Bayesian network is a sparse way of writing down a joint probability distribution. They are white box models which is important when encountering medical problems. When confronted with a certain question they give a complete overview of the uncertainty of the answer. Moreover such models allow exploring the underlying mechanism that generated the data and discover new (possibly causal) connections between the variables. Additionally Bayesian networks allow to incorporate prior knowledge (e.g. expert information, relevant literature) into the model. This allows guiding model training in the right direction. The usage of prior knowledge is very important with small and medium sample sizes that often occur in medical problems.
Microarray data and proteomic data are high dimensional and require pre-processing steps to extract the features most relevant to the class distinction. Classical statistical techniques could be used to accomplish this task. The different data types will then be combined by learning separate Bayesian networks for each data type. Subsequently, these separate models will be combined to construct a complete Bayesian network describing all the variables from the different data types using a small data set consisting of patients for whom all the different data types are available (around 70 cases – these patients are usually sparse). Note that we could also incorporate prior knowledge to improve learning of the models for each data type and of the model with all data types. The final Bayesian network model will then be used to prospectively tackle the problems concerning the management of ovarian cancer.

olivier.gevaert@esat.kuleuven.ac.be Olivier Gevaert

Nonrandom divergence of gene expression following whole genome duplications in plants

Casneuf T, Raes J, De Bodt S, Maere S, Van de Peer Y
Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Ghent, Belgium

Recent genome projects have revealed a surprisingly large number of duplicated genes in eukaryotic genomes, many of which seem to have arisen from large-scale, or even genome-wide duplication events. Whole genome duplication is particularly prominent in plants, as it is believed that up to 80% of the angiosperms are ancient polyploids or paleopolyploids, including a large proportion of our most important crops. For over 100 years, gene and genome duplication have been considered important for adaptive radiations of species and has been linked to the origin of evolutionary novelties, as it provides a source of genetic material on which evolution can work. Despite the large amount of research done on this subject, the consequences that duplication brings about for a duplicated gene pair remain poorly understood. In the past, four possible fates have been suggested for a gene and its duplicate. Ohno put in 1970 that loss or /nonfunctionalisation/ is the most likely fate of a gene of a duplicated pair, wile in rare cases one of the two duplicates acquires a new function (/neofunctionalization/). S/ubfunctionalization/ of multifunctional genes in which both gene copies lose a complementary set of regulatory elements and thereby divide the ancestral gene's original functions, forms a third potential fate. A fourth fate is genetic redundancy, where both duplicate genes are kept – while they keep very similar functions - so as to provide the organism with robustness against harmful mutations.

Recently, retention of duplicated genes was shown to be dependant on their function (Maere et al., 2005). However, it is not known whether the function also directs the fate of the retained gene pairs. Here, by using microarray expression data to study the evolution of duplicated genes in /Arabidopsis thaliana/, we show that the mode of duplication, the function of the genes involved, and the time since duplication all play important roles in the divergence of gene expression. Surprisingly, duplicates that have been created by large-scale duplication events and that can still be found in duplicated regions on the genome have more correlated expression patterns than those that were created by small-scale duplications or than those that do no longer lie in duplicated segments. Despite the mode of duplication or the functional class a gene belongs to, the younger the gene pair, the more correlated their expression patterns will be. A strong bias in divergence of gene expression was observed towards gene function and the biological process genes are involved in. For instance, genes involved in protein modification, signal transduction, kinase activity, carbohydrate and drug transporter activity, transferase activity, nucleotide, oxygen and drug binding have expression patterns that considerably diverged after duplication. In contrast, proteins involved in conserved processes, such as photosynthesis, cell cycle, DNA metabolism, protein biosynthesis and energy pathways, and genes involved in response to stress diverged less significantly after duplication.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M. & Van de Peer, Y. (2005) Modeling gene and genome duplications in eukaryotes. Proc. Natl. Acad. Sci. USA (In Press)

ticas@psb.ugent.be Tineke Casneuf

novoSNP 3: Sequence annotated variation detection

Peter De Rijk, Jurgen Del-Favero
Department of Molecular Genetics VIB8, Applied Molecular Genomics Group, Flanders Interuniversity Institute for Biotechnology, University of Antwerp, Antwerpen, Belgium

Resequencing in different individuals followed by comparison of sequences traces is the golden standard for sequence variation discovery. We have automated the data analysis for this approach with the program novoSNP. The program detects both single nucleotide (SNP) as well as insertion-deletion variations (INDELs) by aligning the traces to a reference sequence and scoring each position for a number of parameters. novoSNP scored significantly better than PolyPhred and PolyBayes finding all variations in a typical mutation analysis data set and nearly all in a large genomic resequencing data set (Weckx et al.2005, Genome Research 15:436-442). Version 3 of novoSNP has had many improvements, ranging from an upgrade of the database backend to various interface enhancements. Configurable clipping parameters now allow the use of suboptimal sequences, at the expense of more false positives. The most important upgrade is the addition of sequence annotation: The reference sequences can be extensively annotated, indicating positions of genes or other regions of interest. Annotation can be imported from GFF, EMBL, Genbank files or by providing the program with an mRNA sequence. The annotations can be viewed, edited and filtered from within the program and exported in GFF or Genbank format. All information about the potential variations with regards to these annotations is generated automatically. The variation annotation includes information such as the kind of the variation (non-synonymous, synonymous, frame shift, intronic, UTR, etc.), type of change resulting from the variation, and the position of the variation in the mRNA and/or exon and/or protein.

Peter.DeRijk@ua.ac.be Peter De Rijk

OligoFaktory: A Web Portal for Interactive Oligonucleotide Design

Colas Schretter, Laurent Gatto and Michel C. Milinkovitch
Unit of Evolutionary Genetics, Free University of Brussels

OligoFaktory is a web-based set of applications for the design, on an arbitrary number of target sequences (such as ORFs), of high-quality oligonucleotides for PCR and micro-array experiments. The plugin-based architecture allows to easily embed new modules: i.e. automated amplicons determination or design methods. The user-centered web interface conforms to state-of-the-art webstandards. An XHTML form is associated with each tool to fetch parameters from the user as input. A unified presentation of results provides overviews with distribution charts and relative location bar graphs, as well as detailed features for each oligonucleotide. Input and output files conform to a common XML interchange file format to allow both automatic generation of input data, archiving, and post-processing of results. We also describe a lightweight queuing system that distributes the resource intensive calcutations among multiple hosts.

lgatto@ulb.ac.be Laurent Gatto

PlaNet, a network of European plant databases

Eric Bonnet, Stéphane Rombauts, Pierre Rouzé and Yves Van de Peer
University of Ghent / VIB

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 bioinformatics 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 utilize the network. The partners implement local data collections within their field of interest and and provide access to their databases via web services.
The connection between the individual resources is realized with BioMOBY, which provides architecture for the discovery and distribution of biological data through web services.
A total of 119 different web services are available at the moment. A service aggregator is also available on the main planet website () that allow the user to compile the output of several web services for a given gene of interest.
The PlaNet is a European funded project (QLRI-CT-2001-00006).
The PlaNet consortium is composed of the John Innes Centre (UK), Nottingham Arabidopsis Stock Centre (UK), Centro Nacional de Biotecnologia (Spain), Flanders Interuniversity Institute for Biotechtonoly (Belgium), Plant Research International (The Netherlands), Munich Information Centre for Protein Sequences (Germany), Genoplante-Info (France).

eric.bonnet@psb.ugent.be Eric Bonnet

qBASE: an Excel application for the management and automatic analysis of real-time quantitative PCR data

Jan Hellemans, Geert Mortier, Anne De Paepe, Frank Speleman, Jo Vandesompele
Center for Medical Genetics Ghent, Ghent University Hospital, Belgium

Gene expression analysis is becoming increasingly important in biological research and clinical decision making, with real-time quantitative PCR becoming the method of choice for expression profiling of selected genes. Maturation of chemistry and hardware has made the practical performance of real-time quantitative PCR measurements feasible for the majority of molecular and genetic labs. However, accurate and straightforward mathematical and statistical analysis of the raw data (cycle threshold values) as well as the management of growing data sets have become the major hurdles in gene expression analyses.
Since the software provided along with the different detection systems does not provide an adequate solution for these issues, we developed qBASE: an Excel tool for the management and automatic analysis of real-time quantitative PCR data. The qBASE Browser allows data storage and annotation while keeping track of all real-time PCR runs by hierarchically organizing data into projects > experiments > plates. It is compatible with the export files from many currently available PCR instruments and provides easy access to all your data, both raw and processed. The qBASE Analyzer contains an easy plate editor, performs quality control, converts Ct values into normalized and rescaled quantities with proper error propagation, and displays results both tabulated and in graphs. The program does not limit the number of samples, genes and replicates, and allows data from multiple runs to be combined and processed together. The possibility to use up to 5 reference genes allows reliable and robust normalization of gene expression levels. qBASE allows easy exchange of data between users, and exports tabulated data for further statistical analyses using dedicated software. The application is free for non-commercial use, and intends to be an open source project, in that other interested parties can write their own analysis or visualization plug-ins.
We believe that by facilitating data management and providing tools for automatic data analysis, qBASE addresses one of the major problems in doing real-time quantitative PCR based nucleic acid quantification.

Jan.Hellemans@UGent.be Jan Hellemans

Quality control and quality improvement of 'omics' data

Rachel I.M. van Haaften1, Arie van Erk1, Cristina Luceri2, Kaatje Lenaerts3, Magali Jaillard1, Chris T.A. Evelo1
1 Maastricht University, BiGCaT Bioinformatics UM & TU/e, 2 University of Florence, department of Pharmacology, 3 Maastricht University, department of Human Biology

Omics technology used for large scale measurements of gene expression at the mRNA (transcriptomics, e.g. microarrays) and protein level (proteomics, e.g. antibody arrays) in biomedical research is rapidly evolving. Studies using such technologies yield huge amounts of data which have to be analyzed in a correct way to eventually give some useful information about the physiological outcome of the experiment. The output of array experiments is often accepted as correct at the general level when the overall image and the individual spots appear to be of satisfactory quality.
Clustering is a useful method of detecting array signal irregularities in multiple array experiments. It is especially suitable for dealing with large data sets. We used it both for comparing the different arrays used in the same experimental group, as well as for evaluating the behavior of the individual genes over the arrays within and between the experimental groups. It enabled us to correct a number of local signal aberrations.
In parallel to this, we did an in depth quality control on both two color microarrays and on antibody arrays using two color dyes. For this quality control we used the analysis and visualization tool Spotfire DecisionSite to replot the raw and normalized data (of the intensity and background signals of both colors) back in a matrix that corresponds to the original array location.
After replotting of the arraydata in the original physical layout we often see bad parts of the array which can be recognized by non-random localization of the background signals or non-random localization of differentially expressed genes/ proteins in one part of the array. These irregularities can influence the outcome of the complete study which can result in disturbed expression profiles.
The cause of the abnormalities can be very divers e.g. bad spotting of the array (especially for home-spotted arrays), there can be a scratch or a hair on the array or the cause can be in the processing of the array i.e. washing, centrifugation or drying steps.
When such an abnormality is recognized by visualization of the results, you can decide to throw out part of the array and start over the normalization and analysis without the bad part. In some cases you can decide to choose another, more localized, method of normalization by which the problem can be solved.
In a first example, an antibody array study, we saw strong artifacts which may have been caused by centrifugation steps which lead us to improvements of the technology used.
The second example was a microarray study directed towards red wine antioxidant treatment. Replotting of the data showed that one spotted block contained mostly low quality spots. Throwing out the whole block removed some data that had barely survived original quality control. Clustering of the eight biological replicate arrays from this study showed a cluster of genes that strongly differed between the arrays. Pathway analysis of those differing genes seemed to indicate that the animals used had differences in the expression of cell adhesion and cell-cell communication genes. Visualization of the differing gene cluster on the array showed that all those genes were in fact spotted in just two blocks that showed a differing average fold change value. Pathway analysis of the blocks showed that these block actually contained most of the cell-cell communication genes and most of the cell-adhesion genes (in fact pointing to a bad design). We were able to normalize these blocks separately (they were not bad just different) which made the problem disappear. After that there were no longer any differences between parallel treated animals, and we were able to see the effects of the treatment (which were, as expected, related to anti-oxidant pathways).
It can be concluded that rigid quality control, using combinations of smart visualization, clustering and pathway analysis can show major problems that will lead to rejection of study results or alternatively can lead to improved data treatment and normalization thus enabling the detection of real biological results in stead of technical artifacts.

Rachel.vanHaaften@BIGCAT.unimaas.nl Rachel I.M. van Haaften

Simulating genetic networks made easy: network construction with simple building blocks

Steven Vercruysse, Martin Kuiper
Computational Biology Division, Department of Plant Systems Biology, VIB / Ghent University

We present SIM-PLEX, a genetic network simulator with a very intuitive interface by which a user can easily specify interactions as simple "if - then" statements. The simulator is based on a mathematical model that approximates gene regulatory interactions as acting in a switch-like manner (called Piecewise Linear Differential Equations).

As gathering data on interactions in genetic networks today is almost trivial, the need for efficient software tools to model and simulate the network behaviour is becoming increasingly important. E.g. in the study of the regulatory interaction network of cell cycle control, yeast-2-hybrid experiments, protein-DNA interaction array experiments, and transcript microarrays yield data that may supplement well-described interactions between core components. In a Systems Biology approach, the objective is to integrate these new pieces of knowledge into a functional model of the regulatory process. This typically involves mathematical modeling of a (piece of the) genetic network to simulate its behaviour, and then to use the simulation results to design the next experiments. However, due to a lack of precise parameter information the use of exact differential equations (as used in Novak et al. (2001)) in practice rarely is realistic or feasible. Moreover, the use of such complex equations also puts such exact mathematic modeling beyond the reach of researchers primarily trained in biology.

We present a simulator based on PLDEs, a simplifying mathematical model where genes 'turn each other on or off' like switches (see de Jong et al., 2004). This made it possible to build an intuitive interface by which the user can declare each interaction as a simple if-then statement. The result of a quantitative simulation can be presented in the form of a graph of product amounts varying over time. The overall shape of the plots and the relative product amounts can in fact be much more important than the exact values, although these can be tuned too when more exact measurements become available. In this way, our simulator can serve as a bridge between qualitative (states, relative values only) and quantitative (numerical values) modeling, whereas its intuitive interface may bridge a gap between computational modeling and wet-lab biologists.

stcru@psb.ugent.be Steven Vercruysse

Statistical Data Fusion to prioritize lists of genes

Bert Coessens, Stein Aerts, Diether Lambrechts, Yves Moreau, Bart De Moor
BioI@SCD (SISTA), Departement Elektrotechniek (ESAT), Katholieke Universiteit Leuven

In the field of linkage analysis and association studies researchers are often confronted with large lists of candidate disease genes, especially when investigating complex multigenic diseases. Investigating all possible candidate genes is a tedious and expensive task that can be alleviated by selecting for analysis only the most salient genes.

In this context we developped a method for the computational prioritisation of a list of candidate disease genes based on multiple heterogeneous information sources. It is a generic approach building on the assumption that a new candidate disease gene has `similar properties' as a set of other genes that represents a certain biological case (a disease, for instance). The training set is seen as a model for the biological case comprising multiple submodels, one for each information source. A list of candidate (test) genes is ranked according to the similarity with the training genes. The rankings according to the different information sources are then combined using order statistics to obtain an overall rank, hence the term `statistical data fusion'.

bert.coessens@esat.kuleuven.ac.be Bert Coessens

Statistical methods for detecting heterotachy

G. Baele, Y. Van de Peer, J. Raes, S. Vansteelandt
Department of Applied Mathematics and Computer Science, Department of Plant Systems Biology

The substitution rate of a given site in a molecule is (1) not always constant through time and (2) can differ for different phylogenetic groups. Such behaviour is called heterotachy or within-site rate variation. Indications for heterotachy have already been given in the 1970s, when Fitch demonstrated that substitutions in cytochrome c occur at different positions in fungi and metozoa (Fitch 1971). More recently, Germot and Philippe have shown that the number of variable positions can also be different between lineages (Germot et al 1999).

Uncovering those positions that are heterotachous (i.e. which evolve according to the principle of heterotachy) is important for several reasons. First, it is a useful aid in the study of functional constraints. Second, it may provide additional proof for recently developed models that take into account heterotachy or covarion(-like behaviour). Third, it may deliver new insights that can help in building even more accurate evolutionary models for the reconstruction of phylogenetic trees. Finally, testing whether a given alignment contains heterotachous positions is important because the presence of heterotachy in a set of related sequences may disturb a reliable inference of phylogenetic relationships.

In our study, we develop a statistical method (1) to detect heterotachous positions in a given alignment and (2) to detect which phylogenetic groups are primarily responsible for it. Our method is insensitive to the tree lengths of the monophyletic groups in our alignment. Furthermore, it acknowledges that thousands of positions are simultaneously being tested. This is done by incorporating powerful corrections for multiple testing.

Detecting heterotachous positions in an alignment by applying the standard or modified chi-square procedure to thousands of positions simultaneously, is problematic. Indeed, traditional methods that test each site separately by comparing with a p-value cutoff of 0.01 or 0.05, are likely to generate an abundance of false positive results (i.e. positions that are falsely identified as heterotachous) as a result of multiple testing. To control the proportion of false positive hits, we use recent methods from statistical genetics that have proved powerful in other settings. In particular, using ideas from Storey (Storey et al 2003) and Benjamini and Hochberg (Benjamini et al 1995), we design our testing procedure such that it controls the so-called False Discovery Rate, which is defined as the proportion of false positives among all positions that are declared heterotachous by our testing procedure.

We will apply our method to a large set of SSU rRNA sequences derived from many different eukaryotes and eukaryotic taxa. Having detected the heterotachous positions, we will study their locations and compare these between different monophyletic groups using the secondary structure of our sequence alignment. The results of our method will be interpreted along with the variability map(s) of the given sequence alignment, since the degree of general positional variability of a position cannot be inferred from the degree of heterotachy, and vice versa. In addition, we will examine the impact of heterotachous positions on the performance of tree reconstruction algorithms for our data.

References

W. M. Fitch. The non-identity of invariable positions in the cytochromes c of different species. Biochem. Genet. 5:231-241. 1971.
A. Germot, and H. Philippe. Critical analysis of eukaryotic phylogeny: a case study based on the HSP70 family. J. Eukaryot. Microbiol. 46:116-124. 1999.
J. D. Storey, and R. Tibshirani. Statistical significance for genomewide studies. PNAS 100(16):9440-9445. 2003.
Y. Benjamini, and Y. Hochberg. Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. J. R. Statist. Soc. B 57(1):289-300. 1995.

guy.baele@ugent.be Guy Baele

T-profiler: Scoring the activity of pre-defined groups of genes using gene expression data

André Boorsma, Barrett C. Foat, Daniel Vis, Frans Klis, Harmen J. Bussemaker
Swammerdam Institute for Life Sciences-Microbiology, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands

One of the key challenges in the analysis of gene expression data is how to relate the expression level of individual genes to the underlying transcriptional programs and cellular state. Here we describe T-profiler, a tool that uses the t-test to score changes in the average activity of pre-defined groups of genes. The gene groups are derived from Gene Ontology, ChIP-chip experiments, and consensus transcription factor binding motifs; by scoring entire chromosomes, T-profiler can also detect aneuploidy. If desired, an iterative procedure can be used to select a single, optimal representative from sets of overlapping gene groups, and a jack-knife procedure makes calculations more robust against outliers. T-profiler makes it possible to interpret microarray data in a way that is both intuitive and statistically rigorous, without the need to combine experiments or choose parameters. Currently, gene expression data from Saccharomyces cerevisiae and Candida albicans are supported.

boorsma@science.uva.nl Andre Boorsma

The poplar genome project

Lieven Sterck, Stephane Rombauts, Pierre Rouzé and Yves Van de Peer
Bioinformatics and Evolutionary Genomics, Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

Poplar trees have been used all over the world to produce a large variety of wood-based products such as timber, pulp, and paper. Besides their great economical value, poplars are also rapidly becoming the model organism for forest biology and tree biotechnology. They can be easily transformed and vegetatively propagated and have a rapid growth. Another major advantage of the poplars is that they have a modest genome size organised in 19 chromosomes. It is therefore not surprising that, in 2001, poplar was selected as the first woody plant to have its genome sequenced.
Now that the poplar genome is publicly available, genome-wide bioinformatics analyses of the genome sequence can take a start. These include different kinds of analyses such as uncovering remnants of old duplication events, gene and gene family evolution, identifying transposable elements, and so on.

Recent analyses based on a Ks-dating of paralogous pairs of unigenes (clustered ESTs) as well as genomic data (predicted gene-models) suggest that poplar, like Arabidopsis, is an ancient polyploid. Results also suggested that the rate of synonymous substitutions in poplar is quite different from all other model-organisms, probably due to its slower generation time. Preliminary analyses between poplar and Arabidopsis indicate that approximately 70% of the poplar genome is collineair (conserved in gene content and order) with one or more Arabidopsis regions.
Annotating and classification of poplar-specific transposable elements and subsequent masking with these transposable elements has lead to the observation that almost half of the genome is made up of (relicts of) transposons.

Comparing gene-family sizes and looking for lineage-specific gene family expansions between poplar and Arabidopsis sheds light on the physiological differences between poplar and Arabidopsis.

References:

Lieven Sterck, Stephane Rombauts, Stefan Jansson, Fredrik Sterky, Pierre Rouzé, & Yves Van de Peer.(2005) EST data suggest that poplar is an ancient polyploidy. New Phytol. In press.

lieven.sterck@psb.ugent.be lieven Sterck

Using Molecular Dynamics to Identify Proton Pathways in Protein Disulphide Isomerase

Marc F. Lensink, Andre H. Juffer, Shoshana J. Wodak, Lloyd W. Ruddock
SCMBB, Universite Libre de Bruxelles, Belgium; Dept. of Biochemistry and Biocenter Oulu, University of Oulu, Finland

A key step in the folding of a protein into its native state, is theformation of native disulphide bonds. Even though Protein DisulphideIsomerase (PDI) was one of the first catalysts of protein foldingfound, identified nearly 40 years ago, still many questions regardingits mechanism of action remain to be answered. The active site of thecatalytic a' domain of PDI shows the motif -WCGHC-. Diluting the fullreaction mechanism down to its most basic element identifies bothcysteines as being involved in a proton abstraction from thesubstrate.

We have performed extensive Molecular Dynamics (MD) simulations of thecatalytic a' domain of PDI, in both reduced and oxidized form, as wellas of several mutants proteins. The influence of the protein dynamicson the pKa's of individual residues shows a shift in reactivity fromthe one active cysteine to the other. The analysis of the resultingtrajectories identifies a conserved hydrogen bonding network, capableof transporting the abstracted proton from the core of the protein tothe water phase.

lensink@scmbb.ulb.ac.be Marc F. Lensink

wrappers4EMBOSS, a fast-and-easy way to integrate BLAST and other 3rd party software under EMBOSS

Guy Bottu and Martin Sarachu
Belgian EMBnet Node, Université Libre de Bruxelles, Boulevard du Triomphe, 1050 Bruxelles and AR.EMBnet, Instituto de Bioquímica y Biología Molecular, Universidad Nacional de La Plata, Calle 115 entre 49 y 50, 1900 La Plata, ARGENTINA

EMBOSS (1) is a freeware OpenSource package for sequence analysis. It contains a lot of programs and is constantly expanding, but of course it is not possible to (re)code all the algorithms and functions the users might need. Therefore people have turned to integrating 3rd party software. One way to do this is to take the original source code and modify ("embossify") it ; usually it are the input and output modules that are changed. A second way is to write an EMBOSS "wrapper" program, which launches the actual program. One advantage of this second approach is that at an upgrade of the underlying software one only needs to make minor changes to the wrapper code in order to adjust for changed parameters, if any. This can be done in less time than (re)embossifying the source code, which you would have to do even for small changes in the original source. Another advantage is that you can integrate 3rd party software for which distributing (a modified version of) the source code is not allowed because of a license or simply impossible because the code is not available.

wrappers4EMBOSS is the joint development of the Belgian and Argentinian EMBnet nodes. It is distributed as freeware together with wEMBOSS, a Web interface for EMBOSS, but they can be installed and used independently from each other. wrappers4EMBOSS contains an installation script, which detects the location of the underlying software (which should have been installed before), asks interactively a few questions (like which sequence databanks are available and where temporary files should be written), then modifies the configuration files and source code accordingly and recompiles EMBOSS. The advantage of this approach is that the software can be run easily, without a properly configured UNIX shell, under wEMBOSS or any other graphical user interface for EMBOSS. The current version of wrappers4EMBOSS supports the following software : - the BLAST software for searching a sequence against a databank by similarity (including PSI-BLAST, PHI-BLAST, 2 sequences alignment and the possibility to search a personal databank) - the fastA software for searching a sequence against a databank by similarity (including the fasts tool for identifying a protein from its peptides, 2 sequences alignment and the possibility to search a personal databank) - the CLUSTAL software for multiple sequence alignment and Neighbour-Joining phylogenetic tree building - a tool to search a protein against PROSITE motif databank - a tool to search a protein or nucleic acid against Blocks motif databank - an interface to the SRS software for searching databanks by keyword, which produces Lists with sequence names directly usable by EMBOSS

wrappers4EMBOSS can be downloaded from the wEMBOSS portal (2)

(1) "EMBOSS: The European Molecular Biology Open Software Suite". Rice, P., Longden, I. and Bleasby, A. Trends in Genetics 16(6), 276-277 (2000).
(2) http://www.wemboss.org

gbottu@ben.vub.ac.be Guy Bottu

Participant List

Aerts Stein
K.U.Leuven
Herestraat 49, bus 602
3000 Leuven
Belgium
stein.aerts@med.kuleuven.ac.be

Alard Philippe
Algonomics
Technologiepark
9058 Ghent
Belgium
philippe.alard@algonomics.com

Allemeersch Joke
K.U.Leuven
Kasteelpark Arenberg 10
3000 Leuven
Belgium
joke.allemeersch@esat.kuleuven.ac.be

Ameloot Paul
Dep Molecular biomedical research
Technologiepark
9052 Ghent
Belgium
paul.ameloot@dmbr.ugent.be

Antezana Erick
PSB - Computational Biology Division
Roodebeeksesteeweg 83
1200 Brussels
Belgium
erant@psb.ugent.be

Aurélie Defferard
Bayer BioScience
Technologiepark 38
9052 Ghent
Belgium
aurelie.defferrard@bayercropscience.com

Ayoubi Torik
Maastricht University
Universiteitssingel 50
6200 MD Maastricht
The Netherlands
torik.ayoubi@gen.unimaas.nl

Baele Guy
UGent
Meerstraat 54
9860 Oosterzele
Belgium
guy.baele@ugent.be

Baurain Denis
University of Liège
Department of Life Sciences, Institute of Botany B22
B-4000 Liège
Belgium
denis.baurain@ulg.ac.be

Bekaert Sofie
Ghent University
Coupure Links 653
B-9000 Ghent
Belgium
sofie.bekaert@UGent.be

Berger Fabrice
FUNDP (URBM)
Av. de Bruxelles, 61
5000 Namur
Belgium
f.berger@opentech.be

Blom Evert-jan
Molecular Genetics
Kerklaan 30
9751 NN Haren
The Netherlands
e.j.blom@rug.nl

Blomme Tine
BioInformatics & Evolutionary Genomics - VIB/UGENT
technologiepark 927
9000 Ghent
Belgium
tiblo@psb.ugent.be

Boekhorst Jos
CMBI, Radboud University Nijmegen
Toernooiveld 1
6525ED Nijmegen
The Netherlands
J.Boekhorst@cmbi.ru.nl

Bonnet Eric
BioInformatics & Evolutionary Genomics - VIB/UGENT
Technologiepark 927
9052 Ghent
Belgium
eric.bonnet@psb.ugent.be

Bonvin Alexandre
Bijvoet Center fo Biomolecular Research, Utrecht University
Padualaan 8
3584 CH Utrecht
The Netherlands
a.m.j.j.bonvin@chem.uu.nl

Boorsma André
University of Amsterdam
Nieuwe achtergracht 166
1018 WV Amsterdam
The Netherlands
boorsma@science.uva.nl

Bottu Guy
Univesrsité Libre de Bruxelles, Belgian EMBnet Node
U.L.B. CP257, campus de la Plaine, blv. du Triomphe
1050 Brussels
Belgium
gbottu@ben.vub.ac.be

Bruynseels Koen
Cropdesign NV
Technologiepark 3
9052 Ghent
Belgium
koen.bruynseels@cropdesign.com

Casneuf Tineke
BioInformatics & Evolutionary Genomics - VIB/UGENT
Technologiepark 927
9052 Zwijnaarde
Belgium
ticas@psb.ugent.be

Coenye Tom
Ghent University
Ledgeanckstraat 35
9000 Ghent
Belgium
Tom.Coenye@UGent.be

Coessens Bert
K.U.Leuven
Kasteelpark Arenberg 10
3001 Leuven - Heverlee
Belgium
bert.coessens@esat.kuleuven.ac.be

Coornaert David
Univesrsité Libre de Bruxelles, BEN
12 r des Profs Jeener & Brachet
6041 Gosselies
Belgium
dcoorna@dbm.ulb.ac.be

Couche Fabian
Univesrsité Libre de Bruxelles, SCMBB
Avenue du Boulevard du Triomphe - CP 263
1050 Brussels
Belgium
fcouche@scmbb.ulb.ac.be

Cuelenaere Koen
Dalicon BV
PO box 354
6700 AJ Wageningen
Netherlands
koenc@dalicon.com

De Bleser Pieter
Department of Molecular Biomedical Research VIB-Ghent Universi
Technologiepark 927
B-9052 Zwijnaarde
Belgium
pieterdb@dmbr.ugent.be

De Bodt Stefanie
BioInformatics & Evolutionary Genomics - VIB/UGENT
Technologiepark 927
9052 Ghent
Belgium
stefanie.debodt@psb.ugent.be

De Jong Anne
University of Groningen
Kerklaan 30
9751 NN Haren
The Netherlands
Anne.de.Jong@rug.nl

De Keyzer Simon
Univesrsité Libre de Bruxelles, SCMBB
Bd. du Triomphe - CP 263
1050 Brussels
Belgium
simon@scmbb.ulb.ac.be

De Laet Jan
Univesrsité Libre de Bruxelles, Belgian Biodiversity Platform
ULB CP257 2 N4 115C
1050 Brussels
Belgium
jdelaet@naturalsciences.be

De Maeyer Marc
K.U.Leuven
Lab. Biomoleculaire modellering, Celestijnenlaan 200D
B-3001 Leuven - Heverlee
Belgium
marc.demaeyer@fys.kuleuven.ac.be

De Preter Katleen
Center for Medical Genetics
De Pintelaan 185, MRB 2nd floor
9000 Ghent
Belgium
katleen.depreter@ugent.be

De Rijk Peter
VIB8 / UA
Universiteitsplein 1
2610 Antwerpen
Belgium
Peter.DeRijk@ua.ac.be

De Smet Frank
K.U.Leuven, ESAT-SCD
Kasteelpark Arenberg 10
3001 Leuven - Heverlee
Belgium
frank.desmet@esat.kuleuven.ac.be

De Wit Vanessa
K.U.Leuven
NAAMSESTRAAT 59
3000 Leuven
Belgium
Vanessa.DeWit@bio.kuleuven.ac.be

Degroeve Sven
Universiteit Gent
Coupure Links
9000 Ghent
Belgium
sven.degroeve@ugent.be

Dehertogh Benoit
FUNDP (URBM)
Av. de Bruxelles, 61
5000 Namur
Belgium
benoit.dehertogh@fundp.ac.be

Depiereux Eric
FUNDP
61 rue de Bruxelles
5000 Namur
Belgium
eric.depiereux@fundp.ac.be

Derix Ton
unimaas
Universteitssingel 50
6200 MD maastricht
Belgium
ton.derix@gen.unimaas.nl

Dessailly Benoit
Service de Conformation des Macromolécules Biologiques, Univers
CP-263, Campus La Plaine (ULB), Bld du Triomphe
1050 Ixelles
Belgium
benoit@scmbb.ulb.ac.be

Dorca Fornell Carmen
VIB
Krijgslaan 250
B-9000 Ghent
Belgium
carmendorca@hotmail.com

Dugas Olivier
Univesrsité Libre de Bruxelles
70 rue Meurein
59000 Lille
France
odugas@dbm.ulb.ac.be

Eijssen Lars
Universiteit Maastricht
Universiteitssingel 50
6229ER Maastricht
The Netherlands
lars.eijssen@gen.unimaas.nl

Engelen Kristof
K.U.Leuven, BIO@SCD
Kasteelpark Arenberg 10
3001 Leuven - Heverlee
Belgium
kristof.engelen@esat.kuleuven.ac.be

Fadda Abeer
K.U.Leuven
Vlamingenstraat 87, box 305
3000 Leuven
Belgium
abeer.fadda@student.kuleuven.ac.be

Fahmi Saleh
Cairo University- Faculty of Medicine
116 el hegaz st, heliopolis
0000 Cairo
Egypt
smfahmi@hotmail.com

Fays Frédéric
Univesrsité Libre de Bruxelles, SCMBB
Boulevard du Triomphe - CP 263
1050 Brussels
Belgium
frederic@scmbb.ulb.ac.be

Fiers Mark
Plant Research International
PO Box 16
6700 AA Wageningen
The Netherlands
mark.fiers@wur.nl

Florquin Kobe
BioInformatics & Evolutionary Genomics - VIB/UGENT
Technologiepark 927
B-9052 Ghent
Belgium
kobe.florquin@pandora.be

Gatto Laurent
Unit of Evolutionary Genetics
rue Jeener et Brachet, 12
6041 Gosselies
Belgium
lgatto@ulb.ac.be

Gevaert Olivier
K.U.Leuven
Kasteelpark Arenberg 10
3000 Leuven
Belgium
olivier.gevaert@esat.kuleuven.ac.be

Givi Omid
Hanze University Groningen
Zernikeplein 11
9747 AS Groningen
The Netherlands
o.givi@pl.hanze.nl

Glassee Wim
UA
Universiteitsplein 1
2610 Antwerpen
Belgium
wim.glassee@ua.ac.be

Glenisson Patrick
ESAT
Kasteelpark arenberg 10
B-3001 Leuven
Belgium
patrick.glenisson@esat.kuleuven.ac.be

Gonze Didier
Univesrsité Libre de Bruxelles
Bvd Triomphe, Campus Plaine
1050 Brussels
Belgium
dgonze@ulb.ac.be

Guillaume Karen
VUB
Pleinlaan 2
1050 Brussels
Belgium
kguillau@vub.ac.be

Harding Boris
UA
universiteitsplein 1
2610 Antwerpen
Belgium
boris.harding@ua.ac.be

Hasan Shariful
Computer's Ozone
Plot # 3 Muslim Colony near FTC building Shahra e Faisalu
74400 Karachi
Pakistan
sharif_danish@yahoo.com

Hassan Anerhour
Univesrsité Libre de Bruxelles
Bld du Triomphe
1050 Brussels
Belgium
hassan@amaze.ulb.ac.be

Hellemans Jan
Center for Medical Genetics Ghent, Ghent University Hospital
De Pintelaan 185, MRB
9000 Ghent
Belgium
Jan.Hellemans

Hermans Filip
ugent
technologiepark 927
9052 Ghent
Belgium
fiher@psb.ugent.be

Herzog Robert
Univesrsité Libre de Bruxelles, BEN
Rue Prof. Jeener&Brachet 12
6041 Gosselies
Belgium
rherzog@ulb.ac.be

Hubaut Olivier
Univesrsité Libre de Bruxelles, SCMBB
Bvd du Triomphe - CP 263
1050 Brussels
Belgium
olivier@scmbb.ulb.ac.be

Jaillard Magali
BiGCaT
Universiteit Maastricht (UNS50, Box 18) - P.O. Box 616
6200 MD Maastrich
Netherlands
magali.jaillard@bigcat.unimaas.nl

Janky Rekin's
Univesrsité Libre de Bruxelles, SCMBB
Service de Conformation de Macromolecules Biologiques et de Bioi
1050 Brussels
Belgium
rekins@scmbb.ulb.ac.be

Janssen Paul
SCK-CEN
Boeretang 200
2400 Mol
Belgium
pjanssen@sckcen.be

Joosten Henk-jan
WUR
dreijenlaan 2
6703HA Wageningen
The Netherlands
henk-jan.joosten@wur.nl

Kamuzinzi Richard
Univesrsité Libre de Bruxelles
Rue des prof. Jeener et Brachet 12
1410 Waterloo
Belgium
rkamuz@dbm.ulb.ac.be

Krols Luc
Peakadilly N.V.
BioIncubator, Technologiepark 4
9052 Zwijnaarde
Belgium
Luc.Krols@peakadilly.com

Kuiper Martin
VIB-PSB
Technologiepark 927
9052 Zwijnaarde
Belgium
makui@psb.ugent.be

Kupper Cardozo Alessandra
Univesrsité Libre de Bruxelles
Route de Lennik, 808 - CP 618
1070 Brussels
Belgium
akupperc@ulb.ac.be

Kuramae Eiko
Centraalbureau voor Schimmelcultures
Uppsalalaan 8
3584 CT Utrecht
The Netherlands
kuramae@cbs.knaw.nl

Kwasnikowska Natalia
Limburgs Universitair Centrum
Ottergemsesteenweg 272
9000 Ghent
Belgium
natalia.kwasnikowska@luc.ac.be

Ledent Valérie
Univesrsité Libre de Bruxelles
Rue des prof. Jeener et Brachet 12
1410 Waterloo
Belgium
Valerie.Ledent@ulb.ac.be

Lemer Christian
SCMBB
bd du Triomphe - CP 263
1050 Brussels
Belgium
chris@amaze.ulb.ac.be

Lemmens Karen
K.U.Leuven, ESAT-SCD
Kasteelpark Arenberg 10
3001 Leuven - Heverlee
Belgium
karen.lemmens@esat.kuleuven.ac.be

Lens Pierre
Bayer CropScience
Technologiepark 38
9052 Ghent
Belgium
pierre.lens@bayercropscience.com

Lensink Marc
Univesrsité Libre de Bruxelles, SCMBB
Bd du Triomphe - CP 263
1050 Brussels
Belgium
lensink@scmbb.ulb.ac.be

Leunissen Jack
Wageningen Universiteit
Dreijenlaan 3
6703 HA Wageningen
The Netherlands
jack.leunissen@wur.nl

Lima Gipsi
SCMBB-ULB
Boulevard du Triomphe
1050 Brussels
Belgium
gipsi@scmbb.ulb.ac.be

Lindsey Patrick
University of Maastricht
Universiteitssingel 50, UNS 50 , postvak 16, Postbus 616
6200 MD Maastricht
The Netherlands
patrick.lindsey@gen.unimaas.nl

Liu Feng
Limburg University
Diepenbeek
3590 Diepenbeek
Belgium
feng.liu@luc.ac.be

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

Maere Steven
UGent/VIB
Technologiepark 927
9052 Zwijnaarde
Belgium
stmae@psb.ugent.be

Maniraja Jesintha Mary
Univesrsité Libre de Bruxelles
SCMBB, CP 263, Blvd Triomphe
1050 Brussels
Belgium
jesintha@amaze.ulb.ac.be

Marchal Kathleen
K.U.Leuven, CMPG/ bioinformatics
KasteelPark Arenberg 20
3001 Leuven
Belgium
kathleen.marchal@biw.kuleuven.be

Maree Raphael
University of Liège
Methodes Stochastiques, Montefiore Institute (B28),
Grande Trave
4000 Liège
Belgium
raphael.maree@ulg.ac.be

Martens Cindy
Ghent University
Technologiepark 927
9052 Zwijnaarde
Belgium
cimar@psb.ugent.be

Maurer-stroh Sebastian
IMP Institute of Molecular Pathology
Dr.Bohrgasse 7
1030 Vienna
Austria
stroh@imp.univie.ac.at

Mátrai Janka
K.U.Leuven
Celestijnenlaan 200D
3001 Leuven - Heverlee
Belgium
janka.matrai@fys.kuleuven.ac.be

Méndez Raúl
Univesrsité Libre de Bruxelles, SCMBB
Université Libre de Bruxelles. Campus Plaine, Bvd. du Triomphe-
1050 Brussels
Belgium
raul@scmbb.ulb.ac.be

Meeus Jeroen
University of Ghent / VIB
Department of Plant Systems Biology, Technologie park 927
9052 Ghent
Belgium
jeroen.meeus@ugent.be

Menten Björn
Center for Medical Genetics, Ghent University Hospital
De Pintelaan 185
9000 Ghent
Belgium
bjorn.menten@ugent.be

Michoel Tom
PSB, VIB/UGent
Technologiepark 927
9052 Ghent
Belgium
tom.michoel@psb.ugent.be

Moreau Yves
K.U.Leuven
Kasteelpark Arenberg 10
B-3001 Leuven
Belgium
moreau@esat.kuleuven.ac.be

Muratet Michael
University of Alabama/Huntsville
308 Sparkman Dr
35899 Huntsville
United States
mimur@psb.ugent.be

Nakatani Andreia
Centraalbureau voor Schimmelcultures
Uppsalalaan,8
3584CT Utrecht
The Netherlands
aknakata@cbs.knaw.nl

Nap Jan-peter
Hanzehogeschool Groningen
PO Box 3037
9701 DA Groningen
The Netherlands
j.p.h.nap@pl.hanze.nl

Nathanson Jason
Univ Calif San Diego
10010 N Torrey Pines Rd
92037 La Jolla
United States
nathanson@salk.edu

Nooren Irene
Devgen
Technologiepark 30
9000 Ghent
Belgium
Irene.Nooren@devgen.com

Nuytinck Jorinde
Univesrsité Libre de Bruxelles,Belgisch Biodiversiteitsplatform
cp 257- Boulevard du Triomphe- Campus Plaine-building NO
1050 Brussels
Belgium
jnuytinc@ulb.ac.be

Ogao Patrick
University of Groningen, Mathematics and cComputing Science Depa
P.O. box 800
9700AV Groningen
The Netherlands
ogao@cs.rug.nl

Pattyn Filip
Ghent University
De Pintelaan 185
9000 Ghent
Belgium
Filip.Pattyn@UGent.be

Perez-rueda Ernesto
IBT-UNAM
Av. Universidad 1001
62210 Cuernavaca
Mexico
erueda@ibt.unam.mx

Pierre Geurts
University of Liège
Institut Montefiore, Sart Tilman B28
B4000 Liège
Belgium
p.geurts@ulg.ac.be

Pletinckx Jurgen
AlgoNomics
Technologiepark 4
9052 Zwijnaarde
Belgium
jurgen.pletinckx@algonomics.com

Pochet Nathalie
K.U.Leuven, BioI, ESAT-SCD
Kasteelpark Arenberg 10
3001 Leuven - Heverlee
Belgium
nathalie.pochet@esat.kuleuven.ac.be

Posada Esmeralda
Bayer Cropscience
Wolterslaan 195
9040 Ghent
Belgium
elpos38@hotmail.com

Raes Jeroen
BioInformatics & Evolutionary Genomics - VIB/UGENT
Technologiepark 927
9052 Ghent
Belgium
jerae@psb.ugent.be

Ren Xin-ying
Plant Research International
Droevendaalsesteeg 1
6708PB Wageningen
The Netherlands
xinying.ren@wur.nl

Renwart Benjamin
University of Liège
Institut Montefiore, Sart Tilman B28
B4000 Liège
Belgium
ben.renwart@skynet.be

Rigali Sébastien
University of Liège
Institut de chimie B6a
B-4000 Liège
Belgium
srigali@ulg.ac.be

Rigali Sébastien
University of Liège
Institut de Chimie, B6a
B-4000 Liège
Belgium
srigali@ulg.ac.be

Roest Mark
VORtech Computing
P.O.Box 260
2600 AG Delft
The Netherlands
roest@vortech.nl

Russo Christophe
FUNDP (URBM)
Av. de Bruxelles, 61
5000 Namur
Belgium
crusso@urbm.fundp.ac.be

Sand Olivier
Univesrsité Libre de Bruxelles
Bvd du Triomphe CP 263
1050 Brussels
Belgium
oly@scmbb.ulb.ac.be

Sclep Gert
VIB
Technologiepark 927
9052 Zwijnaarde
Belgium
Gert.Sclep@psb.ugent.be

Shahalizadeh Solmaz
University of Tehran
No 234.Fazle Elahi Alley(38th),Farhang Sqr.Seyed Jamaledin St.
1436683578 Tehran
Iran
solmaz.sh@gmail.com

Simillion Cedric
BioInformatics & Evolutionary Genomics - VIB/UGENT
Technologiepark 927
9000 Ghent
Belgium
cedric.simillion@psb.ugent.be

Robbens Steven
BioInformatics & Evolutionary Genomics - VIB/UGENT
Technologiepark 927
9052 Ghent
Belgium
steven.robbens@psb.ugent.be

Rombauts Stephane
BioInformatics & Evolutionary Genomics - VIB/UGENT
Technologiepark 927
9000 Ghent
Belgium
strom@psb.ugent.be

Saeys Yvan
BioInformatics & Evolutionary Genomics - VIB/UGENT
Technologiepark 927
9052 Ghent
Belgium
yvan.saeys@psb.ugent.be

Schrevens Eddie
K.U.Leuven
Willem de Croylaan 42
3001 Leuven - Heverlee
Belgium
eddie.schrevens@agr.kuleuven.ac.be

Semay Elke
NetApp
Grétrystraat 22
2018 Antwerpen
Belgium
elke@netapp.com

Sheng Qizheng
Department of Electrical Engineering, ESAT-SCD, Katholieke Unive
Kasteelpark Arenberg 10
3001 Leuven - Heverlee
Belgium
qizheng.sheng@esat.kuleuven.ac.be

Sirota Fernanda L
Univesrsité Libre de Bruxelles, SCMBB
Boulevard du Triomphe - CP 263
1050 Brussels
Belgium
fernanda@scmbb.ulb.ac.be

Stassen Fons
AZM/Genome Center Maastricht
Universiteitssingel 50, postvak 16
6200 MD Maastricht
The Netherlands
fons.stassen@gen.unimaas.nl

Sterck Lieven
BioInformatics & Evolutionary Genomics - VIB/UGENT
technologiepark 927
9000 Ghent
Belgium
lieven.sterck@psb.ugent.be

Szikora Jean-pierre
UCL
74, av Hippocrate - UCL 7459
1200 Brussels
Belgium
szikora@icp.ucl.ac.be

Thomas Gregoire
Peakadilly nv
Technologiepark 4
9052 Zwijnaarde
Belgium
gregoire.thomas@peakadilly.com

Thomas-chollier Morgane
VUB - U.L.B.
Rue des prof. Jeener et Brachet 12
1410 Waterloo
Belgium
mthomas@dbm.ulb.ac.be

Todt Tilman
HAN
Laan van Scheut 2
6503 CG Nijmegen
The Netherlands
tdt@ft.han.nl

Tran Joseph
Univesrsité Libre de Bruxelles, SCMBB
SCMBB - Université Libre de Bruxelles. Campus Plaine. CP 263. B
1050 Brussels
Belgium
jtran@scmbb.ulb.ac.be

Tsiporkova Elena
VIB / University of Ghent
Department of Plant Systems Biology, Technologie park 927
9052 Ghent
Belgium
elena.tsiporkova@psb.ugent.be

Turatsinze Jean Valery
Univesrsité Libre de Bruxelles
Campus Plaine. Boulevard du Triomphe, Université Libre de Buxel
1050 Brussels
Belgium
jturatsi@ulb.ac.be

Tylzanowski Przemko
K.U.Leuven
Herestraat 49
3000 Leuven
Belgium
przemko@med.kuleuven.ac.be

Van Aelst Stefan
UGENT
Krijgslaan 281 S9
B-9000 Ghent
Belgium
Stefan.VanAelst@UGent.be

Van Breukelen Dr. Ir. Bas
Utrecht University
Sorbonnelaan 16
3584 CA Utrecht
The Netherlands
b.vanbreukelen@pharm.uu.nl

Van Daelen Raymond
Keygene N.V.
P.O. Box 216
6700 AE Wageningen
The Netherlands
Raymond.van-daelen@keygene.com

Van De Peer Yves
Departement of Plant Systems Biology, Ghent University
Technologiepark 927
9052 Ghent
Belgium
yves.vandepeer@psb.ugent.be

Van De Plas Raf
K.U.Leuven
Kasteelpark Arenberg 10
B-3001 Leuven - Heverlee
Belgium
raf.vandeplas@esat.kuleuven.ac.be

Van Delm Wouter
K.U.Leuven, ESAT/SCD
Kasteelpark Arenberg 10
, B-3001 Leuven - Heverlee
Belgium
wvandelm@esat.kuleuven.ac.be

Van Den Bulcke Tim
K.U.Leuven, ESAT - SCD
Kasteelpark Arenberg 10
3001 Leuven - Heverlee
Belgium
tim.vandenbulcke@esat.kuleuven.ac.be

Van Der Burgt Ate
Plant Research International
Droevendaalsesteeg 1
6708 PB Wageningen
The Netherlands
ate.vanderburgt@wur.nl

Van Der Weken Dietrich
Ghent University
Krijgslaan 281
9000 Ghent
Belgium
dietrich.vanderweken@ugent.be

Van Dijk Marc
Department of NMR spectroscopy, Utrecht University
Padualaan 8
3584 CH Utrecht
The Netherlands
mvdijk@nmr.chem.uu.nl

Van Dijk Aalt-jan
Utrecht University
Padualaan 8
3584 CH Utrecht
The Netherlands
a.j.vandijk@chem.uu.nl

Van Droogenbroeck Bart
VIB2
Technologiepark 927
9052 Zwijnaarde
Belgium
badro@psb.ugent

Van Durme Joost
IRIBHM, ULB
Lenniksebaan 808
1070 Brussel
Belgium
jvdurme@ulb.ac.be

Van Eijsden Rudy
Maastricht University, Dep. of Genetics & Cell Biology
Universiteitssingel 50
6229 ER Maastricht
The Netherlands
rudy.vaneijsden@gen.unimaas.nl

Van Ham Roeland
Plant Research International
Droevendaalsesteeg 1
6708 PB Wageningen
The Netherlands
roeland.vanham@wur.nl

Van Handenhove Jack
Bayer CropScience
Techologiepark 38
9052 Ghent
Belgium
jack.vanhandenhove@bayercropscience.com

Van Hellemont Ruth
K.U.Leuven
kasteelpark Arenberg 10
3001 Leuven
Belgium
ruth.vanhellemont@esat.kuleuven.ac.be

Van Hijum Sacha
Department of Molecular Genetics; University of Groningen
Kerklaan 30
9751 NN Haren
The Netherlands
s.a.f.t.van.hijum@rug.nl

Van Hummelen Paul
MicroArray Facility, VIB
UZ Gasthuisberg, Herestraat 49
3000 Leuven
Belgium
paul.vanhummelen@vib.be

Van Kampen Antoine
Academic Medical Center
Meibergdreef 9
1105 AZ Amsterdam
The Netherlands
a.h.vankampen@amc.uva.nl

Van Leemput Koenraad
Universiteit Antwerpen
Middelheimlaan 1
B-2020 Antwerpen
Belgium
koen.vanleemput@ua.ac.be

Van Loo Peter
K.U.Leuven
Herestraat 49, bus 602
3000 Leuven
Belgium
Peter.VanLoo@med.kuleuven.ac.be

Van Oeveren Jan
Keygene N.V.
P.O. Box 216
6700 AE Wageningen
The Netherlands
Jan.van-oeveren@keygene.com

Van Roy Nadine
Dept. Medical Genetics, UGent
de pintelaan 185
9000 Ghent
Belgium
nadine.vanroy@UGent.be

Van Vooren Steven
K.U.Leuven
Kasteelpark Arenberg 10
3001 Leuven - Heverlee
Belgium
Steven.VanVooren@esat.kuleuven.ac.be

Van Walle Ivo
AlgoNomics
Technologiepark 4
9052 Zwijnaarde
Belgium
ivo.van.walle@algonomics.com

Van Wiemeersch Luc
PSB
Technologypark
9052 Ghent
Belgium
luwie@psb.ugent.be

Van Yper Stefan
Ghent University
Coupure Links 653
9000 Ghent
Belgium
Stefan@biomath.UGent.be

Vandepoele Klaas
BioInformatics & Evolutionary Genomics - VIB/UGENT
Technologiepark 927
9000 Ghent
Belgium
klaas.vandepoele@psb.ugent.be

Vandesompele Jo
Ghent University Hospital
De Pintelaan 185
9000 Ghent
Belgium
joke.vandesompele@ugent.be

Vercruysse Steven
Plant Systems Biology / UGent
Technologiepark 927
9052 Ghent
Belgium
stcru@psb.ugent.be

Vernaillen Frank

Molenstraat 25
9550 Herzele
Belgium
frank.vernaillen@telenet.be

Verstegen Harold
Keygene N.V.
P.O. Box 216
6700 AE Wageningen
The Netherlands
harold.verstegen@keygene.com

Veugelers Mark
VIB
Rijvisschestraat 120
9052 Ghent
Belgium
mark.veugelers@vib.be

Vlieghe Dominique
DMBR VIB/UGhent
Technologiepark 927
B-9052 Ghent
Belgium
dominique.vlieghe@dmbr.ugent.be

Vlietinck Robert
university of Maastricht
P.O. Box 616 #16
6200 MD Maastricht
The Netherlands
robert.vlietinck@med.kuleuven.ac.be

Voet Arnout
K.U.Leuven, Laboratory for biomolecular modelling.
Celestijnenlaan 200d
3001 Leuven - Heverlee
Belgium
arnout.voet@fys.kuleuven.ac.be

Voorbraak Frans
Academic Medical Center
Meibergdreef 15
105AZ Amsterdam
The Netherlands
f.p.voorbraak@amc.uva.nl

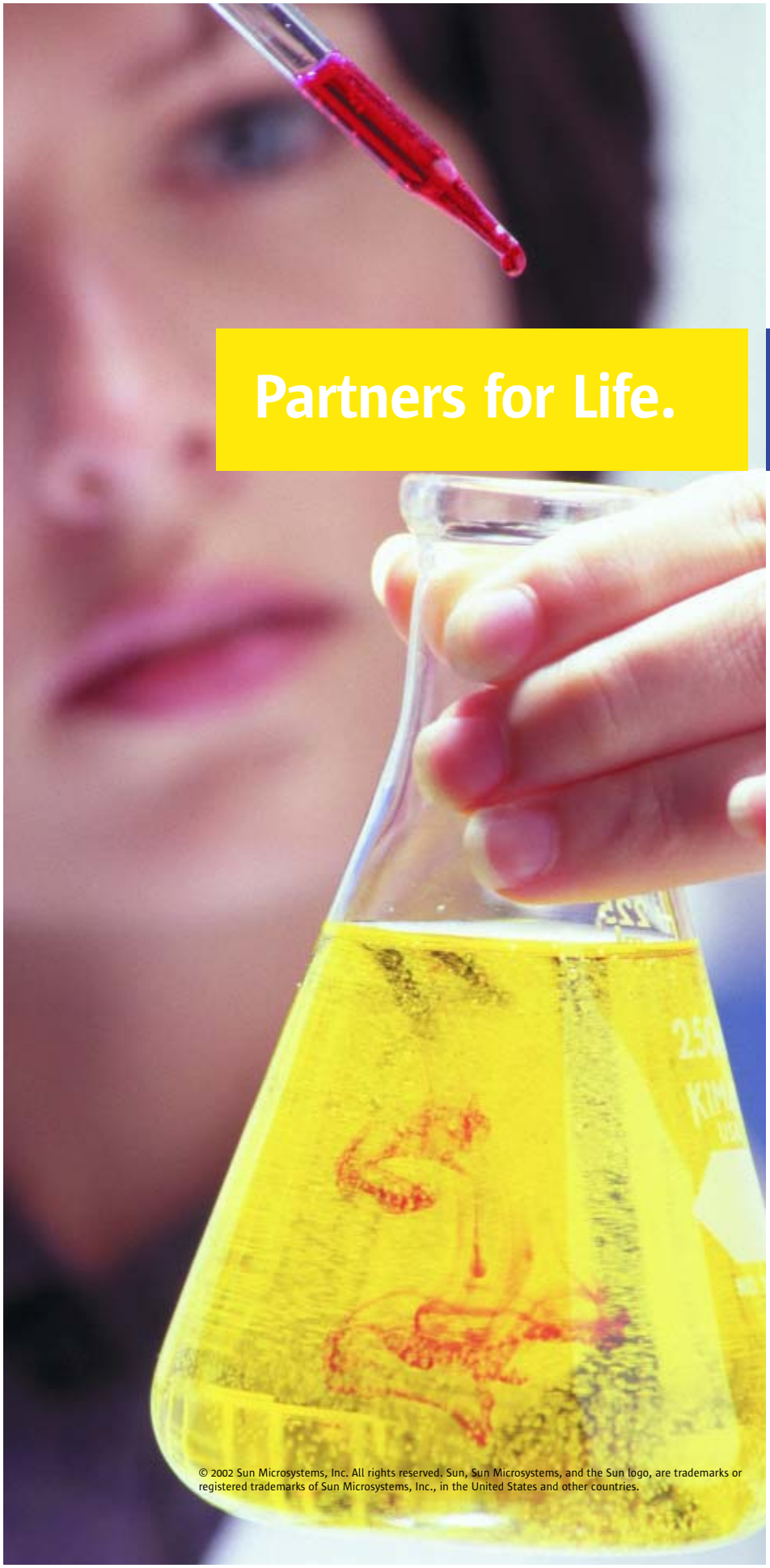
Warmoes Marc
Agendia
Louwesweg 6
1066 EC Amsterdam
The Netherlands
marc.warmoes@agendia.com

Weckx Stefan
VIB MicroArray Facility
KULeuven, onderwijs en navorsing, Herestraat 49, box 816
3000 Leuven
Belgium
stefan.weckx@vib.be

Wels Michiel
WCFS, p.a. CMBI - Radboud universiteit nijmegen
Postbus 9010
6500 GL Nijmegen
The Netherlands
mwels@cmbi.ru.nl

Wuyts Jan
BioInformatics & Evolutionary Genomics - VIB/UGENT
Technologiepark 927
9052 Ghent
Belgium
jan.wuyts@psb.ugent.be

Zamar Ruben
UBC
Agricultural Road 6356
V6T 1Z2 Vancouver
Canada
Stefan.VanAelst@UGent.be



Partners for Life.

Sun microsystems

At Sun Microsystems,™ we recognize that no company works alone in the life sciences community. That's why, as a leader in the field, we work with the foremost hardware and software vendors, systems integrators, standards bodies, academic organizations, and customers in developing solutions that deliver IT therapeutics for the global life science community.

www.sun.com/lifesciences

Keygene, based in Wageningen, the Netherlands, is a dynamic and leading international R&D company that provides genetic analyses and contract research in the support of commercial breeding and selection programs.

Keygene's business is based on a broad technology platform in the field on molecular genetics, functional genetics, and bioinformatics and is known world-wide as the inventor of the powerful AFLP® and cDNA-AFLP® technologies.

Keygene currently employs close to a hundred scientists.

Keygene N.V.
Postbus 216
6700 AE Wageningen

Agro Business Park 90
6708 PW Wageningen
The Netherlands

t +31 317 466 866
f +31 317 424 939

info@keygene.com
www.keygene.com

Keygene is a biotechnology company that carries out innovative research to advance commercial breeding by using a number of proprietary genetic technologies and know-how. To expedite creation of new varieties by its customers, Keygene's main focus is the analysis, creation and exploitation of genetic variation. Recently, we have initiated a number of new and ambitious programs. In the Applied Research unit we offer a

Postdoc position

Vacancy number 05002

Position

You will be working on scientific innovations in the unit Applied Research, with the aim to improve commercial (plant) breeding and selection processes by using molecular markers. The focus will be on the development of new strategies for using genetic markers in breeding populations, including software tools to analyze marker-trait associations.

Candidate

You have a Ph.D. degree with a background in the application of genetic markers for plant genome analysis, critical thinking ability in experimental design and interpretation and good communication skills at both written and oral levels. You are creative, ambitious, result-oriented and strive for efficiency improvement of breeding processes. You have international scientific experience and you have an interesting publication list.

Postdoctoral program

You will participate in our postdoctoral program. This program provides scientifically interesting topics and a thorough base for individual training and development to prepare for a biotechnological and/or academic career. In your day-to-day work you will be coached by a mentor. The postdoctoral program offers you possibilities to write publications and patents, visit seminars, give presentations and to gain experience in a commercial environment.

Keygene

Keygene offers you a position in a challenging and dynamic research company with an internationally recognized position in plant genetics research. You will be supported by highly qualified technicians and well-equipped research facilities. Keygene offers you the possibility to do high-level scientific research in a stimulating environment and an informal setting. Keygene is located in Wageningen, a renowned centre of agricultural research, nearby internationally attractive cities such as Amsterdam, Brussels and Cologne.

More information

Additional information can be obtained from Anker Sørensen, Innovation coordinator Applied Research, telephone +31 317 466866, email Anker.Sorensen@keygene.com.

If you are interested in this position you can send your letter of application and resume to: Keygene N.V. Attn. Ms A. Philipsen, HR manager, P.O. Box 216, 6700 AE Wageningen, The Netherlands or by email: Angelique.Philipsen@keygene.com Please mention the vacancy number 05002 in your letter of application.

NetApp simplifies enterprise data management.

Your customers. Our solutions. NetApp solutions simplify everyone's life, starting with yours. A single unified storage platform handles all your compliance and data protection needs. Now enterprise data is secure, accessible and recoverable - so you mitigate risk and maximize data availability. NetApp helps you keep the enterprise up and running with integrated hardware, software, and services delivering comprehensive data management solutions.

See how NetApp customers tell the story at www.netapp.com/go/enterprise

NetApp®
The evolution of storage.™

Bayer CropScience

To reinforce our research organization in Gent, we are currently looking for a

Senior Programmer-Analyst (m/f) and a **Programmer-Analyst** (m/f)

The candidates will join the BioData, Architecture & Services group that implements, maintains and supports a global framework of structured data management solutions to handle biological material and experimental data in support of BioScience Research, Development and Technology Management activities.

Senior Programmer-Analyst

Job Description

- You will further develop with our team the platform architecture by analyzing, designing, developing and implementing new bio-information components and solutions.
- You will collect, analyze and complement (key) user requirements, and elaborate the design and implementation plans.
- You will perform administration management on dedicated bio-information servers.

Profile

- You have a master in Informatics or equivalent through experience.
- You have 5-8 years of experience with relational databases (Sybase, Oracle) and with one or more programming tools (PowerBuilder, Java, Perl).
- Industrial and/or research experience in the development of large databases with complex data structures is an asset.
- You enjoy working in an international team environment and you have excellent communication and reporting skills.
- You are very service and customer oriented.

Programmer-Analyst

Job Description

- You will contribute to the development of the platform architecture by developing and implementing new bio-information components and solutions.
- You will develop and implement key user requirements.
- You will perform administration management of specific field trial database applications.

Profile

- You have a master in Informatics, Agronomy or equivalent through experience.
- You have 3-5 years of experience with relational databases (Sybase, Oracle) and programming tools (PowerBuilder, Java, Perl).
- Good knowledge in plant breeding, experience in statistics and data mining is an asset.
- You enjoy working in an international team environment and you have excellent communication and reporting skills.
- You are very service and customer oriented.

Further inquiries and applications, contact
Peggy Poelmans, HR Manager • +32 (0)9 243 04 72 • peggy.poelmans@bayercropscience.com
Bayer BioScience N.V., Technologiepark 38, 9052 Gent (Zwijnaarde)
Timing: Application will be received until April 29th, 2005

Bayer BioScience N.V. located in Gent - Zwijnaarde, Belgium,
is part of the business group BioScience of Bayer CropScience AG and is one of its main biotech innovation centers.
This centre of excellence integrates plant biotechnology research and business support functions
such as Legal and Intellectual Property, to optimize the innovation process in a responsible and sustainable approach.
The company currently employs little over 200 employees.