December 14-15 2009
University of Liège (Belgium)
Benelux Bioinformatics Conference

Louis Wehenkel
Pierre Geurts
Eric Bullinger
Kristel Van Steen
**BBC09**, the fifth Benelux Bioinformatics Conference takes place in Liège (Belgium) on December 14-15 2009 in the Academic Room of the University of Liège.

The Benelux Bioinformatics Conference is a two-day international conference on bioinformatics. Computational methods and mathematical models are playing an ever increasing role in large-scale "omics" molecular biology, leading to the emergence of bioinformatics as a scientific discipline within Systems Biology, with major applications in health and green biotechnology. The Benelux Bioinformatics Conference is the primary community event for bioinformatics researchers in the Netherlands, Belgium, and Luxemburg

**Conference Chairs**
Louis Wehenkel, Pierre Geurts, Eric Bullinger, Kristel Van Steen
Department of EE and CS and GIGA-R, University of Liège, Belgium

**Organizing Support**
Michèle Delville, Céline Dizier
Association des Ingénieurs de Montefiore (A.I.M.)

**Programme Committee**
Francisco Azuaje, Research Centre for Public Health (CRP-Santé), Luxembourg
Eric Bullinger, University of Liège, Belgium
Chris Evelo, Maastricht University, The Netherlands
Pierre Geurts, University of Liège, Belgium
Martijn Huynen, Nijmegen Centre for Molecular Sciences, The Netherlands
Jack Leunissen, Wageningen University, The Netherlands
Kathleen Marchal, K.U.Leuven, Belgium
Yves Moreau, K.U.Leuven, Belgium
Jan-Peter Nap, Hanze University of Groningen, The Netherlands
Jan Van den Bussche, Universiteit Hasselt & Transnationale Universiteit Limburg, Belgium
Yves Van de Peer, Gent University, Belgium
Kristel Van Steen, University of Liège, Belgium
Louis Wehenkel, University of Liège, Belgium
Lodewijk Wessels, NKI-AVL, The Netherlands
Sponsors

The organizers gratefully acknowledge the following sponsors who have provided financial support or/and precious help for the organization of the workshop.
## Contents

- Organization 3
- Sponsors 5
- Schedule 8
- Invited talks 10
- Oral Presentations 13
- Posters 33
- List of Authors 133
Schedule

Monday, December 14th

9h45-10h00 Welcome

10h00-11h00 Invited Talk

M. Madan Babu (MRC Laboratory of Molecular Biology, University of Cambridge), *Intrinsically Unstructured Proteins: Regulation and Disease*  \( p^{10} \)

11h00-12h20 Session 1

11h00-11h20 L. Palmeira, S. Penel, V. Navratil, J. Pellet, B. de Chassey, L. Meynier, V. Lotteau, C. Gautier, and C. Rabourdin-Combe. *Virus/host relations: analysing their interactions and co-evolution through data integration*  \( p^{24} \)

11h20-11h40 M. Schrynemackers, P. Geurts, L. Wehenkel, M. Madan Babu. *Prediction of genetic interactions in Yeast using machine learning*  \( p^{26} \)


12h00-12h20 M. Yousef, M. Ketany, L. Manevitz, L. C Showe, M. K. Showe. *Classification and biomarker identification using gene network modules and support vector machines*  \( p^{22} \)

12h20-14h30 Lunch break and poster session

14h30-15h50 Session 2

14h30-14h50 T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys. *Robust biomarker identification for cancer diagnosis using ensemble feature selection methods*  \( p^{25} \)

14h50-15h10 A. Daemen, M. Signoretto, O. Gevaert, J. AK Suykens, B. De Moor. *Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data.*  \( p^{15} \)

15h10-15h30 A. Essaghir, F. Toffalini, L. Knoops, A. Kallin, J. van Helden, J.B. Demoulin. *Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data.*  \( p^{16} \)


15h50-16h10 Coffee break

16h10-16h50 Session 3

16h10-16h30 E. Costa, C. Vens, H. Blockeel. *Top-Down Phylogenetic Tree Reconstruction*  \( p^{30} \)

16h30-16h50 V. Storms, M. Claeys, A. Sanchez, B. De Moor, A. Verstuyf, K. Marchal. *The Effect of Orthology and Coregulation on Detecting Regulatory Motifs*  \( p^{27} \)

16h50-17h50 Invited Talk
Pierre Rouzé (BioInformatics & Systems Biology, VIB Department of Plant Systems Biology, Ghent, Belgium). From protists to plants, fungi and animals: eukaryote genomes are not born equal

17h50 Welcome reception

Tuesday, December 15th

9h30-10h30 Invited Talk

Thomas Sauter (Life Sciences Research Unit, Université du Luxembourg). Modeling and Analysis of Metabolic and Signaling Networks using Bottom-Up Approaches

10h30-11h50 Session 4

10h30-10h50 D. Fey and E. Bullinger. Identification of biochemical reaction systems using semi-definite programming


11h10-11h30 I. Nepomuceno, F. Azuaje, P. V. Nazarov, A. Muller, Y. Devaux, L. Vallar, J. S. Aguilar-Ruiz, D. R. Wagner. Supervised prediction of heart failure through transcriptional association networks


11h50-14h00 Lunch break and poster session

14h00-15h20 Session 5

14h00-14h20 T. Cattaert, J.M. Mahachie John, K. Van Steen. Selecting multiple epistatic models using MB-MDR


14h40-15h00 C. Klijn, J. Bot, D.J. Adams, M. Reinders, L. Wessels, J. Jonkers. Identification of a network of co-occurring, tumor-related DNA copy number changes using a genome-wide scoring approach

15h00-15h20 K. Ye, K. Walter, M. Hurles, Z. Ning. Detecting Breakpoints of Large Deletions and Medium Sized Insertions on the Low Coverage Samples of 1000Genomes Project from Pair-end Short Reads

15h20-16h20 Invited Talk

Jeroen Raes (Vrije Universiteit Brussel). Untangling microbial ecosystems using metagenomics: from the oceans to human disease

16h20-16h30 Closing
Intrinsically Unstructured Proteins: Regulation and Disease

M. Madan Babu
MRC Laboratory of Molecular Biology, University of Cambridge, UK

Abstract
Altered abundance of several intrinsically unstructured proteins (IUPs) has been associated with perturbed cellular signaling that may lead to pathological conditions such as cancer. Therefore, it is important to understand how cells precisely regulate the availability of IUPs. We observed that regulation of transcript clearance, proteolytic degradation, and translational rate contribute to controlling the abundance of IUPs, some of which are present in low amounts and for short periods of time. Abundant phosphorylation and low stochasticity in transcription and translation indicate that the availability of IUPs can be finely tuned. Fidelity in signaling may require that most IUPs be available in appropriate amounts and not present longer than needed.

From protists to plants, fungi and animals: Eukaryote genomes are not born equal

Pierre Rouzé
BioInformatics & Systems Biology, VIB Department of Plant Systems Biology, Ghent, Belgium

Abstract
Since the incipience of genome-wide sequencing, more than a thousand genomes from eukaryotes have been sequenced and the low cost of the “new sequencing” technologies will suddenly bring many more on the shelves. There is a clear issue in making the best use of these data, finding and annotating the genes and other features from these new genomes. This issue has mainly been seen from a computer science perspective. I would like here to pinpoint another issue which has to do with biology. Most of the organisms which have been sequenced until lately were either fungi, animals or plants. Although all model organisms used and documented from a cell and molecular biology perspective are among these ones, this is nevertheless a minor fraction of the eukaryote phylogenetic spectrum. Annotation traditionally proceeds by analogy, either by searching for genes that are known or found elsewhere or ab initio by looking at recurrent features according to the knowledge we have of the molecular mechanisms of genome expression. Do we care about and know well enough these features and mechanisms, i.e. the way the information is structured and the way it is encoded in the lesser documented organisms which are going to be the bulk genome sequences soon? Having been involved in the annotation of such organisms, e.g. green algae, brown algae, diatoms and haptophyte we indeed came across unexpected findings which come as a warning of our capability to properly decipher the genome information content in such organisms.
Modeling and Analysis of Metabolic and Signaling Networks using Bottom-Up Approaches

Thomas Sauter
Life Sciences Research Unit, Université du Luxembourg

Abstract
The talk will introduce and discuss some key methods in mechanism based modeling of biological systems (“bottom-up systems biology”): Detailed ODE, constraint based and Boolean modeling. A variety of illustrative examples from metabolic and signaling networks will be given including carbohydrate metabolism in E. coli and apoptotic signaling in mammalian cells. To overcome the challenge of insufficient experimental data concerning model validation and parameter determination, possible strategies include identifiability analysis, reduced and coarse-grained modeling. In addition the newly founded “Luxemburgish Center for Systems Biology” will be introduced to the audience.

Untangling microbial ecosystems using metagenomics: from the oceans to human disease

Jeroen Raes
Vrije Universiteit Brussel, Belgium

Abstract
Metagenomics allows a large-scale, unbiased insight in the functional and phylogenetic composition of complex microbial communities. However, given the complexity of these data, computational analysis is a major bottleneck. Here, I will discuss our experiences in developing methods for the analysis of metagenomics data and their application in eg. the global ocean and human-associated communities (the human microbiome) such as the intestinal tract. These approaches should lead the way towards an ’eco-systems biology’ approach to the study of microbial communities and provide great opportunities for the study of the role of commensals in human disease.

References:
• Raes & Bork, Nat Rev Microbiol 6, 2008
• Gianoulis*, Raes* et al, PNAS 106, 2009
Oral Presentations

Tom Cattaert, Jestinah M. Mahachie John, Kristel Van Steen. Selecting multiple epistatic models using MB-MDR .......................................................... 14
Anneleen Daemen, Marco Signoretto, Olivier Gravez, Johan AK Suykens, Bart De Moor. Improved microarray-based decision support with graph encoded interactome data ............................................ 15
A. Essaghir, F. Toffalini, L. Knoops, A. Kallin, J. van Helden and J. B. Demoulin. Transmission factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data. .................................................. 16
Dirk Fey and Eric Bullinger. Identification of biochemical reaction systems using semi-definite programming ................................................................. 17
Christiaan Klijn, Jan Bot, David J. Adams, Marcel Reinders, Lodewyk Wessels, Jos Jonkers. Identification of a network of co-occurring, tumor-related DNA copy number changes using a genome-wide scoring approach .............................................................. 21
Malik Yousef, Mohamed Ketany, Larry Manevitz, Louise C Showe, Michael K. Showe. Classification and biomarker identification using gene network modules and support vector machines ........................................................................................................ 22
I. Nepomuceno, F. Azuaje, P. V. Nazarov, A. Muller, Y. Devaux, L. Vallar, J. S. Aguilar-Ruiz, D. R. Wagner. Supervised prediction of heart failure through transcriptional association networks ......................................................................................... 23
L. Palmeira, S. Penel, V. Navratil, J. Pellet, B. de Chassey, L. Meyniel, V. Lotteau, C. Gau-tier, and C. Rabourdin-Combe. Virus/host relations: analysing their interactions and co-evolution through data integration ........................................................................................................... 24
Thomas Abeel, Thibault Helleputte, Yves Van de Peer, Pierre Dupont, and Yvan Saeys. Robust biomarker identification for cancer diagnosis using ensemble feature selection methods ........................................................................................................ 25
Marie Schrynemackers, Pierre Geurts, Louis Wehenkel, M. Madan Babu. Prediction of genetic interactions in Yeast using machine learning ......................................................................................................................... 26
Valerie Storms, Marleen Claeyss, Aminael Sanchez, Bart De Moor, Annemieke Verstuyf, Kathleen Marchal. The Effect of Orthology and Coregulation on Detecting Regulatory Motifs ............................................................................................... 27
Eduardo Costa, Celine Vens, Hendrik Blockeel. Top-Down Phylogenetic Tree Reconstruction ........................................................................................................... 30
Kai Ye, Klaudia Walter, Matthew Hurles and Zemin Ning. Detecting Breakpoints of Large Deletions and Medium-Sized Insertions on the Low Coverage Samples of 1000Genomes Project from Pair-end Short Reads ........................................................................ 31
Selecting multiple epistatic models using MB-MDR
Tom Cattaert, Jestinah M. Mahachie John, Kristel Van Steen
Montefiore Institute-Bioinformatics, Université de Liège, Belgium

When searching for epistasis (gene-gene interaction), parametric regression approaches have severe limitations when there are many independent variables compared to the number of observed outcome events. Alternatively, one can use the non-parametric Multifactor Dimensionality Reduction method (MDR, Ritchie et al., 2001), which tackles the dimensionality problem by pooling multi-locus genotypes into two risk groups: a high and a low risk set. The MDR method and its numerous extensions are all extremely computer-intensive. Models are evaluated on the basis of cross-validation and prediction accuracy. Only the best model is singled out and its significance assessed through permutation testing.

We recently proposed FAM-MDR (Cattaert et al., 2009, in preparation), a novel unified MDR strategy for epistatic association analysis capable of handling both unrelateds and families of any structure, different outcome types (e.g. categorical, continuous or survival type). It allows for adjustment for lower order effects or confounding factors. In comparison with MDR, it is less computationally intensive since its core is based on Model-Based MDR (MB-MDR, Calle et al., 2007). The latter is a semi-parametric MDR method to detect epistasis in unrelated individuals. Models are evaluated via association tests on the final one-dimensional construct.

In this work, we address the issue of identifying several good models using MB-MDR methodology and borrowing ideas from FAM-MDR. To this end, alternatives for the identification of high and low multi-locus risk cells are discussed, because optimal definitions may substantially improve the power of our method and may dramatically reduce computation time. Furthermore, we validate and evaluate our epistasis detection method in a simulation study, by computing power and type I error under a variety of scenarios. Emphasis will be placed on multiple co-occurring epistatic models.

Improved microarray-based decision support with graph encoded interactome data

Anneleen Daemen, Marco Signoretto, Olivier Gevaert, Johan AK Suykens, Bart De Moor

Department of Electrical Engineering ESAT/SCD, Katholieke Universiteit Leuven, Leuven, Belgium

Motivation: Within the domain of cancer, microarray technology has earned a prominent place for its capacity to characterize the underlying tumor behavior in detail, leading to an improved diagnostic and prognostic capability. In the majority of studies, however, genes are treated as single entities without regard to their neighbors in the interactome network. Moreover, due to the noisy characteristics of microarray data and the limited overlap between gene signatures, prior biological knowledge should be incorporated in classification models.

Methods: We investigated interaction and pathway information that is available on different aspects of biological systems, obtained from 15 databases such as KEGG, REACTOME and OPHID. Ideally, one wants a classifier that does not make a distinction between a pathway that is switched on either by activation of gene A or inactivation of gene B. Relations between genes with similar functions but active in alternative pathways were therefore exploited and incorporated in a support vector machine based on spectral graph theory. Where traditionally with kernel methods the similarity between patients is calculated with a similarity measure based on the patients’ gene expression profiles, interactome data from any database can be used to improve the manner how patient similarities are calculated and thus also to improve classification performance.

Results: Using 10 microarray data sets, we first reduced the number of data sources that are relevant for multiple cancer types and outcomes. Three sources (one on metabolic pathway information, one on protein-protein interactions and one on miRNA-gene targeting) outperformed the other sources with regard to the considered class of models. Secondly, the prior knowledge database that will lead to the largest increase in classification performance is not known beforehand and depends on the specific microarray data set or type of predicted outcome. Moreover, inclusion of irrelevant prior knowledge may deteriorate the results. Both fixed and adaptive approaches were therefore considered to combine the three corresponding models. Averaging the predictions of these models performed best and significantly better than the model based on microarray data only. These results were confirmed on 6 validation microarray sets.

Conclusion: Irrespective of microarray technology or cancer type, integrating interactome data thus improves outcome prediction and classification decision making.
Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data.

A. Essaghir, F. Toffalini, L. Knoops, A. Kallin, J. van Helden and J. B. Demoulin

Université catholique de Louvain, Belgium

Deciphering transcription factor networks from microarray data remains an unresolved issue. Predictions based on in silico promoter analysis of mammalian genomes still produce high background. The activation of well-characterized target genes has been previously used to infer the regulation of p53, FOXO and SREBP1. By generalizing this approach, this study presents a simple method to mine microarray data for regulated transcription factors. We generated a catalogue containing 352 transcription factors associated with 2,721 well-characterized target genes, covering a total of 6,422 regulations. When it was available, a distinction between transcriptional activation and inhibition was included for each regulation. Next, we built a tool (www.TFactS.org) that compares new submitted gene lists with target genes in the catalogue, to detect regulated transcription factors. We validated TFactS with published lists of regulated genes in various models. Our results correctly predicted all the transcription factors reported in these studies. By contrast, looking for transcription factor enrichment based on in silico promoter analysis of the same genes was less informative (we tested: GSEA, TFM-Explorer, CORE, TF, CRSD and oPOSSUM). We also validated our method by a systematic analysis of NCI60 cancer microarray dataset which revealed the regulation of SOX10, MITF and JUN transcription factors in melanomas in relationship with BRAF mutations. We then performed microarray experiments comparing gene expression response of human fibroblasts stimulated by growth factors. TFactS predicted the activation of SREBP and the inhibition of FOXO, in line with our published results. In addition, it predicted the activation of STAT transcription factors by PDGF-BB, but not by other growth factors, which we confirmed experimentally. Our results show that the expression level of transcription factor target genes constitute a robust signature for transcription factor regulation, and can be efficiently used for microarray data mining.
Identification of biochemical reaction systems using semi-definite programming
Dirk Fey and Eric Bullinger
Systems & Modelling, Institute Montefiore, Université de Liège, Belgium

Estimation of kinetic parameters is a key step in modelling biochemical reaction networks as, often, their direct estimation is expensive, time-consuming or even infeasible. Here we present two complemental methods which explicitly take into account particular structure and dynamics of biological models as arising from mass action, Michaelis-Menten and Hill kinetics.

1. Limiting the parameter search space. The first approach proves inconsistency of entire parameter regions for a given model structure and data set, by formulating a semi-definite feasibility problem. The feasibility problem is only feasible for parameter regions that contain consistent parameter values. Thus, checking feasibility for different upper and lower bounds on the parameter values using semi-definite programming identifies inconsistent parameter regions. This drastically reduces the parameter search space, such that subsequent parameter estimation methods can disregard the inconsistent parameter regions. In contrast to similar approaches in the literature, the here presented approach does not require a steady state assumption, nor a discretisation of the system. Measurement uncertainties are dealt with using upper and lower bounds and regular sampling times are not required.

2. Estimating the parameter values. The second approach estimates the kinetic parameters using a nonlinear observer, i.e. a mathematical system that feeds back the error of prediction and measurement. The approach relies on a nonlinear state space extension transforming the system into a parameter-independent form. This state space representation facilitates the design of a nonlinear observer based on dissipativity arguments using linear matrix inequalities. An observer is a dynamical system performing the actual estimation by feeding back the error of prediction and measurement. To ensure the convergence of the estimate, certain observability conditions must hold. These are related to the identifiability of model and data, i.e. the existence of a unique solution for the parameter estimate.

Both methods are illustrated using simple, yet biological significant examples.
Unsupervised microarray analysis of glioma reveals prognostically relevant subgroups


Dept. Electrical Engineering (ESAT-SCD), KU Leuven, Belgium; Depts. Neurology, Bioinformatics, Biostatistics, Erasmus University Medical Center, Rotterdam, the Netherlands

Unsupervised analysis has already been applied successfully in many different cancer sites. However, in most cases hierarchical clustering methods are used which have a number of disadvantages. In a hierarchical clustering every sample has to belong to a cluster and there is no straightforward manner to estimate the uncertainty. In addition, the point where to cut the hierarchical tree is often chosen arbitrary. We used a different approach by applying a hybrid clustering algorithm called Hierarchical Ordered Partitioning And Collapsing Hybrid (HOPACH) which tackles these issues. HOPACH builds a hierarchical tree of clusters and combines the strengths of partitioning (e.g. K-means) and agglomerative clustering (e.g. hierarchical clustering) methods. This algorithm iteratively splits a cluster into smaller clusters with an enforced ordering followed by collapsing clusters if the average silhouette increases to correct for errors in the previous step. This is repeated until a complete hierarchical tree is constructed. We applied this algorithm on a large cohort of approximately 300 glioma samples of all histological subtypes and grades. Gliomas are the most common primary brain tumors with heterogeneous morphology and variable prognosis. Treatment decisions in patients rely mainly on histological classification and clinical parameters. However, differences between histological subclasses and grades are subtle, and classifying gliomas is subject to a large inter-observer variability. Unsupervised microarray analysis can provide a more objective classification of glioma. Our results showed that seven distinct molecular subgroups exist that correlate with survival. Bootstrapping analysis showed that the clustering was robust and included two favorable prognostic subgroups, two with intermediate prognosis, two with poor prognosis, and one control group. These intrinsic molecular subtypes of glioma are different from histological subgroups and correlate better to patient survival. The prognostic value of molecular subgroups was further validated on three large independent sample cohorts from The Cancer Genome Atlas project (TCGA), the Rembrandt project and the GSE12907 data set. In addition, other important genetic markers for glioma such as EGFR amplification, IDH1 mutation and 1p/19q loss of heterozygosity segregate in distinct molecular subgroups. Our results show that advanced clustering methods are able to identify novel clinically relevant subgroups of glioma that correlate with prognosis. In addition, expression profiling appears to be a more accurate and objective method to classify glioma samples compared to histological classification. Molecular classification therefore can have an important influence on diagnosis and can guide clinical decision making.
Mutation prioritization for next-generation sequencing
Department of Human Genetics, Nijmegen Centre for Molecular Life Sciences, Radboud University Nijmegen Medical Centre, Nijmegen, the Netherlands

With the development of efficient enrichment strategies for next-generation sequencing it is now possible to screen large numbers of patients for causative mutations in candidate (gene) regions. Pinpointing the exact causative mutation is however hampered by the sheer amount of normal variation between individuals at the base pair level and by our limited knowledge on genotype-phenotype correlation. Although some biological hallmarks for pathogenic mutations exist they are not easily available for prioritizing large lists of candidate mutations as revealed by next generation sequencing. Furthermore, any one hallmark has only limited power in discriminating a benign variation from a pathogenic mutation. For this reason we developed a computational pipeline that prioritizes candidate mutations based upon biological relevant information such as evolutionary base pair conservation, local genomic architecture, and effects at the amino acid level. We apply this prioritization approach to next generation sequencing data from 7 patients with autosomal recessive ataxia. For these patients we enriched genomic material harbouring 10 known ataxia genes, which we sequenced on the Roche 454 Titanium system. For all of these patients we were able to successfully prioritize the causative mutation among the thousands of benign base pair variations.
Mechanisms of redundancy between signaling pathways
Holstege Lab, Department of Physiological Chemistry, UMC Utrecht, The Netherlands

Phosphorylation-based signaling is crucial for almost all cellular processes. For instance, signaling pathways are essential for translating environmental changes into appropriate cellular responses and thus allow cells to deal with a multitude of environmental conditions and external stimuli. Perturbations in signaling pathways often lead to aberrant behaviour or diseases such as cancer. A better understanding of these pathways is therefore crucial towards understanding important cellular processes and diseases. Here, we systematically investigate all known signaling pathways in Saccharomyces cerevisiae by expression profiling deletion mutants of all non-essential kinases and phosphatases that make up the various components of these pathways. Analysis of these gene expression profiles groups kinases and phosphatases functioning within the same pathway together, reveals antagonistic behaviour between different pathways and identifies the various cellular processes that these pathways target. About a third of the deletion mutants do not show any phenotypical effect. One possible explanation for the lack of observed phenotypes is a high degree of redundancy or cross-talk between the pathways. To investigate this, we first analyzed a set of synthetic genetic interactions between all kinases and phosphatases for extracting redundant kinase/phosphatase pairs. Putative redundant pairs where then further investigated by expression profiling the double deletions. This reveals three distinct types of redundancy: complete redundancy; partial redundancy and quantitative redundancy, each with its own biological implications. A number of putative functional roles derived from the redundant pairs are verified by additional follow-up experimentation. In addition, by combining synthetic genetic data and gene expression profiling, we can more precisely pinpoint the proteins involved in the cross-talk between different pathways. Follow-ups directed at establishing the exact mechanism of the cross-talk are provided for the cell wall integrity and osmosensing pathway.
ORAL PRESENTATIONS

Identification of a network of co-occurring, tumor-related DNA copy number changes using a genome-wide scoring approach
Christiaan Klijn, Jan Bot, David J. Adams, Marcel Reinders, Lodewyk Wessels, Jos Jonkers
Netherlands Cancer Institute, Delft University of Technology, Wellcome Trust Sanger Institute

Tumorigenesis is a multi-step process in which normal cells transform into tumors following the accumulation of mutations that enable them to evade the checkpoints that would normally suppress their growth or result in apoptosis. Copy Number Alterations (CNAs) are one of the mechanisms which tumor cells employ to activate oncogenes and inactivate tumor-suppressor genes. Often, multiple genes need to be simultaneously activated or suppressed to initiate tumorigenesis. We hypothesize that such synergism between cancer genes might be identified by looking for regions of co-occurring gain and/or loss.

To find CNAs that are co-occurring we developed a genome-wide scoring framework to separate truly co-occurring aberrations from the noisy passenger and dominant single signals present in the data. Our score weighs both the intensity and the co-variance of the CNAs of a pair of loci in a set of samples. Using 2d Gaussian kernel convolution we were able to define peaks of co-occurrence in a 2d pairwise genomic space. The resulting regions of high co-occurrence were investigated for inter-region functional interactions. Given the computational complexity of doing a whole-genome pairwise comparison at a high resolution we have used a distributed computing approach to make the calculations feasible.

Analysis of high-resolution DNA copy number data from a panel of 95 hematological tumor cell lines correctly identified co-occurring recombinations at the T-cell receptor and immunoglobulin loci in T- and B-cell malignancies, respectively. This demonstrates that we can recover truly co-occurring genomic alterations. The other loci showing a strong co-occurring loss are significantly enriched for tumor suppressor genes and the co-occurring gained loci show an enrichment for oncogenes.

Using hierarchical clustering on the single locus pairs that are detected on the lowest genomic scale of our analysis, we were able to construct a network of co-occurring genomic locations. The nodes in this network represent genomic loci clustered together and the edges represent co-occurrences of copy number changes in the data. The genes in a node show a very strong, significant increase in functional interactions with genes in connected nodes. Up to 11% of all genes are reported to have a functional interaction with genes in a connected node as opposed to 1% in a random situation.

We prune our complete networks to focus on the most important interactions. In this way we define a core network of co-occurring losses centred around chromosome 17p. This network of losses is independent of the canonical cancer genes present in the loci described by the nodes.

Our results shows that in the samples we analyzed, large scale, low intensity copy number changes affecting many functionally related genes are an important feature of cancer development or maintenance. They might affect the gene dosages of a large interconnected network functionally related genes, whose loss or gain is not necessarily driven by a few canonical cancer genes.
Classification and biomarker identification using gene network modules and support vector machines
Malik Yousef, Mohamed Ketany, Larry Manevitz, Louise C Showe, Michael K. Showe
The Institute of Applied Research- The Galilee Society, Shefa-Amr, Israel
Al-Qasemi Academic College, Baqa Algharbiya, Israel
Computer Science Department, Haifa University, Haifa, Israel
Molecular Oncogenesis/Systems Biology Program, The Wistar Institute, Philadelphia, PA 19104, USA
Corresponding Author

Abstract Background: Classification using microarray datasets is usually based on a small number of samples for which tens of thousands of gene expression measurements have been obtained. The selection of the genes most significant to the classification problem is a challenging issue in high dimension data analysis and interpretation. A previous study with SVM-RCE (Recursive Cluster Elimination), suggested that classification based on groups of correlated genes sometimes exhibits better performance than classification using single genes. Large databases of gene interaction networks provide an important resource for the analysis of genetic phenomena and for classification studies using interacting genes. We now demonstrate that an algorithm which integrates network information with recursive feature elimination based on SVM exhibits good performance and improves the biological interpretability of the results. We refer to the method as SVM with Recursive Network Elimination (SVM-RNE)

Results: Initially, one thousand genes selected by t-test from a training set are filtered so that only genes that map to a gene network database remain. The Gene Expression Network Analysis Tool (GXNA) is applied to the remaining genes to form n clusters of genes that are highly connected in the network. Linear SVM is used to classify the samples using these clusters, and a weight is assigned to each cluster based on its importance to the classification. The least informative clusters are removed while retaining the remainder for the next classification step. This process is repeated until an optimal classification is obtained. Conclusion: More than 90% accuracy can be obtained in classification of selected microarray datasets by integrating the interaction network information with the gene expression information from the microarrays.
Supervised prediction of heart failure through transcriptional association networks
I. Nepomuceno, F. Azuaje, P. V. Nazarov, A. Muller, Y. Devaux, L. Vallar, J. S. Aguilar-Ruiz, D. R. Wagner
Dpto. Lenguajes y Sistemas Informáticos, University of Seville, Spain
Laboratory of Cardiovascular Research, CRP-Santé, Luxembourg
Microarray Center, CRP-Santé, Luxembourg
School of Engineering, Pablo de Olavide University, Spain
Division of Cardiology, Centre Hospitalier, Luxembourg

Motivation: more than 60% of myocardial infarction (MI) patients develop heart failure (HF) during the six years post-MI. For this reason, a major contribution in cardiovascular research would be the identification of key molecular mechanisms that drive HF in post-MI patients together with the early identification of patients at risk of developing HF. In this study, we analyzed gene expression biosignatures of blood cells to identify potential biomarkers in a group of patients with high and low ejection fraction (EF). In order to identify mechanisms potentially driving HF, i.e. low EF, we developed a novel method to infer gene co-expression networks from patients with high and low EF.

Method: as a first step, we present a method to find putative, direct regulatory relationships between genes, which allowed us to build gene association networks specific to each clinical category. Unlike correlation-based methods, which analyze each pair of genes, our method analyzes each gene by taking into account the remaining genes as inputs to estimate the value of the gene under analysis. The latter is done with linear regression functions using model trees. Model trees are usually more precise than linear regression models because they can adjust different linear regressions to separate areas of the search space and, hence, favouring localized similarities over global similarities. Moreover, the networks obtained can be seen as a classification model taking into account the estimated and true values of the genes in the network. This model was used to classify low and high EF patients.

Results: we analyzed gene expression profiles from blood cells taken at the time of MI in two groups of 16 patients each with high EF and low EF. The proposed method was applied to both groups independently. The method, being applied to 15307 input genes identified a network with 17 edges (gene-gene associations) and 19 nodes (genes) in high EF patients, and a network with 59 edges and 48 genes in low EF patients. These networks have 8 genes in common, and were compared with those obtained by correlation-based methods. Topologically, the networks obtained by our approach are significantly more compact, which facilitates expert interpretation and future validations. Furthermore, the networks generated by our method can be applied to automatically predict clinical outcome. An average classification accuracy of 72% was obtained after a cross-validation.

Conclusions: gene association networks provide new insights into underlying molecular mechanisms leading to HF after MI and set the basis for potential novel prognostic models in HF. Moreover, the developed approach is not limited to HF and can be used for the gene network reconstruction based on microarray data in general.
Virus/host relations: analysing their interactions and co-evolution through data integration
L. Palmeira, S. Penel, V. Navratil, J. Pellet, B. de Chassey, L. Meyniel, V. Lotteau, C. Gautier, and C. Rabourdin-Combe
Laboratoire de Biométrie et Biologie Evolutive, Lyon, France

Viruses are responsible for a large number of human infectious diseases and cancers. Understanding how they assemble and how they interact with their hosts remains a major challenge. We are mainly interested in studying how these complex interactions between several organisms are acquired and maintained throughout evolution. We therefore developed a global approach aimed at providing accurate evolutionary and phylogenetic information to tackle these questions as well as experimental data from both the transcriptome and interactome.

We are predominantly interested in understanding viral evolution and viral adaptation. Indeed, the mechanisms behind fast viral adaptation are far from being elucidated. Determining, for instance, whether viruses evolve as a pool of interchangeable genetic modules allowing for a rapid evolution is still an open question. Identifying modularity in viral protein evolution as well as identifying modules or genes that could have been horizontally transmitted between viruses and hosts is also a key question. In order to answer these questions, we are working on (1) determining homologies between viruses and between viruses and their hosts and (2) identifying small modules in all available viral proteins. We have thus constructed a database of families of homologous proteins between fully sequenced viruses, human and insects vectors of pathogens. Access to these families is available through the PhEVER database – http://pbil.univ-lyon1.fr/databases/phever/.

Further work on the identification of modules inside proteins is still under way.

Subsequently, this information is crossed with data from the interactome and from the transcriptome of different virus-infected cells to determine how similar pathways can be the target of viruses containing homologous sequences. In a previous analysis of hepatitis C virus interaction network, for instance, we were able to show that human cellular proteins interacting with viral proteins are enriched in highly central and interconnected proteins of the network. Therefore, identifying to what extent different viruses can target the same central proteins in the network is one of our major goals. We have hence developed a strategy of mapping of interactions not only between viral proteins but also between viral and human proteins. Data generated by yeast two-hybrid screening between all viral proteins and all human proteins is already available for at least 20 viruses including Flaviviridae and has been compiled with protein-protein interaction curated from the literature.

We present here an integrated approach of genomic and post-genomic data to analyse virus/host co-evolution. This global approach based on data from the interactome, transcriptome and genome of viruses, human and insects vectors of viral pathogens will undoubtebly provide valuable tools for a better understanding of the mechanisms of co-evolution.
Robust biomarker identification for cancer diagnosis using ensemble feature selection methods
Thomas Abeel, Thibault Helleputte, Yves Van de Peer, Pierre Dupont, and Yvan Saeys

Department of Plant Systems Biology, VIB, Technologiepark 927, 9052 Gent, Belgium
Department of Molecular Genetics, Ghent University, Gent, Belgium
Department of Computing Science and Engineering INGI, Universiteit catholique de Louvain, Belgium
Machine Learning Group, Universiteit catholique de Louvain, Belgium

Biomarker discovery is an important topic in biomedical applications of computational biology, including applications such as gene and SNP selection from high dimensional data. Surprisingly, the stability with respect to sampling variation or robustness of such selection processes has received attention only recently. However, robustness of biomarkers is an important issue, as it may greatly influence subsequent biological validations. In addition, a more robust set of markers may strengthen the confidence of an expert in the results of a selection method.

Our first contribution is a general framework for the analysis of the robustness of a biomarker selection algorithm. Secondly, we conducted a large scale analysis of the recently introduced concept of ensemble feature selection, where multiple feature selections are combined in order to increase the robustness of the final set of selected features. We focus on selection methods that are embedded in the estimation of support vector machines (SVMs). SVMs are powerful classification models that have shown state-of-the-art performance on several diagnosis and prognosis tasks on biological data. Their feature selection extensions also offered good results for gene selection tasks. We show that the robustness of SVMs for biomarker discovery can be substantially increased by using ensemble feature selection techniques, while keeping the same classification performances.

The proposed methodology is evaluated on four microarray data sets showing increases of up to 27% in robustness of the selected biomarkers. The stability gain obtained with ensemble methods is particularly noticeable for small signature sizes (a few tens of genes), which is most relevant for the design of a diagnosis or prognosis model from a gene signature.
Prediction of genetic interactions in Yeast using machine learning
Marie Schrynemackers, Pierre Geurts, Louis Wehenkel, M. Madan Babu
Department of EE and CS & GIGA-R, University of Liège, Belgium
MRC Laboratory of Molecular Biology, Cambridge, UK

The inference of the genetic interaction network of an organism is an important challenge in systems biology. An interaction (epistasis) exists between two genes when the presence or the absence of one gene modifies the effect of the other genes. The knowledge of these interactions is very important to understand the functions of the genes and their products. In Yeast S. cerevisiae, the "Epistatic miniarray profile" (E-MAP) technique allowed to measure some interaction subnetworks on two subsets of respectively 424 and 743 genes (Schuldinger et al. 2004, Collins et al., 2007). For the time being, it remains however impossible to test experimentally the 18 millions potential interactions between the 6000 yeast genes. In this work, we propose to use computational techniques based on machine learning to complete the experimentally confirmed interactions.

From a machine learning point of view, the goal is, from the known subnetwork, to infer a model able to predict the presence or absence of an interaction between two genes from some features of those genes. We proposed several strategies to transform this problem into one or several standard supervised classification or regression problems and exploited tree-based ensemble methods to learn the models. To validate the approach, we used the two available E-MAPs as our training subnetwork and considered as input gene features several expression data and chemo-genomic profiles (Hillenmeyer et al., 2008).

The evaluation of the models by cross-validation shows that we are able to predict new interactions with a reasonable accuracy. Through feature relevance ranking, the approach should also shed new light on the biology behind the interactions. Future works will consider the addition of new input features and a thorough analysis of genome-wide predictions.
The Effect of Orthology and Coregulation on Detecting Regulatory Motifs

Valerie Storms, Marleen Claeys, Aminael Sanchez, Bart De Moor, Annemieke Verstuyf, Kathleen Marchal

CMPG, Department of Microbial and Molecular Systems, K.U.Leuven, Leuven 3001, Belgium
Laboratory of Molecular Biology, Institute of Plant Biotechnology, Central University ‘Marta Abreu’ of Las Villas (UCLV), Santa Clara 54830, Cuba
Department of Electrical Engineering ESAT-SCD, K.U. Leuven, Leuven 3001, Belgium
Departement Experimentele Geneeskunde, LEGENDO, K.U. Leuven, Leuven 3000, Belgium

Despite the enormous collection of tools available for the discovery of transcription factor binding sites (motifs), the detection of motifs remains a challenging problem. Initially motif discovery was performed on a set of genes, known to be coregulated and thus containing binding sites for the same transcription factor. But with the growing number of sequenced genomes, detecting motifs through ‘phylogenetic footprinting’ became feasible and the next generation of motif discovery tools has therefore integrated the use of orthology in addition to the coregulation information. While these tools treat the orthologous sequences as independently, and thus ignore the underlying phylogeny that describes their relatedness, the most advanced motif discovery tools explicitly incorporate these phylogenetic relations by means of an evolutionary model. Despite the potential of comparative data in addition to coregulated data, so far no independent study has evaluated the extent of information contained within either the coregulation or the orthologous space, and the conditions under which complementing both spaces becomes useful. In this study we performed such analysis by 1) designing appropriate datasets (real and synthetic) and 2) applying two of the more advanced probabilistic methods for motif discovery: Phylogibbs[1] and Phylogenetic sampler[2], which both model the evolutionary dependency in the orthologous space. To set the base line performance we included MEME[3], as a representative of algorithms that cannot explicitly incorporate phylogenetic relations. Our results show that the success rate of combining coregulation and orthology information depends on the complex relation between the algorithm and the dataset. The performed tests illustrate that the nature of the used algorithm is crucial in determining how to exploit multiple species data in the best way to improve motif discovery performance. The results of this unbiased comparison also point out the strengths and weaknesses of current implementations, information which is useful for both developers and users.

A generic gene regulatory network reconstruction method: application to Lactococcus lactis MG1363


Center for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Center, the Netherlands
Kluyver Centre for Genomics of Industrial Fermentation, Delft, the Netherlands
NIZO food research BV, Ede, the Netherlands
Molecular Genetics group, GBB, RUG, Kerklaan 30, Haren, the Netherlands
NBIC, Toernooiveld 1, 6525 ED, Nijmegen, the Netherlands
TI Food and Nutrition, Wageningen, the Netherlands

Gene regulatory networks (GRN) are used by prokaryotes to adapt to changing environment by regulating gene expression levels. The basic building blocks of GRN are regulons, which consist of a transcriptional regulator and a number of target genes. In prokaryotes these target genes are organized into operons. We develop a generic method to predict gene regulatory networks for any prokaryote. As a model system for predicting the GRN the dairy bacterium Lactococcus lactis MG1363 was chosen. Our method consists of four major steps: (i) training a Random Forest based classification model on the known regulatory interactions of both Escherichia coli and Bacillus subtilis. For the classification, not only regulatory interaction between transcriptional regulator and its target are used, but any interaction occurring between genes part of the same regulon. Features used to predict interactions between two genes include operons, coexpression in DNA microarray data, gene ontologies, metabolical pathways; (ii) classifying interactions between any two genes in L. lactis MG1363; (iii) identifying quasi-cliques, so called dense sub-graphs, these are parts of the network that show a higher connectivity than its surrounding. These quasi-cliques are assumed to be regulatory modules, regulons or combinations of regulons; and finally (iv) identifying putative transcriptional regulators that govern the expression of the genes part of a quasi-clique. A Random Forest classification model is trained based on the known regulator-target interactions to predict what gene is actually the regulator within a quasi-clique. The result of applying these four steps onto gene-pairs of L. lactis MG1363 is a putative gene-regulatory network. Preliminary results indicate favorable performance of our method: in general, 60% of the known L. lactis regulatory interactions are correctly identified. When considering the smaller regulons, with member sizes of 6 to 14 members accuracy increases to over 80%. In future, we will further develop the method and implement it into user-friendly software to accurately determine the GRN for any prokaryote.
Anatomically Guided Differential Analysis of Neurodegenerative Alterations in Tissue via Mass Spectral Imaging and Spatial Querying

Raf Van de Plas, Kristiaan Pelckmans, Thomas Philips, Bart De Moor, and Etienne Waelkens

Katholieke Universiteit Leuven, Dept. of Electrical Engineering (ESAT), SCD-SISTA (BIOI), Kasteelpark Arenberg 10, B-3001 Leuven (Heverlee), Belgium
Katholieke Universiteit Leuven, Dept. of Molecular Cell Biology, Sec. Biochemistry, O&N, Herestraat 49 - bus 901, B-3000 Leuven, Belgium
Katholieke Universiteit Leuven, ProMeta, Interfaculty Centre for Proteomics and Metabolomics, O&N2, Herestraat 49, B-3000 Leuven, Belgium
Katholieke Universiteit Leuven, Dept. of Experimental Neurology, O&N 2, Herestraat 49 - bus 1022, B-3000 Leuven, Belgium

Imaging mass spectrometry or mass spectral imaging (MSI) is a developing technology that adds a spatial dimension to mass spectrometry-based biochemical analysis. It delivers insight into the local distribution of biomolecules (e.g. proteins, peptides, and metabolites) throughout an organic tissue section. MSI has particular merit in disease-driven searches for tissue biomarkers and the study of localized chemical interactions. It can serve both as an exploratory and confirmatory instrument.

Current approaches for the examination of the vast amounts of data generated in a MSI experiment have either focused on providing a spatial answer to a mass-driven hypothesis (e.g. ion images of a single mass species) or have focused on disassembling the measurements into a set of tissue region trends in a data-driven manner (e.g. PCA, ICA,...). The opportunity to examine the tissue and look for ions of interest using a morphological region-driven hypothesis rather than a molecule-driven hypothesis has been unavailable due to lack of a suitable method.

In this work we introduce one of the first efficient computational methods that enables the researcher to interrogate MSI data from a spatial standpoint. The method considers the MSI data set as a whole and is able to answer questions such as which ions show upregulation (or down-regulation) specific to a particular area of the tissue (e.g. the hippocampus region in brain). Additionally it can also examine complexer hypotheses involving relative abundance levels across different anatomical regions (e.g. abundance gradients). Spatial-driven questions arise, for example, in pathomechanisms that show location-specific behavior (e.g. Parkinson’s and Huntington’s disease) and in the search for molecules specific to a certain anatomical region.

The utility of the proposed spatial querying method is demonstrated on a set of neural tissue case studies. First, we demonstrate the exploratory use of the technique by extracting a list of ions that only appear in a specified region of mouse brain. Then we extend the paradigm by differentially examining anatomically identical tissue regions in control versus disease. This generates a differential ion profile which is demonstrated on mouse spinal cord.
Top-Down Phylogenetic Tree Reconstruction
Eduardo Costa, Celine Vens, Hendrik Blockeel
Department of Computer Science, Katholieke Universiteit Leuven, Belgium
Institut National de Recherche Agronomique, UMR 1301, Sophia-Antipolis, France
Leiden Institute of Advanced Computer Science, Universiteit Leiden, The Netherlands

We propose a novel distance based method for reconstruction of phylogenetic trees. Our method is based on a conceptual clustering method that extends the well-known decision tree learning approach. Our method starts from a single cluster and repeatedly splits it into subclusters until all sequences form a different cluster. We assume that a split can be described by referring to particular polymorphic locations, which makes such a divisive method computationally feasible. To define the best split, our method uses a criterion that is close to Neighbor Joining’s (NJ) optimization criterion, namely, constructing a phylogenetic tree with minimal total branch length. Our approach has a number of important advantages. First, by listing the polymorphic locations at the internal nodes, it provides an explanation for the resulting tree topology. Second, the resulting trees can be used as classification trees to add new sequences to the phylogeny. Third, the top-down tree growing process can be stopped before a complete tree is generated, yielding an efficient gene or protein subfamily identification approach. Our algorithm, called Clus-ϕ, can be found at http://www.cs.kuleuven.be/dtai/clus.

We tested Clus-ϕ on a number of synthetic datasets by simulating an evolutionary process, using a coding sequence simulation program. By default, the simulator produces symmetric trees, i.e., binary trees having all leaves at the same depth. However, also completely random tree topologies can be produced. Our trees were compared to those obtained by NJ in terms of similarity with the true tree topology and computational cost.

For datasets based on symmetric tree topologies, Clus-ϕ outperforms NJ. Sometimes, the trees produced by Clus-ϕ are identical to the true topologies, even for datasets with many sequences. However, for datasets based on random trees, NJ outperforms Clus-ϕ. In order to verify how the results change for topologies in between those two extremes, we generated datasets with other topologies. We started from the symmetric case and performed modifications in the topology to draw it towards the random case. The results revealed that the performance of Clus-ϕ decreases as the number of modifications increases. However, this decrease occurs gradually, which means that Clus-ϕ also presents good results for almost symmetric trees.

Regarding computational cost, Clus-ϕ is slower than NJ when analyzing a relatively small number of sequences, but clearly outperforms NJ for datasets with a large number of sequences. However, because it investigates polymorphic locations at each internal node, Clus-ϕ scales worse in the length of the sequences.

We propose Clus-ϕ not as a substitute for NJ or other standard methods for phylogeny, but as a method to be used complementarily to these methods. The fact that Clus-ϕ behaves differently than NJ (better in some cases; worse in other cases), in terms of performance and computational cost, shows that their results can be used to complement one another.

Acknowledgements: E.C. is supported by the Research Foundation - Flanders (FWO) and the GOA Probabilistic Logic Learning. C.V. is a Postdoctoral Fellow of the Research Foundation - Flanders (FWO).
Detecting Breakpoints of Large Deletions and Medium Sized Insertions on the Low Coverage Samples of 1000Genomes Project from Pair-end Short Reads

Kai Ye, Klaudia Walter, Matthew Hurles and Zemin Ning
Molecular Epidemiology Section, Leiden University Medical Center, Einthovenweg 20, Leiden, The Netherlands
The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom

There is a strong demand in the genomic community to develop effective algorithms to identify genomic variants from multiple samples, in order to investigate disease-related variants and genetic survey of large populations. We developed a so-called Pindel method to identify breakpoints of large deletions (1bp-10kb) and medium sized insertions (1-20bp) at base level precision from 36 bp paired-end short reads. In the preprocessing step, all reads are mapped to the reference genome. Then the mapping results are examined to select those paired reads that only one end can be mapped. For each of those remaining pairs, the mapped reads must be uniquely located in the genome while their mates cannot be mapped to anywhere in the genome under a given threshold alignment score. Our Pindel program uses the mapped read to determine the anchor point on the reference genome and the direction of unmapped read. Knowing the anchor point, the direction to search for the unmapped read and the user defined Maximum Deletion Size, a sub-region in the reference genome can be located, where Pindel will break the unmapped reads into 2 (deletion) or 3 (short insertion) fragments and map the two terminal fragments separately.

In this study we developed a novel procedure based on Pindel algorithm to process multiple samples. We add tags to the reads to indicate their sources. Then we run Pindel using the entire pool of reads as the input. We modified our Pindel program to report the sample sources of the supporting reads for each identified indel event. With such information we are able to discern which samples have what indels. We also adapted the algorithm to include detection of deletions with small insertions of non-templated sequences at the breakpoint, and to report a confidence score that is monotonically related to false discovery rate. We demonstrated our method with the low coverage data of human chr1 from 170 individuals in the 1000genomes project. We extracted 1,141,756,691 chr1 related one-end mapped reads from the BAM files of 170 individuals from 4 populations (CEU, CHB, JPT, YRI). It took Pindel less than 2 hours to predict 60,663 1bp-10kb deletions and 49,903 1-60bp insertions on a single CPU.

Applying the new method on the sequence data of human chr1 from 170 low coverage individuals, we obtained several intriguing features. First, there are more than 35k 1bp deletions in chr1 among these 170 individuals. Secondly from 1 bp to 100 bp, the frequency decreases as the size of deletion increases. Thirdly, there is a peak around 300bp, which may contain dozens of putative SINEs. Last but not least, there are periodicities of 2 and 4 when deletion size is smaller than 100bp. There are a clear peak at 4bp and dominating periodicity of 2bp for deletions larger than 8bp.
Posters

P1: Thomas Abeel, Thomas Van Parys, James Galagan, Yves Van de Peer. GenomeView: visualizing the next generation of data ................................................................. 37
P2: Sanne Abeln, Daan Frenkel. Information about solvent contacts improves foldability in knowledge-based protein model ......................................................... 38
P3: Michiel E. Adriaens, Andrea Romano, Nigel J. Belshaw, John C. Mathers and Chris T.A. Evelo. A bioinformatician’s toolbox for ChIP & DNA methylation microarray analysis ................................................................. 39
P4: Johan Desmet and Philippe Alard. Improving antibody properties by in-silico engineering: a realistic perspective ................................................................. 40
P5: Anagha Joshi, Yves Van de Peer and Tom Michoel. Modules, Network motifs and Hierarchical organization of post-transcriptional regulatory network ......................... 41
P6: Andra Waagmeester, Eduard Hoenkamp and Chris Evelo. Review: 50 years of knowledge extraction and its application in biology ........................................ 42
P7: Lieven Baeyens and Yvan Saeys. A comparative evaluation of discretisation methods for Bayesian classifiers on high-dimensional cancer data. ......................... 43
P8: Eric Bareke, Bertrand De Meulder, Michael Pierre, Eric Depiereux. HUBase: A Human-based Expression Array Data Builder .............................................. 44
P9: Samuel Hiard, Carole Charlier, Wouter Coppieeters, Michel Georges and Denis Baurain. Patrocles: a database of polymorphic miRNA-mediated gene regulation in vertebrates Lactococcus lactis strains ......................................................... 45
P11: Julien Becker, Louis Wehenkel, Pierre Geurts. Enzymatic protein prediction from structure using a segment and combine approach ........................................ 47
P12: Jan Bot, Marc Hulsman, Marcel Reinders. The Ibidas data integration and accession platform ................................................................. 48
P13: Vincent Botta, Pierre Geurts, Louis Wehenkel. Using trees as test nodes to model haplotype blocks in GWAS ................................................................. 49
P14: Magdalena Calusinska, Annick Wilmotte. The diversity of clostridial hydrogenases revealed by genome sequencing projects ........................................ 50
P16: Susan LM Coort, Thomas Kelder, Chris TA Evelo. Regulatory gene interaction network visualization using Cytoscape .............................................. 52
P18: Bertrand De Meulder, Benoît De Hertogh, Fabrice Berger, Michael Pierre, Eric Bareke and Eric Depiereux. Methodology for benchmarking microarray data analysis methods ................................................................. 54
P19: Riet De Smet, Hong Sun, Karen Lemmens, Inge Thijs, Kathleen Marchal. A query-based framework to validate ChIP-chip targets ........................................ 55
P20: Jingyuan Fu, Patrick Deelen, Rudolf Fehrman, Gerard te Meerman, Marcel Wolfs, Cisca Wijmenga, Lude Franke. Complexity of human gene expression traits in different primary tissues ................................................................. 56
P23: Thomas Kelder, Alexander Pico, Martijn van Iersel, Kristina Hanspers, Bruce Conklin and Chris Evelo. WikiPathways, community-based curation for biological pathways ................................................................. 59
P24: F. Qiang, K. Engelen, P. Meysman, K. Lemmens, R. Desmet, C. Fierro, I. Thijs, K. Marchal. Colombos: the access port to the cross-platform microarray compendium ................................................................. 60
P30: Raoul Frijters, Marianne van Vugt, Ruben Smeets, René van Schaik, Jacob de Vlieg, Wynand Alkema. *CoPub Discovery: a literature mining tool for hidden biological knowledge* ......................................................... 66
P31: Omar Gaci and Stefan Balev. *How to Fold a Protein using Amino Acid Interaction Networks* ......................................................... 67
P34: Ying He, Grigoris Amoutzias and Yves Van de Peer. *Evolution of protein-protein interaction networks for plant kZIP transcription factors* ......................................................... 70
P35: Fabien Heuze and Benoît Charleux. *Structural features and interaction defects of mutated proteins associated with human Mendelian disorders* ......................................................... 71
P36: Hong Sun, Karen Lemmens, Tim Van den Bulcke, Kristof Engelen, Bart De Moor and Kathleen Marchal. *ViTraM: Visualization of Transcriptional...* ......................................................... 72
P38: Ivan Ishchukov, Inge Thijs, Kathleen Marchal. *In silico study of the regulation of sRNAs in Escherichia coli* ......................................................... 74
P39: Jeroen de Ridder, Anthony Uren, Jaap Kool, Marcel Reinders, Lodewyk Wessels. *Inferring Cancer Pathways From Retroviral Insertional Mutagenesis Screens* ......................................................... 75
P41: Peter Konings & Yves Moreau. *wikiRobot* ......................................................... 77
P43: Yao-Cheng Lin, Jean-Marc Frigerio, Cyril Cabane, Pierre Rouzé, Antoine Kremer, Christophe Plomion and Yves Van de Peer. *In silico screening of Conserved Orthologous Sequences (COS) from Fagaceae ESTs resources* ......................................................... 79
P46: Philip Lijnzaad, Thanasis Margaritis, Dik van Leenen, Diana Bouwmeester, Patrick Kemmeren, Sander van Hoof and Frank Holstege. *Gene-specific dye bias correction of two-colour microarray data* ......................................................... 82
P48: Marc Jan Bonder, Rudolf Fehrmann, Lude Franke, Gerard te Meerman, Cisca Wijmenga. *Predicting the primary origin of metastasized tumors by using a principal component analysis of >33,000 microarrays* ......................................................... 84
P49: Cindy Martens and Yves Van de Peer. *The Hidden Duplication Past of the Plant Pathogen Phytophthora and its Consequences for Infection* ......................................................... 85
P51: Pieter Monsieurs, Hugo Moors, Rob Van Houdt, Paul Janssen, Max Mergeay, Natalie Leys. Heavy metal resistance in Cupriavidus metallidurans: a complex evolutionary and transcriptional process .......................................................... 87
P52: Petr V. Nazarov, Arnaud Muller, Viktor Khuiko, Loïc Couder and Laurent Vallar. CoExpress: a tool for an effective co-expression analysis of microarray data sets .......................................................... 88
P53: Huu Minh Nguyen, Yvan Saëys, Luc Knockaert and Tom Dhaene. Comparison of embedded feature selection algorithms for splice site detection ..................................................... 89
P54: Jurgen Nijkamp, Wynand Winterbach, Marcel van den Broek, Jean-Marc Daran, Jack Pronk, Marcel Reinders and Dick de Ridder. Combining a de novo assembly with mappings against multiple reference genomes ................................................. 90
P56: Catharina Olsen, Patrick E. Meyer and Gianluca Bontempi. Inferring causal relationships using information-theoretic measures .......................................................... 92
P57: Lex Overmars, Tom Groot Kormelink), Mark de Been, Roland J. Siezen and Christof Franke. Microbial Genome Viewer 2.0: Facilitating the sequence-based reconstruction of specific bacterial regulatory mechanisms ................................................. 93
P59: Dorien A. M. van Dartel, Jeroen L. A. Pennings, Frederik J. van Schooten, Aldert H. Piersma. Identification of Developmental Toxictants by their interference with Cardiomyocyte Differentiation of Embryonic Stem Cells Studied by Transcriptomics 95
P61: Jan Ramon. Efficient search in molecular graph space to recognize mass spectra .......................................................... 97
P62: Koos Rooijiers, Carolin Kolmeder, Peter Schaap. Efficient use of metagenome data in shotgun metaproteomics .......................................................... 98
P63: Jahn-Takeshi Saito, Adem Bilican, Chris Evelo. Querying the Nutritional Phenotype Database dbNP - A study querying tool for the new globally developed systems biology database for nutrigenomics ........................................................................................................ 99
P64: Katrin Sameith*, Eva Apweiler*, Marian Groot Koerkamp, Dik van Leenen, Patrick Kemmeren, Frank C.P. Holstege. Systematic investigation of the glucose signalling pathways in Saccharomyces cerevisiae through mRNA expression profiling .......................................................... 100
P65: M. Sanders, J. Klomp, S. Verhoeven, C. de Graaf, J. de Vlieg. Structure based GPCR pharmacophore construction using sequence information .......................................................... 101
P66: Shi Yu, Tillmann Falck, Anneleen Daemen, Johan Suykens, Bart De Moor, Yves Moreau. Non sparse kernel fusion and its applications in genomic data integration .......................................................... 102
P67: Shi Yu, Leon-Charles Tranchevent, Bart De Moor, Yves Moreau. Gene prioritization and clustering by multi-view text mining .......................................................... 103
P68: Olivier Sperandio, Christelle H Reynès, Anne-Claude Camproux, Bruno O Villoutreix. Rationalizing the chemical space of protein-protein interaction inhibitors .......................................................... 104
P70: S. Brohée, R. Barriot, J. Breekpot, B. Thiempont, S. Van Vooren), B. Coessens, L- C. Tranchevent, P. Van Loo, M. Gewilling, K. Devriendt, Y. Moreau. CHDWiki: An online collaborative and interactive data repository dedicated to congenital heart defects .......................................................................................... 106
P73: Léon-Charles Tranchevent, Stein Aerts, Peter Van Loo, Bert Coessens, Bassem A. Hassan and Yves Moreau. How to get the best from your array CGH data? .......................................................... 109
P74: Laura Trotta, Eric Bullinger and Rodolphe Sepulchre. Performance criteria and robustness in biological bistable models at the example of the apoptotic process. .......................................................... 110
P75: Sofie Van Landeghem, Yvan Saeys, Bernard De Baets, Yves Van de Peer. Training a text miner to summarize various bio-molecular events found in research articles.


P77: Hanka Venselaar. Project HOPE: Providing the last piece of the puzzle.


P82: Michiel Wels, Peter Bron, Anne Wiersma, Maria Marco and Michiel Kleerebezem. A functional-genomics fermentation platform to identify and optimize industrial-relevant properties of Lactobacillus plantarum.


P84: Elisabeth Wischnitzki, Yves Van de Peer, Klaas Vandepoele. Conservation of transcription factor binding sites in plants and its application to reduce false positive target genes.

P85: Peyman Zarrineh, Carolina Fierro, Bart De Moor, Kathleen Marchal. Fusion of sequence data and microarray data, a systematic approach toward cross-species comparison.

P86: Mattias de Hollander, Erwin Datema, Harrold van den Burg, Ioannis Stergiopoulos, Pierre de Wit, Roeland van Ham. Genome-sequencing of the tomato pathogen Cladosporium fulvum and comparative secretome analysis with other fungal species.


P91: Herman van Haagen, Peter-Bram ’t Hoen, Antoine de Morree, Barend Mons, Martijn Schuemie. Improving the prediction of protein-protein interactions by combining different biological sources.


P95: Bastiaan van den Berg. Logic gene network design: a CAD tool based on modularity and standardization.

P96: Tom van den Bergh, Remko Kuipers, Henk-Jan Joosten, Peter J. Schaap. Mutator a fully automated method to scan and extract mutations from full text articles.
Currently, sequencing centers all around the world (Sanger, NHGRI, Broad Institute, Beijing Genome Institute) are churning out billions of nucleotides a day. Vast amounts of data are available, but unfortunately visualization tools are lagging to keep up to give you a good look at all this data.

We present GenomeView, a stand-alone genome browser designed to visualize genomic data. There are a number of interesting features which make GenomeView a unique tool to browse many types of data.

First it has the features you would expect from any regular genome browser such as browsing your sequence with annotation, but GenomeView also allows you to do it interactively. You can load any additional data you have on your computer that is in one of the 20 or so supported file formats (fasta, gff, ...). Second, it can display multiple alignments ranging from high-level conservation plots, down to the nucleotide level. Multiple alignments of complete genomes will only become more important as more and more related genomes are being sequenced and people want to transfer knowledge from one genome to the next. Finally, GenomeView can visualize data obtained from next-generation sequencing experiments. The read-alignments are displayed as individual reads mapped to the genome, down to the nucleotide level, but also as coverage plots which may be useful in estimating relative expression levels in transcriptomics experiments.

As a genome browser, it has been integrated in the Tuberculosis Database (http://www.tbdb.org/) to visualize SNPs detected in Illumina short-read data.

Besides being able to visualize all kinds of data that can be mapped to the genome, GenomeView can also be used to edit the data. As such, it has been integrated in the BOGAS annotation platform (http://bioinformatics.psb.ugent.be/webtools/bogas/) as a gene curation tool.
P2: Information about solvent contacts improves foldability in knowledge-based protein model
Sanne Abeln, Daan Frenkel
Vrije Universiteit Amsterdam, University of Cambridge

MOTIVATION
Statistical pair potentials have successfully been used to simulate protein folding with designed proteins using simplified structural models. The interaction energies between amino acids for these potentials may be derived from statistical analysis of experimental protein structures in the Protein Databank (PDB). The resulting pair potentials may also be used in threading algorithms. In their natural cellular environment folded proteins must not aggregate. Aggregates may be cytotoxic or compromise the biological function of the proteins. Evolutionary pressure generally ensures that proteins do not aggregate in their natural biochemical environment. However, using simple pair potentials between amino acids to design foldable model proteins, leads to model proteins that aggregate at the folding temperature when multiple proteins are simulated simultaneously. We show that by including solvent interactions in the pair potentials the designed proteins remain soluble at their folding temperatures and below.

METHODS
Pair potentials between amino acids can be established by considering the number of pairs of amino acids that are close in experimental protein structures. The interaction energies between to amino acids may be estimated by considering the ratio between the number of observed close amino acid pairs and the number of expected pairs based on amino acid abundance. Here, this method is extended to include explicit solvent contacts, creating additional interaction energies for each amino acid type with water.

A simple lattice model is used to test the refoldability and aggregation of designed proteins. The solvent interactions can be included through interactions with empty lattice sites. To design foldable proteins we use a Monte Carlo optimisation of the interaction energies by changing the sequence, while keeping the structure fixed in a compact conformation (Coluzza & Frenkel 2003).

RESULTS
Large aggregates are observed at the folding temperature, when multiple proteins are simulated simultaneously with a pair potential based on amino acid interactions alone (e.g. Betancourt & Thirumalai 1999). When foldable proteins are designed and simulated with the pair potential that includes the solvent, the proteins do not aggregate at temperatures significantly lower than the folding temperature. The proteins designed with the solvent potential also show better defined hydrophobic cores.

SIGNIFICANCE
Protein aggregation, in particular amyloid formation, is associated with several neurodegenerative disease. There is therefore a high interest to model such the early aggregation stages of amyloid forming proteins. However, under physiological conditions most proteins should not aggregate; when modelling aggregation leading to amyloid formation, first an appropriate model of non-aggregating proteins should be established. The pair potential presented here allows for qualitatively realistic simulations of the aggregation behaviour in protein systems. The calculation of the potential is simple to implement, and the essential idea of adding solvent interaction is easily adaptable to more complex interaction potentials and more detailed protein structure models. Moreover, the proteins designed with the solvent included potential have a better defined hydrophobic core. This suggests, that it may be useful to include solvent-amino acid interactions in potentials for structure prediction, for example as a pair-potential for threading methods.
P3: A bioinformatician’s toolbox for ChIP & DNA methylation microarray analysis

Michiel E. Adriaens*, Andrea Romano†‡, Nigel J. Belshaw§, John C. Mathers¶ and Chris T.A. Evelo*

*Department of Bioinformatics - BiGCaT, University of Maastricht, Maastricht
†Department of OB/GYN
‡GROW - School for Oncology and Developmental Biology
§Institute of Food Research, Norwich
¶Human Nutrition Research Centre, Newcastle University, Newcastle upon Tyne

Microarray technology comes in many flavours. ChIP-on-chip is used to determine potential binding sites of DNA binding proteins genome wide. The same microarray technology and analogous protocols can be applied to detect the DNA methylation status of thousands of CpG rich regions, such as CpG islands, in one go. The underlying technology and study designs of these types of microarray analyses are extremely diverse. Hence, flexible and extendible software is needed. R is an open source statistical analysis environment and programming language. It is easily extended with additional functionality through packages. The Bioconductor project is a great example of this, which is a collection of packages to suit all your bioinformatics and biostatistics needs. R combined with Bioconductor is the open source standard for microarray data analysis, which makes R a logical choice for ChIP and DNA methylation microarray analysis. There are however many minor and major annoyances in R. Although it is flexible, it is also unstable and ever changing. Packages that your analysis workflow relied on previously may be altered beyond recognition in an updated version. In addition, some essential functionality for ChIP and DNA methylation microarray analysis is often missing or suffering from cluttered interfaces, such as retrieving sequences of enriched regions from a sequence database. We have designed an analysis toolbox in R that facilitates the analysis and biological interpretation of ChIP and DNA methylation microarray data. The package wraps around a collection of stable Bioconductor packages and is optimized for the popular NimbleGen platform. We are currently adding functionality to enable comprehensive analysis of modern high-throughput sequencing technology such as ChIP-seq.
P4: Improving antibody properties by in-silico engineering: a realistic perspective
Johan Desmet and Philippe Alard
Algonomics

We will illustrate how in silico tools facilitate epitope discovery and antibody engineering citing two recent papers in PLoS One and PNAS. In the former we found an antibody epitope on hemagglutinin of influenza which explains the multiple strains specificity. The 3D model of the complex was confirmed by mutagenesis. In the latter we performed in silico affinity maturation of anti-gastrin antibody to obtain a potential therapeutic agent. Suboptimal interactions were identified in the antibody-antigen complex model and experimentally optimized.
Since the number of RBPs encoded in eukaryotic genomes approaches that of transcription factors, suggest that the regulatory program that controls the posttranscriptional fate of mRNAs their localization, translation, and survival may prove to be nearly as diverse and complex as the regulation of transcription itself. The technology has made it feasible to generate large scale data at posttranscriptional level. There is a need to develop bioinformatics methods to build posttranscriptional regulatory network integrating diverse data sources. In this paper we demonstrate that though many methods are developed to infer regulatory network at transcriptional level, they can be used to unravel different aspects of regulation at posttranscriptional level. We show that posttranscriptional regulation is similar to the transcriptional regulation in many global properties such as scale free nature or hierarchical organization of the network. We make an attempt to unravel various aspects of posttranscriptional regulation by integrating data incrementally from different sources. We used this data to build regulatory networks at transcriptional and posttranscriptional levels. The clusters obtained from posttranscriptional profiles are functionally more coherent. The regulatory network obtained from posttranscriptional profiles is equally informative about the regulation at transcriptional level as transcriptional profiles moreover it also provides information about posttranscriptional regulation thus we suggest that conditional sampling at posttranscriptional level should be preferred than at transcriptional. We then integrated expression data with known RNA binding protein targets for functional characterization of RNA binding proteins. We further build an integrated network by adding known transcription factor target information. Finally we find 3 overrepresented network motifs in integrated networks including a novel ‘feedback loop’ specific to posttranscriptional regulations.
P6: Review: 50 years of knowledge extraction and its application in biology

Andra Waagmeester∗†, Eduard Hoenkamp† and Chris Evelo∗

∗ Maastricht University, Department Bioinformatics - BiGCaT, P.O.Box 616, NL-6200 MD Maastricht, the Netherlands
† Maastricht University, DKE, P.O.Box 616, NL-6200 MD Maastricht, the Netherlands

The increase of scientific literature - in the form of abstracts (PubMed) or Full text (PMC and others) - requires automatic methods in extracting knowledge from text, since it is hardly possible to keep up-to-date by extensive reading (Jensen et al. 2006). It is remarkable that this conclusion appears in a textbook on knowledge retrieval in 1968 (Gerard 1968). So for more than 50 years scientists have been active in developing methods for the automatic extraction of knowledge from text. Knowledge extraction is applied e.g. in biology to find novel pathway parts from PubMed abstracts (Waagmeester et al. 2009). This was achieved by a custom-made framework with various software components each with a specific task. The sequence in which order the various components are applied varies depending on the context. We would like to present three abstraction layers representing a knowledge extraction process from text. These are: (1) A corpus of text, in the form of abstracts or sets of full text, (2) the logical structure, consisting of language models, software components, software framework, thesauri, ontologies, etc. (3) The computer infrastructure used. We propose this abstraction layer as a guideline to study the feasibility of a knowledge extraction process.

References:
P7: A comparative evaluation of discretisation methods for Bayesian classifiers on high-dimensional cancer data.

Lieven Baeyens and Yvan Saeys
Department of Plant Systems Biology, VIB, Gent, Belgium and Department of Molecular Genetics, Ghent University, Gent, Belgium

Microarray and mass spectrometry data have become an indispensable tool in the association of genotypic characteristics to phenotypic differences. In this work, we aim to study the use of Bayesian classifiers to automatically predict phenotypes (e.g. cancer) from experimental data. The K-Dependent Bayesian classifier (K-DB) is a Bayesian classification model that conveniently allows us to travel between the two extremes of the dependency spectrum between the model variables: either totally ignoring all dependencies (Naive Bayes Classifier), or taking into account all possible dependencies (Full Bayesian Classifier). Depending on the computational resources or research specifications, the amount of dependence to be taken into account can be determined by the parameter K of the K-DB model, which imposes an upper bound on the number of variables between which to look for interactions.

Our first contribution is a comparative study on the performance of the K-DB classification model versus other state-of-the-art classification techniques (such as SVM and KNN). Furthermore, we looked at the influence of two (commonly used) discretisation methods on the classification accuracy. We also present a new supervised discretisation method, called K-bin minimal entropy partitioning. Finally, we provide an implementation of the KDB-model which was optimized (reduction of the computational cost in time and memory) in order to process high-dimensional datasets, as commonly used in the Bioinformatics domain.
P8: HUBase: A Human-based Expression Array Data Builder
Eric Bareke, Bertrand De Meulder, Michael Pierre, Eric Depiereux
Laboratoire de Biostatistique et de Bioinformatique, Unité de Recherche en Biologie Moléculaire,
Facultés Universitaires Notre Dame de la Paix (FUNDP)

Public available data of microarray gene expression signals represent an extraordinary opportunity for extracting genomic relevant information and validating biological hypotheses. Through microarray experiments are very popular in life science research, managing existing data is still challenging task for many researchers, and the exploitation of this exceptionally rich mine of information is still hampered by lack of appropriate computational easy-to-use tools, able to manage and to re-organize previously published microarray data in order to overcome the critical issues raised by microarray analysis.

This work presents HUBase, a web-based expression array data manager for researchers who need to organize microarray data efficiently and get new user-defined datasets built instantly. All data management and organization procedures are optimized and automated so that any user can build his/her own microarray dataset(s) for analysis. HUBase provides maximum reproducibility for each microarray dataset building process carried out.

HUBase integrates with Gene Expression Omnibus1 and allows instantaneous re-organization of previously published microarray data.

HUBase is a novel expression array data builder system. By append biological tags to previously published microarray data, HUBase allows: i. Biological tagging of public available microarray data samples in order to build user-customized batches of expression array data. ii. Batch building of expression array data from Gene Expression Omnibus.
P9: Patrocles: a database of polymorphic miRNA-mediated gene regulation in vertebrates

Samuel Hiard, Carole Charlier, Wouter Coppieters, Michel Georges and Denis Baurain

Systems and Modeling, Montefiore Institute and Unit of Animal Genomics, GIGA-R and Faculty of Veterinary Medicine, University of Liège, Belgium

The Patrocles database (http://www.patrocles.org/) compiles DNA sequence polymorphisms (DSPs) that are predicted to perturb miRNA-mediated gene regulation. Distinctive features include: (i) the coverage of seven vertebrate species in its present release, aiming for more when information becomes available, (ii) the coverage of the three compartments involved in the silencing process (i.e. targets, miRNA precursors and silencing machinery), (iii) contextual information that enables users to prioritize candidate 'Patrocles DSPs', including graphical information on miRNA-target coexpression and eQTL effect of genotype on target expression levels, (iv) the inclusion of Copy Number Variants and eQTL information that affect miRNA precursors as well as genes encoding components of the silencing machinery and (v) a tool (Patrocles finder) that allows the user to determine whether her favorite DSP may perturb miRNA-mediated gene regulation of custom target sequences. To support the biological relevance of Patrocles’ content, we searched for signatures of selection acting on ‘Patrocles single nucleotide polymorphisms (pSNPs)’ in human and mice. As expected, we found a strong signature of purifying selection against not only SNPs that destroy conserved target sites but also against SNPs that create novel, illegitimate target sites, which is reminiscent of the Texel mutation in sheep.
P10: Gene-trait matching in Lactococcus lactis strains
J.R. Bayjanov, R.J. Siezen and S.A.F.T. van Hijum

Center for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Center, the Netherlands. NBIC, Toernooiveld 1, 6525 ED, Nijmegen, the Netherlands. TI Food and Nutrition, Wageningen, the Netherlands. NIZO food research BV, Ede, the Netherlands. Kluyver Centre for Genomics of Industrial Fermentation, Delft, the Netherlands.

Identifying genotype-phenotype (gene-trait) relations serves multiple purposes: (i) screen microorganisms for desired traits using gene content, (ii) gain insight in gene function, and (iii) gain more knowledge on the effects of environmental factors that lead to gain or loss of a certain trait in an organism. A very cost-effective way for inferring gene-trait relations is to use high-throughput techniques such as comparative genome hybridization, transcriptomics, metabolomics, and phenotypic microarrays. A major bottleneck lies in the integration of these complex multivariate data. These data can better be studied using available multivariate data analysis methods rather than using conventional correlation measures. However, current methods in multivariate data analysis have in general the following shortcomings: (i) they correlate only one observed variable (e.g., phenotypic measurement) with predictors (e.g., presence or absence of genes), and (ii) the actual relation between gene presence and phenotype is difficult to extract. We are developing a method and software that allows for non-parametric integration of multivariate data to identify many-to-many relationships. The interactions between predictors and the relationship of the predictor(s) with the observed variable are presented to the user, allowing for in-depth mining of such complex datasets. The software will soon be available as a web-tool. Currently, the method is being tested on available genotype data (presence/absence of genes generated by the PanCGH method [1]) and phenotype data (growth of strains on different sugars) for different Lactococcus lactis strains. Our method allowed determining (i) gene clusters associated to a single trait and (ii) gene clusters that were associated to multiple traits. Based on the identified gene-trait relations we were able to (re)annotate gene function in a number of cases.
P11: Enzymatic protein prediction from structure using a segment and combine approach
Julien Becker, Louis Wehenkel, Pierre Geurts
Department of EE and CS & GIGA-R, University of Liège, Belgium

Numerous applications in bioinformatics exploit complex objects often represented by graphs, in which nodes correspond to elementary components and vertices correspond to relations between these components. The study of macroscopic properties of these objects aims to understand the relation that exists between the topology of the graph and the target property. In this work, we are interested in using supervised machine learning techniques to automatically build a predictive model for some target property of a class of graphs from a database of graphs of this class for which this property is known. As an example application, we consider the problem of predicting the function of a protein from its 3D structure.

The proposed method is a first attempt to apply the "Segment and Combine" principle to graphs. It consists of four steps: (i) each protein in the learning sample is converted to a graph using the 3D positions of its amino acids (ii) each graph is segmented into one or more pieces that can be represented by equal-size vectors (iii) a standard supervised learning method is applied on the sample of segments to build a model for predicting the label of each segment (which is directly derived from the label of its parent protein) (iv) Finally, a class prediction for a new protein is computed from the classifications of all its segments by the classifier.

We have explored several alternatives at each of these four steps and compared them with existing methods on a benchmark problem aiming at distinguishing enzymatic and non-enzymatic proteins. Our results in terms of accuracy show that the proposed approach is promising.
P12: The Ibidas data integration and accession platform
Jan Bot, Marc Hulsman, Marcel Reinders
Delft University of Technology

Researchers in the life-sciences are often required to integrate different data types and sources, and due to the nature of research, often in ways not done before. Uniform access to and easy manipulation of these data sources will allow a researcher to focus on the problem at hand in stead of on the data integration problem. Instead of a focus on data storage and integration we take in this work a query-centric approach. Ibidas provides a uniform, command-line based, query mechanism which can apply operations such as data filtering, joining and grouping. Data can come from files, databases or web services. The system automatically determines which part of the query can be performed by the database or web service, and which part should be performed locally. Parsers for a growing number of file formats are available.

The core of the system is written in Python and external communication is done via web-services. This makes sure that the system is easily extensible, either directly through python modules or through interfacing with the web-services. Visualization can be done using Cytoscape, which connects to Ibidas using a plug-in. Other analysis tools can query the system through web-service wrappers which facilitate communication with e.g. Matlab and R.

Ibidas is equipped with its own flexible database scheme, to support the caching of data obtained from files / web services, as well as to allow the researcher to store (intermediate) results. To publish results, a web service can be easily constructed by assigning query operations to web service-functions.
P13: Using trees as test nodes to model haplotype blocks in GWAs
Vincent Botta, Pierre Geurts, Louis Wehenkel
Department of EE and CS & GIGA-R, University of Liège

In the field of genome wide association studies (GWAS), the individuals are described by hundreds of thousands of single nucleotide polymorphisms (SNPs). Due to linkage disequilibrium, these descriptors are correlated, creating the so called haplotype blocks. In a classical GWAS, the SNP independence is (too) often assumed. We propose here to exploit the block structure inside ensemble of decision trees (Random Forest and Extra-Trees) by treating attribute groups instead of single variables within the decision tree test-nodes. We expect that haplotype blocks can be exploited to construct more efficient tests and thus better predictors. Indeed, when a SNP is associated with the disease, its neighbours are expected to be disease associated too.

More precisely, each test-node becomes a small decision tree learned from a subgroup of variables, which, in this case, correspond to a haplotype block. During the learning process, at each node, K decision trees are constructed on K randomly picked blocks, the most informative one is kept to split the learning set given a threshold on the probabilities generated by the resulting tree. Inside these tree test-nodes, the best nodes are expanded first, then tree complexity (called here the internal complexity) can be limited to use only the most informative SNPs from a block.

This approach has been applied on each chromosome of the Crohn’s disease WTCCC dataset. The results are compared to Random-Forest and Extra-Trees testing individual SNPs. The resulting models are far less complex due to the dimension reduction implied by taking into account the haplotype blocks. Also, we observe a predictive power increase with bigger values of the internal complexity. Practically on every chromosome, treating blocks instead of SNPs inside the test-nodes produced greater AUCs. Furthermore, we extended variable importances to block importances which allow us to directly highlight chromosomal regions of interests.

More generally, facing the increasing amount of data in GWAS, what we propose here is to use weak learners as test-nodes into decision trees in order to reduce the problem dimensionality given the data structure. Further works will focus our efforts on exploring the internal complexity parameter space and on testing variant weak learners (like ensemble of stumps, ensemble of trees or HMM models) as test-nodes.
P14: The diversity of clostridial hydrogenases revealed by genome sequencing projects
Magdalena Calusinska, Annick Wilmotte
Centre for Protein Engineering, University of Liege, Belgium

Molecular hydrogen is a key intermediate in metabolomic interactions of a wide range of microorganisms. Hydrogen is also regarded as a key component in future energy systems as it is a sustainable, clean, and transportable energy carrier. Some microorganisms can produce hydrogen during a reversible reduction of protons to dihydrogen, a reaction which is catalyzed by the enzymes hydrogenases. On the basis of their bimetallocenter composition, hydrogenases are divided into three main groups, phylogenetically not related: [NiFe] hydrogenases, [Fe] only hydrogenases and FeS cluster free hydrogenases. The latter were described in methanogenic Archaea only. [NiFe] hydrogenases, composed of at least two subunits are well characterized and widely distributed between Archaea and Bacteria. However, only a few representatives of Clostridium sp. possess this type of enzyme. On the other hand, much less is known about the [Fe] only hydrogenases, that are usually monomeric enzymes and restricted to Bacteria and a few eukaryotic species. Genome sequencing projects gave a completely new insight into the diversity of forms of putative [Fe] only hydrogenases within the genus Clostridium. With the use of bioinformatic tools, we have described the unusual modularity of forms of these enzymes, from monomeric to tetrameric with a different number of accessory domains reacting with diverse redox partners. This fact seems to support the central role of hydrogenases in cell metabolism and quick adaptation of the host to changing environmental conditions. Moreover, the presence of multiple putative operons encoding for multisubunit [FeFe] hydrogenases is highlighting the fact that hydrogen metabolism is very complex in the Clostridium genus.
P15: Comparison of normalization methods for Agilent microarray using single cell DNA

J. Cheng*, E. Vanneste†, P. Konings*, P. Yazbeck†, T. Voet†, J. Vermeesch†, Y. Moreau*

*Katholieke Universiteit Leuven, ESAT-SISTA, Leuven, Belgium
†Katholieke Universiteit Leuven, Center for Human Genetics, 3000 Leuven, Belgium

High density microarrays are more common being used in the recent studies, e.g. Affymetrix array and Agilent array. The advantages of the high density arrays are that they provide higher coverage and higher resolution for the genome wide analysis compared to traditional array, e.g BAC array. However, high density arrays have also brought the extra difficulty to the normalization methods. The goal of this study is to compare the existing normalization methods for the high density oligonucleotide Agilent 244K array using DNA derived from single cells. Cell line Epstein barr virus (EBV) transformed lymphoblastoids were used in this study. Because of the limited amount of the DNA obtained from single cells, amplification of DNA was performed first using a multiple displacement amplification (MDA) approach [1]. The amplified single cell DNA was then hybridized to the non-amplified genomic DNA [2]. The evaluation of the normalization methods was based on the MA plot and the segmentation results. Circular Binary Segmentation method was used to detect CNV. The karyotypes of the EBV cells were considered as the standard of the segmentations. It shows that loess normalization approach performs a little better than the linear polynomial regression provided by the Agilent Company both for the MA plot and segmentation. However, loess normalization reduced the dynamic range of the ratio between test samples and control samples. Novel normalization methods are needed to handle the intensity pattern without destroying the biological dynamic range.

P16: Regulatory gene interaction network visualization using Cytoscape

Susan LM Coort, Thomas Kelder, Chris TA Evelo
Dept. of Bioinformatics (BiGCaT), Maastricht University, Maastricht, the Netherlands

A manner to visualize interactions between genes and/or proteins is by using the free software package Cytoscape[1]. Cytoscape enables users to extend its functionality by creating or downloading plugins. The aim of the present study is to create a regulatory interaction network in which genes will be linked to their potential transcription factor(s). The network will be constructed starting with differentially expressed genes coming from a microarray experiment. Thereafter, several open-access databases will be used to find interactions between genes and transcription factors based on the transcription factor binding site motifs present in the gene. One of the databases used will be JASPAR[2]. The core of JASPAR contains curate, non-redundant matrix models to describe the binding preference to DNA sequences. Based on these models transcription factors can be coupled to the gene. Another database is ORegAnno[3], which is an open-source database and literature curation system for community-based annotation of experimentally identified DNA regulatory regions, transcription factor binding sites and regulatory variants. Importantly, both databases are open-access and can be used in Cytoscape. Once the TF-gene interaction networks are built they will be analyzed and used by biologists to find biological mechanisms of interest.

P17: CORNET: a user-friendly tool for data mining and integration
Stefanie De Bodt, Diana Carvajal, Jens Hollunder, Joost Van den Cruyce, Sara Movahedi and Dirk Inzé
Department of Plant Systems Biology, Flanders Institute for Biotechnology, and Department of Plant Biotechnology and Genetics, Ghent University, 9052 Gent, Belgium

As an overwhelming amount of functional genomics data has been generated, the retrieval, integration and interpretation of these data need to be facilitated to enable the advance of (systems) biological research. For example, gathering and processing microarray data that are related to a particular biological process is not straightforward, neither is the compilation of protein-protein interactions from numerous, partially overlapping databases, identified through diverse approaches. However, these tasks are inevitable to address the following questions: “Does a group of differentially expressed genes show similar expression in diverse microarray experiments?” “Was an identified protein-protein interaction previously detected by other approaches?” or “Are the interacting proteins encoded by genes with similar expression profiles and localization?” We developed CORNET (CORrelation NETworks) as an access point to transcriptome, protein interactome, localization data and functional information on Arabidopsis thaliana. It consists of two flexible and versatile tools, namely the co-expression tool and the PPI tool. The ability to browse and search microarray experiments using ontology terms and the incorporation of personal microarray data are distinctive features of the microarray repository. The co-expression tool enables either the alternate or simultaneous use of diverse expression datasets, whereas the PPI tool searches experimentally and computationally identified protein-protein interactions. Different search options are implemented to enable the construction of co-expression and/or protein-protein interaction networks centered around multiple input genes or proteins. Moreover, networks and associated evidence are visualized in Cytoscape. Localization is elegantly visualized in pie charts, thereby allowing multiple localizations per protein. CORNET is available at http://bioinformatics.psb.ugent.be/cornet.
P18: Methodology for benchmarking microarray data analysis methods
Bertrand De Meulder, Benoît De Hertogh, Fabrice Berger, Michael Pierre, Eric Bareke and Eric Depiereux
FUNDP, University of Namur

Background
There is many ways to analyze microarray data. Usually, one relies on a benchmark to choose the method(s) to use. There are spike-in studies and simulations, but we were not satisfied with the quality of those methods. Thus, we decided to do our own benchmarking methodology, based on biological data.

Methods
We constructed a hybrid dataset, by concatenation of 34 public domain datasets. This huge matrix (called the DB matrix) has 1,292,414 rows (probesets) and 2*15 columns (replicates). We pretreated the data using GCRMA method. We devised a theoretical framework to analyze this dataset, from classical statistical formulas. This gave us a metric (D/S) which enabled us to design subsets matrices from the DB matrix, with a chosen number of replicates, of DE and non-DE Genes and an estimated level of difficulty. We generated matrices with 2 to 10 replicates, with growing difficulty and noise (4 levels). Our metric, by design, is suited for analyzing t-derived statistics. We choose to assess the performances of the Student’s t test with and without Welch’s correction for heteroscedasticity, SAM, Regularized t test, Window t test with and without Welch’s correction, LIMMA and Shrinkage t test.

Results
To control the quality of our approach, we compared volcano plots, M vs A plots and mean VS sd plots for our DB matrix, with an independent biological dataset and with the dataset used in Choe’s spike in study. The plots showed better correlation between our and biological data than with spike-in data. We thus constructed a dataset which is biologically much better representative than previous work. As for the benchmarking results themselves: we observed similar behavior of the tested methods than in other benchmarks: the methods using a window to compute their statistic have better performances: window t test, shrinkage t test and regularized t test. When the number of replicates is high (> 8), most of the methods performances are similar to student t test. Moreover, when the number of replicates is very low (< 3), the Regularized t test and the Window t test are the best choices. In between and overall, the Shrinkage t appears to be the best choice among the methods we tested.

Conclusion & perspectives
We present a novel way to benchmark the microarray analysis methods, based entirely on biological data. We assess the quality of our approach in several ways, each showing superiority of our approach compared to previously published benchmarks. The results we present are in accordance with previous work, with better resolution nonetheless. In the near future, we plan to adapt our method to other pretreatment methods, to add more treatments and to modify our metric to assess non t-derived statistics. We are also working on adapting the methodology to gene set analysis. Finally, we plan to integrate this work in a webtool called Phoenix which we developed in our lab and that is a multi-purpose platform for microarray data generation, retrieval and analysis.
P19: A query-based framework to validate ChIP-chip targets
Riet De Smet, Hong Sun, Karen Lemmens, Inge Thijs, Kathleen Marchal
CMPG/biol, KULeuven

Despite being a useful technology to gain insight into transcription regulation, inference of regulatory interactions by ChIP-chip analysis is troubled by the identification of false positive and non-functional targets. In addition, due to the condition specific nature of the ChIP-chip assay, other targets of the screened regulator might be missed because the correct conditions were not tested for. Gene expression data provides us with an alternative view on transcriptional regulation: genes that are co-regulated are expected to be coexpressed under specific conditions. Therefore explorative tools for gene expression compendia can be used to assess the confidence of obtained ChIP-chip interactions.

Here we present a workflow to interrogate gene expression compendia for condition-dependent coexpression of ChIP-chip targets with known targets of the profiled regulator. The workflow combines query-driven biclustering with consensus clustering in order to obtain a comprehensive set of non-redundant consensus biclusters containing any of the ChIP-chip targets and additional genes with which they are coexpressed. As a proof of concept we applied our approach to ChIP-chip data for FNR in Escherichia coli. For the analysis of the obtained consensus biclusters ViTraM was used. This software allows for visualization of the consensus biclusters and their interpretation through for instance the presence of motifs, known regulatory interactions or condition annotations. We interpreted the list of ChIP-chip targets in terms of the uncovered consensus biclusters. Particularly, probable functional targets amongst the ChIP-chip targets could be identified through the co-clustering of the ChIP-chip targets amongst each other or with known targets of the profiled regulator. In addition, likely false negatives could be pinpointed as genes not identified by ChIP-chip but present in consensus biclusters highly enriched for FNR targets.

Here we proposed an approach to cross check experimental data with publicly available gene expression data. The method can take many input genes and effectively identifies non-redundant biclusters containing any of these input genes. The method was shown to be useful in distinguishing between functional and non-functional FNR ChIP-chip targets and identifying likely false negative targets.
P20: Complexity of human gene expression traits in different primary tissues
Jingyuan Fu, Patrick Deelen, Rudolf Fehrmann, Gerard te Meerman, Marcel Wolfs, Cisca Wijmenga, Lude Franke
Department of Genetics, University Medical Centre Groningen, The Netherlands
Institute for Life Science and Technology, Hanze University Groningen, University for Applied Sciences, The Netherlands

It is known that genetic differences can affect the abundance of gene expression levels (expression quantitative trait loci (eQTL)). However, these effects can differ per tissue. Not only the strength of the effect of an eQTL can be different, but it is also possible that an eQTL shows an opposite allelic effect. For example, for a genetic variant with two alleles (A and B) it can be that the A allele associates with high expression in one tissue but associates with low expression in another tissue, compared to the B allele.

To detect such eQTL a meta-analysis was conducted to detect cis-eQTL by correlating gene expression levels to SNP genotypes (using a 500kb SNP-probe window size) in subcutaneous abdominal tissue, visceral adipose tissue, liver and muscle from the same 86 individuals and peripheral blood from 1,240 individuals.

We observed over 80,000 eQTL in total at an FDR of 0.05. We found that the effects of eQTL can differ substantially among different tissues. Over 100 eQTL showed an opposite allelic effects when comparing these five tissues with each other. We studied the SNPs and genes that constitute these eQTL and what their biological properties are, by assessing selection, methylation, protein-binding and disease association. For some of these properties we observed substantial differences. These results show the complexity of the genetic architecture of gene expression in primary tissues. The opposite allelic effect appear to be rare but seem to be biologically relevant. The results suggest that the usage of relevant tissues is essential when using eQTL or gene expression data to interpret disease pathology.
Background: Iron-Sulfur Clusters (ISC) are prosthetic groups present in metalloproteins that are involved in several cellular mechanisms, namely in electron transfer, enzymatic catalysis and regulation of gene expression. These co-factors are assembled in the Eukaryotic cell via a well structured cascade of events that involves more than 20 proteins and takes place both in the mitochondrion (mit ISC machinery) and in the cytoplasm (CIA machinery). Most of the proteins involved in the biogenesis of these Fe-S clusters are encoded by essential genes, showing the importance of this process for the cell. The question about the phylogenetic origin of the CIA machinery surfaced when researchers realized that the mitochondrial part of the assembly in Eukaryotes is very similar to the one that takes place in Prokaryotic cells, but unparalleled in its cytoplasmatic portion. Hence, identifying its phylogenetic origin might reveal some other direct or indirect functional interactors that can help in the understanding of the importance of this machinery for cell viability, opening new venues of research in the area.

Methods: In order to accomplished the proposed goal, we built a dataset of protein sequences involved in this pathway, followed by homology sequence search and retrieval, multiple sequence alignment and phylogeny building. After the manual curation of the data, an analysis of co-occurrence will be performed aiming at the prediction of novel protein interactors.

Results: Preliminary results indicate that the proteins that comprise the cytoplasmatic part of ISC assembly (4 well characterized and 2 putative functional interactors) don’t have all the same phylogenetic origin. This can be a hint that each of these proteins might have been involved in a different cellular process before becoming involved in the cytoplasmic ISC biogenesis.
P22: A QC pipeline for one and two channel microarrays implemented in R

*Department of Bioinformatics - BiGCaT, Maastricht University, Maastricht, The Netherlands
†Department of Human Biology, Maastricht University, Maastricht, The Netherlands
‡Nutrigenomics Consortium, Top Institute Food and Nutrition, Wageningen, The Netherlands

Accuracy of measurements and data quality are very important aspects in the field of microarray research where massive amounts of data are generated. Data filtering and correction ensures that downstream analysis will not be biased by any technical influences due to lab processing. In contrast to the array platforms, which are very reproducible and comparable as investigated by the Microarray Quality Control Consortium (MAQC), sample preparation, hybridisation and/or washing steps may strongly influence data quality. Software that generates images to help the user deal with global quality assessment of each microarray allows identification of anomalies or groups of outlier reporters. Tools are available that allow flagging certain bad quality areas or spots on the array. Additionally, platform-specific pre-processing tools currently have their own built-in quality checks on both the global array-wide scale as on the individual reporter level. Spot-specific flagging allows inclusion of as many data points as possible in the analysis, while still removing low quality measurements. However, due to lack of automated procedures, users will often not perform (extensive) quality control (QC) or decide to simply discard a complete array from the experiment, even if only parts of it are of low quality, which results in needless loss of data or necessitates redoing the bad arrays with the associated risk of introducing new biases. Proper QC will be relevant not only for newly generated data, but also when reusing or integrating data from array repositories, such as ArrayExpress or Gene Expression Omnibus. Here we present an automated array quality control pipeline (arrayQC) that will facilitate and speed up the process of identification, filtering, and correction of aberrations by integrating the tasks described above. It will produce several user-friendly and easy to interpret images, providing spot, array and overall data-set QC information. These include virtual array plots, summarisation bar charts, MA and density plots, boxplots, heatmaps and many more. Several plots are generated before and after normalisation, allowing users to assess which normalisation is able to remove the aberrations present, if any. In addition, the pipeline will generate weights based on several quality criteria that are included in the text files produced by either software (including spot size and uniformity, expression above background level, and saturation). These weights can be used for flagging in further analysis. The pipeline can be applied to any single or dual channel data set produced with either the Agilent Feature Extraction Software or the GenePix Pro software. Moreover, it has been written in R, a commonly used open source programming environment, facilitating modification, extension, and integration with other code. The pipeline aims at two groups of users: it is implemented to be used as an automated procedure with default settings in an easy way, but it is also prepared to be modified to specific needs by the more experienced user. The pipeline and detailed documentation are available online from http://svn.bigcat.unimaas.nl/r-packages/arrayQC and the R source code is fully accessible.
Assembling biological pathways from information in scientific literature and biological databases is a challenging task. Building a pathway often requires domain knowledge from specialists in various biological research areas. Furthermore, research is generating new biological knowledge continuously, making pathway curation an ongoing process. Classical pathway databases often lack resources to stay up-to-date with this new knowledge. Community-based curation of biological knowledge takes on this challenge by giving the biology community the power to curate their own data repositories [1]. WikiPathways [2] applies this concept on biological pathway information by presenting biological pathways in the form of a wiki. WikiPathways currently has over 900 registered users and contains more than 1200 pathways for various organisms, including human, mouse, rat, fruit fly, worm, zebra fish and yeast. Pathways are stored in an XML-based pathway format (GPML) that can include annotations to gene, protein and metabolite databases, citations to literature and interactions [3]. Biological entities on the pathway can be cross-referenced between various biological databases and experimental systems using synonym databases based on Ensembl and the Human Metabolite Database. All pathways are compatible with a wide range of analysis tools including Cytoscape for network-based analysis and GenMAPP for visualization and analysis pipelines written in e.g. R, Java, Python or Perl. The critical success factor for community curation is an active user community. WikiPathways attempts to build an active community in several ways. Contributing users are credited for their work, for example by presenting an author list on each pathway. Additionally, usability of a pathway as research tool is an important incentive for users to start creating pathways. Therefore, pathways on WikiPathways have a flexible graphical format and can be exported to several graphics formats and analysis software, in order to support a wide range of end-user applications. Finally, WikiPathways aims to lower the threshold for new users to start contributing by providing an easy to use pathway editor and curation tools. Examples of curation tools are the possibility to revert a pathway to any previous version, labeling pathways according to quality standards, email notification on pathway changes and a discussion page for each pathway.

References:
P24: Colombos: the access port to the cross-platform microarray compendium

F. Qiang, K. Engelen, P. Meysman, K. Lemmens, R. Desmet, C. Fierro, I. Thijs, K. Marchal
CMPI-bioi

Public microarray databases, such as Gene Expression Omnibus (GEO) (Barrett et al, 2007), ArrayExpress (AE) (Parkinson et al, 2007), and the Stanford Microarray Database (SMD) (Demeter et al, 2007), contain freely accessible data for thousands of experiments and a multitude of different organisms. These data are not useful as such however, due to a large inherent heterogeneity: different experimenters/origins, different technological platforms, etc. We have created a (semi-automatic) system to create expression compendia, i.e. to collect, homogenize, and annotate public microarray data for a specified organism, which offers the advantage of maximally exploiting publicly available information. Furthermore, we are implementing a web interface called Colombos, which provides a set of convenient interfaces to ease the access of the data, the visualization tools to magnify the result, and the links to external website for extra information.

Data collection takes three steps. First, the data from GEO, AE, and SMD are automatically retrieved, and semi-automatically parsed into a stand format. Then the collected data are homogenized through an array-specific and platform dependent preprocessing step, which renders expression levels comparable between different experimental conditions and platforms. Finally, a formal, parameterized description of the experimental conditions (manually curated) is incorporated in the compendium. It further improves the comparability across experiments.

The generated compendium will be publically accessible through the Colombos web interface. The user could download the whole compendium for their own analysis necessity, or select specific sections of the compendium as a module for more focused studies. The visualization ability is provided to either highlight the user created module, which could provide clues for possible extensions, or an optimized presentation of overlapping modules. Furthermore, external links to other public websites are automatically generated to access extra information presented there.

Compendia have been generated using our methodology for Escherichia coli, Salmonella enterica and Bacillus subtilis. The web interface is in development, and to be expected online this winter. The compendium generation methodology is being extended to handle also eukaryotic organisms.
P25: Characteristics of recent intron-losses in Arabidopsis
Jeffrey Fawcett, Yves Van de Peer
Plant Systems Biology, VIB, Gent University

Despite having no obvious function, introns are a common feature of eukaryotic genomes. Recent studies have suggested that the common ancestor of animals, plants and fungi must have been very intron-rich, and that most lineages have subsequently experienced a much higher rate of intron-loss compared to intron-gain. However, it is still not clear what exactly drives the loss and retention of introns. A large number of recent events is likely to enhance our understanding about these processes. To this end, we studied the gain, loss, and retention of introns in the genomes of two closely related species Arabidopsis thaliana and A. lyrata. Several intron-losses but very few intron-gains in the two species since their divergence were identified, and it was found that A. thaliana has lost about 4 times more introns than A. lyrata. We observed a strong bias towards short introns being more likely to be lost, suggesting that selection for transcription efficiency is unlikely to be the main driving force behind the loss of introns. Our results also suggest that certain genes are more likely to lose introns than others.
P26: Biclustering of Gene Expression Data using Probabilistic Logic Learning
Nima Taghipour, Daan Fierens, Hendrik Blockeel

We consider the problem of discovering biclusters in gene expression data by means of machine learning. The data contains the measured expression levels of the genes of a particular organism under a number of varying conditions. The learning task given such a dataset is to find subsets of genes that are co-expressed under subsets of conditions (such a subset of genes together with the corresponding subset of conditions is called a bicluster).

The problem of biclustering gene expression data has already been tackled using probabilistic model-based biclustering. So far, this approach was implemented in a special-purpose system [1], although there are a number of general-purpose probabilistic modelling systems that also appear suitable for solving this problem. A solution in a general-purpose system would have the advantage of being easily adaptable and extensible, for instance with respect to additional data sources about the considered genes [1]. The goal of this work is to investigate how well the problem of biclustering gene expression data can be solved with a number of general-purpose probabilistic modelling systems. Concretely, we consider so-called probabilistic logic learning (PLL) systems, which use elements of first-order logic for the sake of expressivity. PLL is currently a very popular approach in the artificial intelligence and machine learning community.

In our work, we first made an analysis of the modelling- and learning-features required to solve the biclustering problem (such as the ability to deal with numerical data, with overlapping clusters, etc.). Next, we made an overview of which of these features are supported by which PLL systems. In this analysis, the PLL system Alchemy (that deals with so-called Markov Logic) appeared to be the most promising. Hence, we continued by implementing probabilistic model-based biclustering in this system. This work showed that there are several practical problems that make it impossible to represent the desired model in the Alchemy system. We report the problems encountered (limitations of Alchemy) as well as the aspects of the biclustering task that can easily be modelled in Alchemy (strong points of Alchemy). In the light of these limitations and strong points, we also compare Alchemy to the other PLL systems considered in our initial analysis.

From the perspective of biological applications, our discussion is relevant in the sense that we give some insight into what kind of problems can and cannot easily be tackled using popular general-purpose systems. From the perspective of informatics, in particular machine learning, our discussion is relevant in the sense that we identify a number of shortcomings of existing systems and corresponding directions for future work.

P27: A Systems Bioinformatics Approach For Evaluating And Translating Drug-Target Effects In Disease Related Pathways

Wilco Fleuren, Jacob de Vlieg

Computational Drug Discovery (CDD), Nijmegen Centre for Molecular Life Sciences (NCMLS), Radboud University Medical Centre, Nijmegen, The Netherlands

One of the reasons for the high attrition rates in drug development is the difficulty of translating drug-target effects in disease related pathways from animal models to a human setting. System level knowledge of the genes and pathways involved in human diseases and detection of critical differences between model organisms and human is therefore essential to choose the most appropriate animal model for testing candidate drugs and to correctly interpret data from these experiments in the light of human physiology. In this project, methods will be developed to construct gene networks that are involved in human diseases and to map sequence and -omics data obtained from humans and animal models onto these networks. These methods will be applied for construction and analysis of a network for human Rheumatoid Arthritis (RA). Because of the wide range of physiological processes underlying this disease, this is particularly suitable for a systems analysis approach. The analysis will result in identification of pathways and genes that are active during various stages and disease phenotypes of RA and affected by drug treatment, yielding candidate biomarkers for patient stratification, disease progression and drug efficacy. Additionally the mapping of data of animal models for RA onto this network will provide insight into the relevance of these animal models for studying RA and allow for a rational choice of a specific animal model to study drug responses and pathophysiological processes underlying RA in human. To evaluate and refine this approach, experiments to validate the predictions of candidate biomarkers and drug response in animal models will be performed in relevant in-vitro and in-vivo systems.
The treatment of chronic-like illnesses such as HIV infection, cancer or chronic depression implies long-lasting treatments that can be associated with low quality outcome, painful side effects and expensive costs. To enhance these treatments, clinicians often adopt what we call Dynamic Treatment Regimes (DTRs). DTRs are sets of sequential decision rules defining what actions should be taken at a specific instant to treat a patient based on information observed up to that instant. Ideally, DTRs should lead to treatments which result in the most favorable clinical outcome possible. Since a few years, a growing research community is working on the development of formal methods (mainly issued from mathematics, statistics and control theory) that allow to infer from clinical data high-quality DTRs.

We propose in this framework an algorithm of quadratic complexity that infer from clinical data a sequence of treatment actions. The algorithm (called CGRL for Cautious Generalization for Reinforcement Learning) has cautious generalization properties, i.e. it avoids taking treatment actions for which the sample of clinical data is too sparse to make safe generalization. The algorithm also has consistency properties, which means that when the sparsity of the set of clinical data decreases to zeros, the inferred sequence of treatment actions is actually optimal. Moreover, in some favorable cases, some tight performance guarantees on the inferred sequence of treatment actions can be computed. The algorithm is illustrated using some simulated data dealing with the HIV infection.
Boolean networks (BNs) have been used for several decades as models of biochemical networks, mainly to predict their qualitative properties and, in the case of genetic regulatory networks, to infer the inputs and interaction rules of their nodes from microarray data. An $m$-node BN model consists of $m$ interacting nodes, where the nodes, in the context of biochemical networks, generally represent genes or molecular species (e.g. proteins). The dynamical state of each node is described by a Boolean variable whose value at discrete time $k+1$ depends on the states of some nodes of the network (called the inputs of the node) at previous time $k$ through an interaction function. According to those interaction functions, the state space of the BN, which is of size $E = 2^m$, is most often partitioned into disjoint sets called basins of attraction. Each basin contains an attractor which is commonly interpreted as a functional, phenotypic cellular state. A noisy BN is a BN where each node has probability $p$ to switch its state to the opposite state between any two times $k$ and $k+1$. With such random perturbations, transitions between the basins of the network are allowed and the resulting stochastic process can be described using a discrete-time homogeneous Markov chain with state space size $E$. As the number $m$ of nodes increases, the size $E$ of the state space increases exponentially. This is the so-called state space explosion problem. The presentation will be concerned with the reduction of a noisy BN to a coarse-grained Markov chain with state space size equal to the number of basins of the original network. We will show that in the low $p$ regime and under certain conditions, the transitions between the basins of the network can be described using an approximating system of ordinary linear differential equations. Also we will illustrate the reduction method and give some results concerning the mean sojourn time spent in a basin of attraction.
P30: CoPub Discovery: a literature mining tool for hidden biological knowledge

Raoul Frijters, Marianne van Vugt, Ruben Smeets, René van Schaik, Jacob de Vlieg, Wynand Alkema
Computational Drug Discovery (CDD), Nijmegen Centre for Molecular Life Sciences (NCMLS), Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

The scientific literature represents a rich source for retrieval of knowledge on associations between biomedical concepts such as genes, diseases and cellular processes. A commonly used method to establish relationships between biomedical concepts from literature is via co-occurrence. Apart from the use in knowledge retrieval, the co-occurrence method is also well-suited to discover new, hidden, relationships between biomedical concepts following a simple ABC-principle, in which A and C have no direct relationship, but are connected via shared B-intermediates. In this paper we describe CoPub Discovery, a tool that mines the literature for new relationships between biomedical concepts. Statistical analysis using ROC curves showed that CoPub Discovery performed well over a wide range of settings and keyword thesauri. We subsequently used CoPub Discovery to search for new relationships between genes, drugs, pathways and diseases. Several of the newly found relationships were validated using independent literature sources. In addition, new predicted relationships between compounds and cell proliferation were validated and confirmed experimentally in an in vitro cell proliferation assay.
P31: How to Fold a Protein using Amino Acid Interaction Networks
Omar Gaci and Stefan Balev
Le Havre University, LITIS Laboratory 25 rue Philippe Lebon, 76600 Le Havre France

Proteins are biological macromolecules participating in the large majority of processes which govern organisms. The roles played by proteins are varied and complex. Unlike other biological macromolecules, proteins have complex, irregular structures. They are built up by amino acids that are linked by peptide bonds to form a polypeptide chain. In their natural environment, proteins adopt a native compact form. This process is called folding and is not fully understood. The process is a result of interactions between the protein's amino acids which form chemical bonds. The biological function of a protein depends on its structure.

The experimental (in vitro) methods of structure resolution are still expensive and slow. That is why molecular biology resorts to computational (in silico) methods of structure determining. Our approach consists in representing proteins by amino acid interaction networks. The nodes in these networks correspond to amino acids. Two nodes are linked if the distance between the corresponding amino acids in the folded protein is below a certain threshold. Ignoring details, such as the type and the exact position of each amino acid, this abstract and compact description allows focusing on the interactions' structure and organization. Interaction networks for proteins of known structure can be easily obtained using the information available in Protein Data Bank. Here, we propose a means to fold this type of graph. Thus, we rely on the Levinthal hypothesis, since the folding process is oriented we observe three steps.

First, we limit the topological space by associating to an unknown sequence a structural family by topological inferences. We consider a comparative model which allows associating a structural family in SCOP classification. We evaluate the prediction reliability by computing the RMSD rate between the sequence and proteins from the predicted family. Second, we develop a genetic algorithm to construct the interaction network composed by secondary structure elements. This method enumerates the motifs existing in the predicted family to predict the most compatible with the sequence. Third, we use an ant colony approach to predict the interactions involved in the folded protein. To lead our topological folding process, we rely on a probability that two amino acids interact as a function of their physico-chemical properties.

We fold two hundred unknown sequences by our approach and we measure the prediction reliability by comparing the graphs obtained with the real ones. The results we obtain depend on the sequence size. For small proteins the resulting graphs have a similarity of 70% whereas for sequence counting more than 200 amino acids the similarity decreases.

Our model relies on topological properties to guide a folding simulation in the topological pathway from unfolded to folded state.
P32: Application of WikiPathways and PathVisio for pathway interpretation of toxicogenomics data

P. Giesbertz\(^{*}†‡\), S. Gaj\(^{*}†‡\), D. Jennen\(^{*}†\), T. Kelder\(^‡\), C. Evelo\(^{*}‡\), J. van Delft\(^{*}†\), J. Kleinjans\(^{*}†\)

\(^{*}\)Netherlands Toxicogenomics Centre, Maastricht University, The Netherlands
\(^†\)Dept. of Health Risk Analysis and Toxicology, Maastricht University, The Netherlands
\(^‡\)Dept. of Bioinformatics - BiGCaT, Maastricht University, The Netherlands

Current high-throughput technology in genomics creates a large amount of biological data. Bioinformatics approaches are directed towards understanding such data on a systems biology level. Advanced mathematical methods like principal component analysis, clustering, neural networks, support vector machine (SVM) approaches and neural networks can help to find patterns in the data. However, to really understand the data the patterns must be combined with existing knowledge. One of the approaches to do so is to associate these data to functional classifications such as can be found in the Gene Ontology. Other methods focus on using biological pathways coming from both public and private pathway databases like KEGG, WikiPathways, Reactome, and MetaCore. Some of these pathway databases contain rather aspecific information about which genes are involved in reactions. A single identifier, like an Enzyme Code, may correspond to a full family of enzymes, whereas only one family member is responsible for the reaction in the given biological context. Another problem arises in the availability of relevant biological pathways for both rodents and humans. This is especially true in the field of toxicology, where many pathways are still lacking.

The focus of this project was to create phase I and phase II detoxification pathways using PathVisio, the pathway editor of WikiPathways. New pathway content was generated using information obtained from several trusted resources: a) handbooks, b) manual/automated PubMed literature searches; c) online pathway resources; d) UniProt Knowledgebase; e) EnsEMBL; f) GeneCards. In addition, the content of some existing toxicology pathways was improved. To illustrate the usefulness of the new pathways data from experiments with carcinogenic compounds were retrieved from the main online microarray data repositories Gene Expression Omnibus (GEO) and ArrayExpress. The resulting data were visualized on the new pathways using PathVisio. All these pathways are made available to the community at WikiPathways, where they can be further used for statistical pathway analysis and visualization in the (toxico)genomics field.
P33: Biological network completion with kernel learning algorithms
G. Meyer, P. Geurts and R. Sepulchre
Departement of Electrical Engineering and Computer Science GIGA, Bioinformatics platform
University of Liège

This poster addresses the problem of completing biological networks such as protein-protein interaction networks or enzyme networks using various external sources of experimental data. Typically, this kind of networks is only partially known. The purpose is thus to exploit at best the side information available (expression data, phylogenetic profiles, localization data...) in order to infer the most plausible missing interactions in the network.

When the network is seen as a graph embedded into a kernel structure, the problem can be formulated as the computation of a kernel matrix that satisfies as much as possible a given set of constraints constructed from the external data sets. Similarities defined by the resulting kernel are then used to predict the most promising interactions.

The problem is tackled with a kernel learning algorithm that we recently developed. The algorithm overcomes some major limitations of previously proposed methods. In particular, it scales to high-dimensional problems and handles efficiently low-rank kernel matrices.

Numerical experiments have been carried out on a protein-protein interaction network as well as a metabolic network. Good performance are obtained in preliminary experiments on these two data sets.
The basic leucine zipper (bZIP) protein family is one of the largest dimerizing transcription factor families, found in all eukaryotes. Its name is derived from the highly conserved basic region (BR) and the leucine zipper (LZ) domains, which are responsible for DNA binding and dimerization, respectively. By dimerization, a small number of monomers may combine into a very large number of dimers, either homo- or hetero-, with unique functions. Arabidopsis thaliana bZIP factors (AtbZIP) are known to play a key role in biological processes such as pathogen defence, light and stress signalling, seed maturation and flower development. Previous study on AtbZIP factors identified 75 to 80 distinct members of the bZIP family which were subdivided into ten diverse functional groups. In their study, AtbZIPs were clustered based on sequence similarities of their basic region and groups of bZIPs were defined with a similar basic region and conserved MEME motifs. Using similar approach, 89 bZIP factors from Oryza sativa (OsbZIP) were uncovered. However, apart from identified AtbZIPs and OsbZIPs, bZIP factors from other plants have not been studied thoroughly. The goal of this study is to identify bZIP factors and families in plants and unicellular genomes and the major phases of expansion of the bZIP family and the evolution pattern of dimerization of bZIPs in plants and unicellulars. Different groups of plants bZIP factors are built based on phylogeny as well as domain architecture instead of simply on sequence similarities.
P35: structural features and interaction defects of mutated proteins associated with human Mendelian disorders

Fabien Heuze and Benoît Charloteaux

CHU - GIGA, Liège

Appropriate function of living cells is carried out through an integrated network of interacting macromolecules (DNA, RNA or proteins) and metabolites connected by biochemical or physical interactions, represented in interactome models by “nodes” and links or “edges”, respectively. Alterations of this intricate cellular machinery trigger human genetic diseases. Adequate modeling to decrypt the underlying mechanisms of these disorders is obviously a necessity. Human genetic diseases are thought to result in general from major alterations of genes that lead to the absence of any functional product encoded by those genes. Here, we distinguished two types of interactome perturbations that correspond either to a gross alteration of the gene product in the “loss of gene product” or “node removal” model, or to perturbations of specific interactions in the “edge-specific” or “edgetic” perturbation model. In a test of this model, we built an extensive, non-redundant and representative dataset of 3,664 missense mutations associated with human Mendelian disorders that could be mapped on high quality protein 3D structures. Our hypothesis was that buried mutations preferentially lead to “node removal”, whereas mutations at the surface of the protein preferentially lead to “edgetic perturbations”, causing different diseases. By structural analyses we found that (i) mutated residues mapped on the protein 3D structures are on average more buried than all the residues of these structures, (ii) about one third of the mutations are largely buried while another third are largely exposed, the last third occupying an intermediate position, (iii) mutated residues associated with autosomal dominant diseases are on average more accessible than the ones associated with autosomal recessive disorders. These results demonstrate our model, since both types of network perturbations, “node removal” vs “edgetic perturbations”, are well represented and induce distinct types of diseases, autosomal recessive vs autosomal dominant, respectively.
P36: ViTraM: Visualization of Transcriptional
Hong Sun, Karen Lemmens, Tim Van den Bulcke, Kristof Engelen, Bart De Moor and Kathleen Marchal
KULEUVEN, CMPG/BIOI

1. Previous studies have unveiled the modular organization of the transcriptional network. Module
detection tools are developed to reveal this modularity by identifying biclusters (or modules), i.e.
groups of genes that show a similar expression profile in a subset of experiments. Some transcrip-
tional module detection tools go one step beyond: not only do they search for the modules, but
also the regulatory program responsible for the observed co-expression behavior of the genes in the
module is identified. Having a visual overview of how these modules overlap, gives insight in the
structure of the biological system. The problem with visualizing overlapping modules simultane-
ously, however, is that the overlap in multiple dimensions complicates the choice of an appropriate
layout.

2. We developed ViTraM, a tool that allows visualizing overlapping transcriptional modules in
an intuitive way. By visualizing not only the genes and the experiments in which the genes are
co-expressed, but also additional properties of the modules, such as the regulators and regulatory
motifs that are responsible for the observed co-expression, ViTraM can assist in the biological
analysis and interpretation of the output of module detection tools.

3. ViTraM, is developed in Java (Apache Batik) and uses Scalable Vector Graphics for an attractive
visual layout. Key aspects of this tool include the implementation of a layout algorithm for the
overlapping modules the creation of a visually attractive representation allowing for interactivity
display of additional information on the modules automatic parsing of input from (XMLCreator)

P37: Inferring regulatory networks from expression data using tree-based methods
Vân-Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel and Pierre Geurts
University of Liège

We propose two procedures for inferring a gene regulatory network from expression data that both exploit supervised learning methods, in particular tree-based ensemble methods such as random forests. In the first procedure, a predictive model is learned for each gene in turn using as inputs the expression of all the other genes. The score of each gene as a regulator for the target gene is then computed from the gene ranking derived from the tree-based variable importance measure. The second procedure is an adaptation of the first one to take into account regulatory modules, i.e. groups of genes regulated together by one or several gene(s). The approach combines gradient boosting with multiple outputs trees to learn iteratively the (potentially overlapping) gene modules and their regulators. The performance of the two methods is assessed using 445 Escherichia coli Affymetrix arrays and 3216 known E. Coli regulatory interactions from RegulonDB. Preliminary results show that both approaches are competitive with existing methods on this problem.
P38: In silico study of the regulation of sRNAs in Escherichia coli
Ivan Ishchukov, Inge Thijs, Kathleen Marchal
Department of Microbial and Molecular Systems, K.U.Leuven

Studying the transcriptional regulation of genes, and by extension sRNAs, has great value in the reconstruction of regulatory networks. RegulonDB, for instance, contains significant information on transcription factors (TF) and their target genes in Escherichia coli, but information on the regulation of sRNAs is still short. Moreover, most studies are small scale and focus on the detailed analysis of a particular sRNA by combining a screening for regulatory binding sites with wet lab validation. To study the transcriptional regulation of sRNAs on a genome-wide scale, we used a combined strategy of motif screening and phylogenetic footprinting. We therefore annotated the presence of sRNAs in the genomes of related species and performed phylogenetic footprinting on the upstream regions of the sRNAs to identify known and novel TF binding sites. For phylogenetic footprinting we used the advanced probabilistic method PhyloGibbs that searches for motifs in aligned homologous DNA sequences by means of Gibbs sampling. Retrieved motifs were then compared with known TF motifs (RegulonDB) by means of the STAMP algorithm. In order to validate the identified TF-sRNA interactions, we calculated the functional enrichment of the targets of both the sRNA and the identified TF. We could show that some of the predicted pairs TF—sRNA have enriched GO terms in both target lists (for TF, and for sRNA). 15 novel reliable predictions on TFs regulating sRNAs were found.
Cancers are caused by an accumulation of multiple independent mutations that collectively deregulate cellular pathways, e.g., such as those regulating cell division and cell-death. Retroviral insertional mutagenesis is a screening technique to unravel these pathways by virally inducing tumors in mice. The insertion loci indicate putative cancer genes. Additionally, the presence of multiple independent insertions within one tumor hints towards a cooperation between the insertionally mutated genes.

To analyze data from these high-throughput screens we introduced the Gaussian kernel convolution (GKC) framework, a method to detect Common Insertion Sites (CISs), i.e., regions in the genome that harbor significantly more integrations than expected by chance. A two-dimensional extension of this method is capable of detecting statistically significant co-mutations, indicating putative collaboration between cancer genes.

For a subset of the virally induced tumors transcriptional profiling was performed. This allowed us to associate deregulated expression of genes with the presence of nearby insertions, as well as insertions in upstream regulators. However, due to collaboration and/or mutual exclusivity of integration sites, clear one-to-one associations are not frequently observed. To alleviate this, we propose to combine integration sites through small Boolean logic networks and infer associations between transcriptional state and the output of these networks. In addition to revealing complex associations that would have gone unnoticed, these networks also give insight in which genes may collaborate or substitute each other and through which putative pathway topologies their effect on transcription levels may be established.
Systems-level analyses of biological processes would benefit from accurate definitions of which proteins function as part of which multi-subunit complexes. High-throughput mass-spectrometry data offers the possibility of systematically defining stable protein complexes. The composition of complexes predicted from mass-spectrometry data differs greatly depending on what approach is used, suggesting that a consensus approach is warranted. We determine a consensus map of 409 protein complex compositions for Saccharomyces cerevisiae, by merging previous predictions with a new approach. Various analyses indicate that this consensus is comprehensive and of high quality. For 85 out of 259 complexes not recorded in GO, literature search revealed strong support in the form of co-precipitation. New complexes were verified by an independent in vivo interaction assay and by gene expression profiling of strains with deleted subunits, often revealing which cellular processes complexes are involved in. The consensus complexes are available in various formats including a merge with GO encompassing 518 protein complex compositions. The utility is further demonstrated by comparison with binary interaction data to reveal interactions between stable complexes.
Modern genetics research generates huge amounts of data which can be difficult to report efficiently. We present wikiRobot, an R package providing an interface to mediawiki servers, enabling interactive reporting of statistical and bioinformatics results.

wikiRobot consists of two different parts: an interface to the mediawiki API, and a Sweave driver for mediawiki. Mediawiki is a popular server, used by e.g. Wikipedia. Other R interfaces to wikis exist, but they are either aimed at providing R documentation or require an R installation on the server. In contrast, wikiRobot can be run on a remote computer, for example the analysts laptop.

The first part of wikiRobot is a Sweave driver for the mediawiki markup language. Sweave provides a flexible framework for mixing text and R code for automatic report generation; at run time R code in a template is replaced by its output. This is commonly used to generate reports, enabling reproducible research in statistics and bioinformatics (Leisch, 2002). Up until now drivers existed for noweb, LaTeX, HTML and Open Office. The resulting mediawiki markup files can be uploaded manually or using the mediawiki API.

Mediawiki installations provide an API that can be used by bots - small programs written in a scripting language such as perl. wikiRobot provides an interface to all relevant functions of the mediawiki API. Additionally, file and image uploading are supported, as these are not provided by the API s.s. yet.

For most mediawiki installations, including Wikipedia, one needs to register a user account for the R bot. Apart from this requirement wikiRobot is completely self-contained. We demonstrated its use in an array-CGH experiment for Crohn disease, where all array diagnostics and analysis were reported using the software. Biologists and clinicians annotate interesting results, and their comments can be traced back using the built-in features of the servers, such as edit history per page or per user. The end results can be easily imported into a word processor for inclusion in articles.

wikiRobot is available under the GPL for all platforms that support R and the XML and Rcurl packages.
Ten years of experience with Molecular Class Specific Information Systems (MCSIS) such as with the hand-curated G protein-coupled receptor database (GPCRDB), or the semi-automatically generated nuclear receptor database (NRDB) has made clear that a wide variety of questions can be answered when protein related data from many different origins can be flexibly combined. MCSISes revolve around a multiple sequence alignment that includes “all” available sequences from the entire super-family, and it has been shown at many occasions that the quality of this alignments is the most crucial aspect of the MCSIS approach.

We describe here a system, called 3DM that can automatically build an entire MCSIS. 3DM bases the multiple sequence alignment on a multiple structure alignment, which implies that the availability of a large number of super-family members with a known structure is a requirement for 3DM to succeed well.

Twelve MCSISes were constructed and placed on the internet for examination. These systems have been instrumental in a large series of research projects related to enzyme activity or the understanding and engineering of specificity, protein stability engineering, DNA-diagnostics, drug design, etcetera.
P43: In silico screening of Conserved Orthologous Sequences (COS) from Fagaceae ESTs resources

Yao-Cheng Lin*, Jean-Marc Frigerio†, Cyril Cabane‡, Pierre Rouzé*, Antoine Kremer†, Christophe Plomion† and Yves Van de Peer*

*Bioinformatics and System Biology, VIB Department of Plant Systems Biology, Ghent University, Ghent, Belgium. †INRA, UMR BIOGECO, Cestas, France ‡Université de Bordeaux 2, Bordeaux, France

Background: The conserved orthologous sequences (COS) are single or low copy number genes in a set of plant genomes and have been developed for many plant species as the anchor markers for evolutionary genomics studies and comparative mapping. Two phylogenetically close genera (Quercus and Castanea) show strong macrosynteny among 12 linkage groups based on oak EST-derived markers. With the newly available ESTs resources in Fagaceae family, we identified the putative COS genes among them.

Result: We took advantage of sequence data from four sequenced dicot plant genomes (Arabidopsis thaliana, Poplar trichocarpa, Medicago truncatula and Vitis vinifera) to overcome the potential incompleteness of the ESTs derived unigene sequences. The Fagaceae unigene libraries include Beech (Fagus grandifolia and Fagus sylvatica), Chestnut (Castanea dentate and Castanea mollissima) and Oak (Quercus alba and Quercus rubra). We kept the unigene from each species as individual dataset in further analysis. Dicot genes were served as the proxy sequence during sequence alignment and clustering of the unigene. TribeMCL clustering was used to identify the homologous sequence within and between genera. Two approaches were used to search for COS genes. The bottom-up method started from searching the conserved orthologous sequence among four sequenced dicot plant genomes. It resulted in 856 single copy orthologous genes in the dicot genomes. Each Fagaceae ESTs unigene dataset was searched against the dicot COS dataset to identify the putative COS genes among the Fagaceae dataset. In total, the number of unigenes have hits with each dicot COS ranging from 1 to 888. It represents that some of the dicot COS only exist in one specific Fagaceae species whereas some dicot COS have multiple orthologous genes in Fagaceae.

The top-down method combined all dicot genomes and Fagaceae unigenes during sequence comparison and clustering. The clustering generated 45,450 gene families and the number present in each gene family range from two genes per family to 368 genes in one family. We gradually break down the number of represented Fagaceae COS into following categories: 1) 1,753 COSs contain at least one genes in one of the dicot genome but has less than five genes in each dicot genome (5,384, 5,321 and 7,320 unigenes in Beech, Chestnut and Oak respectively); 2) based on the result of category 1, we evaluated the copy number of unigenes in each library and subtracted the gene family with standard deviation larger than two, 1,631 COSs were left (3,837, 4,836 and 6,504 unigenes in Beech, Chestnut and Oak respectively); 3) there are 2,356 COS presented in the Fagaceae species but are not presented in dicot genomes (4,651, 5,895 and 7,259 unigenes in Beech, Chestnut and Oak respectively). Conclusion: With different approaches and combining different dataset, our work identified the putative COS genes among Oak, Beech and Chestnut. These putative COS genes could be combined with other orthologous molecular markers to provide higher resolution in comparative mapping within the Fagaceae family.
P44: Glycolytic waves: elegant solution of the complex problem
A. I. Lavrova, T. Mair, E. Postnikov, L. Shimansky-Geier
Humboldt University, Magdeburg University, Germany Kursk University, Russia

It is known that glycolysis plays a central role for the energy metabolism of almost any living organism. The generation and propagation of glycolytic waves has been observed in living neutrophil cells [1]. The propagation dynamics may contain information about the state of the system; therefore it has been suggested that these waves may be of relevance for biological information processing, since the direction of wave propagation depends on both, the concentration of metabolites in the cell and the extracellular distribution of the biochemical compounds. Such waves have been also studied in yeast extracts. Using the open spatial reactor it was shown that waves often start to propagate from the border of the medium to its center [2]. A spontaneous change of the direction of wave propagation has been observed during one experiment in spite of the fact that the controlled influx was not changed [3]. In order to understand this non-trivial behavior we have performed a theoretical investigation on this phenomena and compared with experimental data. Also, it is well known that a lot of large-scale models have been developed to describe the spatio-temporal dynamics of glycolysis [4-6]. In contrast to the developed models we describe experimentally observed data using simple Selkov model which describes phosphofructokinase reaction, where the enzyme is inhibited by the substrate (ATP) and activated by the product (ADP). In order to describe observed phenomena we have used an inhomogeneous influx of the substrate. The goal of this study is not only to improve our understanding of the dynamics but to “catch” the biological sense of such behaviour of glycolytic waves for living systems.

P45: GRIND: Gene Regulation INference by Dual thresholding
W. P. A. Ligtenberg, D. Bosnacki and P. A. J. Hilbers
BioModeling & bioInformatics, Department of BioMedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

Whole genome expression analysis are performed to gain insight into diseases. These experiments measure the responses of all genes, to a given stimulus, at once. Previously, this was only possible for a selection of genes. With the appearance of compendia of these whole genome data sets, researchers are trying to elucidate the underlying gene regulation network from the data. For this purpose many algorithms have been developed. For example: Aracne[1] (Mutual Information), Banjo[2] (Bayesian Networks) and NIR[3] (Ordinary Differential Equations). Previously, we presented a benchmark data set to enable the impartial comparison of network inference algorithms. And we concluded that there was still room for improvement, both on the sensitivity of the algorithms as well as on the Positive Predictive Value (PPV) side. Here, we present our own algorithm (GRIND) to infer gene regulatory networks from knock down data. GRIND has been compared to Aracne, Banjo and NIR using the previously presented benchmark data set.
GRIND decides whether or not there is a connection between two genes based on two widely used features. The first feature is the P-value of a statistic test (In the benchmark we used a T-test for the computational data, however this could be improved by using more advanced test statistics.). The second feature is the correlation between two genes. Thresholds are applied to both features and only if both thresholds are met, a connection is assumed. The threshold values were optimized for one test data set (not belonging to the benchmark) and where then kept constant across the other data sets. The threshold values can be varied to focus more on either sensitivity or PPV. The result is a very fast algorithm, even the largest data set is processed within one day. Besides being faster, the results of GRIND are always better than the best result of the other algorithms.

In the future we will investigate other features that might improve results.

P46: Gene-specific dye bias correction of two-colour microarray data

Philip Lijnzaad, Thanasis Margaritis, Dik van Leenen, Diana Bouwmeester, Patrick Kemmeren, Sander van Hooff and Frank Holstege.
UMC Utrecht Netherlands Bioinformatics Center (NBIC)

We recently introduced the GASSCO method (Margaritis et al., Molecular Systems Biology 2009), for the correction of sequence-specific dye bias. This artefact occurs frequently in two-colour microarrays, but often remains unnoticed or untreated, even when serious. The artefact was hitherto not addressed, as it was not previously realized that it arises from two different factors, one that is specific to the hybridization, and one that is intrinsic to the probes themselves (LOESS normalization only corrects for intensity-dependent dye bias, not for gene specific dye bias). The practice of using dye-swapped slides, while advisable, may not eliminate the artefact, as the hybridization-specific factor need not be identical in both hybridizations of a dye-swapped pair. The GASSCO method, available from the Bioconductor package ‘dyebias’, detects and corrects gene-specific dye bias and does so in a general, robust and efficient way. The improvement can be dramatic; we frequently obtain a reduction in variance of M (that is, log2 of Cy5/Cy3) of over 50%. The method works for data from both expression- and ChIP-experiments, and both for oligo- and cDNA-arrays. It can also be used to perform post hoc corrections of publicly available data.
P47: Tree-based machine learning methods for the exploitation of biomedical images
Raphaël Marée, Pierre Geurts, Louis Wehenkel
Dept of EE&CS and GIGA, University of Liege, Belgium

Imaging techniques have been used for years in biology and medicine for research and clinical routine. With the recent advent of biosensors and digital image acquisition technologies (such as high-content screening, whole-slide virtual microscopy, etc.), scientists are now able to routinely generate terabytes of imaging data. Human interpretation of such large-scale datasets is very time-consuming and often biased. This stresses the need for replicable and generic computational methods to automate the extraction of useful, quantitative, information from these images.

In the first part of this presentation, we will give a brief overview of our recent work in the design of tree-based machine learning methods for the automated analysis of imaging data, including methods to index and search similar images in large and distributed databases, classify images into known phenotypes, and segment images into meaningful regions.

In the second part, we will review several of our ongoing projects at the GIGA centre using these methods to exploit large-scale imaging datasets, including automatic cell sorting and counting in chemotaxis assays, zebrafish phenotype classification for the assessment of developmental toxicity of compounds (in collaboration with Marc Muller’s lab, GIGA-R), and high-throughput detection of protein crystals in crystallization droplets for structural genomics and structure-based drug discovery (in collaboration with Kurt Hoffmann’s lab, RWTH, Fraunhofer IME).
P48: Predicting the primary origin of metastasized tumors by using a principal component analysis of >33,000 microarrays.
Marc Jan Bonder, Rudolf Fehrmann, Lude Franke, Gerard te Meerman, Cisca Wijmenga
Department of Genetics, University Medical Centre Groningen, The Netherlands Institute for Life Science and Technology, Hanze University Groningen, University for Applied Sciences, The Netherlands

Introduction
Expression profiles can differ substantially. It is known these are strongly dependent on the type of tissue and whether an individual is healthy or ill. On top of this environmental differences make a considerable difference: The expression profile in peripheral blood from a person who has just eaten will be very different from the same person who hasn’t eaten for a few hours. When comparing these expression profiles it is likely many genes in metabolism will show differential expression. By applying a principal component analysis (PCA) to over 33,000 human Affymetrix arrays, we identified 1,000 components that effectively describe different transcriptional activities. These thousand components can be used to discriminate between many features, as the individual components are orthogonal and much more robust than individual gene expression level measurements. To demonstrate this, we applied these components to discriminate between different non-disease tissue types. Subsequently, we hypothesized that through this approach we might be capable of predicting what the primary origins of metastasized tumors are.

Material and methods
44,000 Affymetrix samples were downloaded from the Gene Expression Omnibus. After quantile normalization and quality control we retained 33,000 samples. We applied PCA, and concentrated on the 1,000 PCAs with the highest eigenvalues.

A tissue prediction model was built using the component scores for a multiple microarray expression set (comprising 1,600 samples) with 65 different non-disease tissue. This prediction model was subsequently tested on 33 samples.

Results
Unsupervised clustering results in very good discrimination between tissue types. Samples from the same tissue cluster together and similar tissues cluster closer to each other than others. When comparing samples from completely different experiments the results are almost the same, indicating the robustness of this approach. We subsequently build a tissue prediction model using the 65 different major tissue types. We applied this model to the test set of 33 samples and observed that 87.8% of the tissues could be predicted correctly (compared to only 0.5% expected by chance). We subsequently used this model to assess a set of metastasized tumors to see whether we could correctly predict the primary tumor place of origin. For 34.1% of the samples we could correctly predict what the primary tumor was. For 14.1% of the samples we predicted that the tissue was identical to where the metastasized tumour was found. By chance only 1.5% of the samples would have been correctly classified as the primary tumor tissue or the tissue the metastasis was found in.

Conclusions
We conclude that the component scores can be used well to discriminate between different tissue types. For the prediction of the primary origin of metastasized tumors our model is promising. Although we can only correctly predict this origin for 34.1% of the metastasized tumors, one needs to bear in mind that our model was trained on healthy non-tumour tissues. We envision that by tailoring the training set to known primary cancer types, our prediction is likely to improve substantially within the near future.
Oomycetes of the genus Phytophthora are pathogens that infect a wide range of plants species. For dicot hosts, Phytophthora is even the most important pathogen. For example, Phytophthora sojae has caused major problems in the agricultural industry by infecting root and stem rot in soybean. Phytophthora infestans has been responsible for the Great Irish Famine in the 19th century and is still the most destructive pathogen of potato crops. Phytophthora ramorum has had devastating effects on the oak populations in California and Oregon, but recently also in Europe, by causing the disease Sudden Oak Death. Previous analyses of Phytophthora genomes unveiled many genes, large gene families and large genome sizes that can partially be explained by significant repeat expansion patterns. Analysis of the complete genomes of three different Phytophthora species, using a newly developed approach, unveiled hundreds of small duplicated blocks, mainly consisting of two or three consecutive genes. Further analysis of these duplicated genes and comparison with the known gene and genome duplication history of ten other eukaryotes including parasites, algae, plants, fungi, vertebrates and invertebrates, suggests that the ancestor of P. infestans, P. sojae and P. ramorum, most likely underwent a whole genome duplication (WGD). The fact that numerous small blocks of duplicated genes are found indicates that the genomes of Phytophthora species have been heavily rearranged following the WGD. The high repeat content in all three genomes may have played an important role in this rearrangement process. As a consequence, the paucity of retained larger duplicated blocks has greatly complicated previous attempts to detect statistically significant remnants of the WGD. However, our newly developed strategy to specifically detect small duplicated blocks might still unveil ancient large-scale polyploidy events. Moreover, we observed that the retained WGD-duplicates are biased to genes important for pathogenicity and thus the infection of the different hosts. This might help to explain the pathogenic success of the Phytophthora species.
P50: Reorganization of nuclear lamina - genome interactions upon differentiation of embryonic stem cells.
W. Meuleman, D. Peric-Hupkes, M. Reinders, L. Wessels, B. van Steensel
Netherlands Cancer Institute & Delft University of Technology

The three-dimensional organization of chromosomes within the nucleus and its dynamics during differentiation are largely unknown. We present a genome-wide analysis of the interactions between chromatin and the nuclear lamina during subsequent differentiation of mouse embryonic stem cells via lineage-committed neural precursor cells into terminally differentiated astrocytes. Chromatin in each of these cell types shows a similar organization into large lamina associated domains, which represent a transcriptionally repressive environment. During sequential differentiation steps, lamina interactions are progressively modified at hundreds of genomic locations. This remodeling involves both individual transcription units and multi-gene regions, and affects many genes that determine cellular identity. Often, genes that move away from the lamina are concomitantly activated; many others however remain inactive yet become unlocked for activation in a next differentiation step. These results suggest that lamina-genome interactions are widely involved in the control of gene expression programs during lineage commitment and terminal differentiation.
P51: Heavy metal resistance in Cupriavidus metallidurans: a complex evolutionary and transcriptional process.
Pieter Monsieurs, Hugo Moors, Rob Van Houdt, Paul Janssen, Max Mergeay, Natalie Leys
Belgian Nuclear Research Center SCK-CEN, Institute EHS, Microbiology Unit

The soil bacterium Cupriavidus metallidurans CH34 shows great promise in light of bioremediation technologies since it contains a significantly high number of genes involved in metal resistance. We designed a whole-genome microarray to investigate the full stress response of C. metallidurans CH34 at the transcriptomic level when challenged to a wide range of heavy metals such as zinc, copper, cadmium, and lead.

By performing a large number of microarray experiments, new metal resistance regions were identified. These loci showed similar expression profiles when cells were exposed to varying combinations of heavy metals thereby pointing to complex cross-talk at transcriptional level between the different heavy metal responses. In addition, while some genes were only switched on by one metal others were activated by a number of different metals.

The highly redundant nature of these metal resistant regions - illustrated by the large number of paralogous genes - combined with the with the phylogenetic distribution of these metal resistance regions within evolutionary related and metal resistant bacteria, sheds light on the recent evolution of this soil bacterium towards a highly metal-resistant species.

This intricate transcriptional behavior towards metal resistance in C. metallidurans represents an ideal test case for us to reconstruct the regulatory networks underlying this response. The complexity of this regulation is underpinned by the complete genome annotation of the C. metallidurans CH34 which allowed us to identify several hundreds of regulatory proteins, many of which are involved in metal detoxification and general metal resistance. Using a combination of microarray results and DNA and protein motifs, we are currently integrating these data to get a better view on this complex response.
P52: CoExpress: a tool for an effective co-expression analysis of large microarray data sets

Petr V. Nazarov*, Arnaud Muller*, Viktar Khutko†, Loïc Couder* and Laurent Vallar*

*Microarray Center, CRP-Sante, Luxembourg
†SUSS MicroTec Test Systems GmbH, Germany

Co-expression (CE) analysis of microarray data may provide interesting insights in understanding the gene and transcript level regulations in biological samples. It allows gene-networks reconstruction, disease pattern recognition, inferring of causal genes, etc. However, due to high computational costs and memory limitations, there is still a need in effective and user-friendly tools for the analysis of CE.

Here we propose a stand-alone software tool CoExpress for the interactive CE analysis of microarray data. The software is a user-friendly and allows on-the-fly study of CE, including (a) expression data preprocessing (b) building and visualization of CE matrix using correlation or mutual information metrics, (c) clustering, visualization and filtering of CE profiles, (d) visualization of co-expression networks for genes of interest. The possibility of the user-defined data processing using R-scripting is realized, providing a powerful tool for advanced users.

Due to its efficiency of memory usage and algorithm optimizations, a Windows version of CoExpress allows simultaneous analysis of CE for up-to 30000 genes or transcripts, measured on a hundred of arrays, in a reasonable time even on a standard PC. For a more time-consuming analysis, i.e. when working with thousands of experiments or/and using mutual information as a metric, a multi-thread command-line version has been developed that can be run on Linux multi-CPU systems.

Due to the specificity of the CE calculation, the growth of productivity is almost linear with the increase of number of CPUs.

The work of the software was tested using data from 2428 Affymetrix HGU133plus2 array experiments, downloaded from public repositories and preprocessed using R/Bioconductor. Data were normalized using RMA and then summarized, using gene symbols as indexes. The resulting data matrix, containing measurements for 19894 unique gene symbols, were analyzed using the multi-thread version of CoExpress. The analysis revealed that 2812 genes are co-expressed with at least one other gene with the absolute correlation higher than 0.8.

The validation of the resulting network was performed using STRING service at string.embl.de. The gene sets with the same co-expression profile were compared with a set of genes randomly selected genes and showed significantly higher level of connectivity.

The up-to-date version of CoExpress and its multi-thread module are freely available for downloading from www.bioinformatics.lu. The multi-thread module is distributed together with its source code under the GPL, which allows to modify, recompile and run it under various OS.
P53: Comparison of embedded feature selection algorithms for splice site detection
Huu Minh Nguyen, Yvan Saeys, Luc Knockaert and Tom Dhaene
UGent - IBBT - Intec

Recent technological advances have led to increasingly large amounts of data becoming available in many bioinformatics applications. The gathered data is not only characterized by a large number of instances but is also typically very high dimensional, containing many thousands of features. As a result analyzing this wealth of data is becoming very resource intensive, requiring significant amounts of processor time and memory capacity. One such application is the in silico detection of splice sites. In eukaryotic genes, splice sites form the boundaries between exons and introns, locating splice sites is therefore an important step in the detection of genes. For example (Degroeve - 2004) uses approximately 12,500 features to identify actual splice sites using a support vector machine (SVM) classifier. However many of these features may be redundant or noisy, using up resources without contributing to classification accuracy in the former case or being detrimental to the performance of the classifier in the latter case. Several feature (subset) selection algorithms have been described in the literature. These algorithms look for possible redundant or noisy features to improve classification performance, reduce the cost of building models and provide insight in processes behind the data. Feature selection algorithms are usually grouped into three categories: filter methods, wrapper methods and embedded methods. Filter methods select features on the basis of the intrinsic properties of the data. They are easy to apply, fast, scale well with dataset size and are model independent. However they do not account for interaction between features, unless explicitly taken into account. Wrapper methods utilizes a specific classification model as a black-box to determine whether or not a feature should be kept. As they are tailored for a specific model they tend to do better for that model type than filter methods and they are able to take into account possible interactions between features. They are however much more computationally expensive than filter methods and have higher risk of overfitting. The third category are the embedded methods which perform feature selection and model selection at the same time. Just like wrapper methods, they are tailored to a type classifier and can also take into account interactions between features, but are much less computationally expensive compared to wrapper methods. We focus on two SVM-based embedded methods, namely recursive feature selection (RFE) and AROM-l2 to improve the performance of the SVM splice site classifier in (Degroeve - 2004). These algorithms have shown to have good performance and are tailored to SVMs, utilizing the weights of a SVM to decide whether features should be kept or discarded. RFE discards a fixed percentage of lowest weighted features at each iteration, whereas AROM-l2 keeps all features until they have a weight equal to zero. We compare the resulting classification performances using the area under the curves of ROC (AuROC) and precision-recall curves (AuPRc). We also take into account the time complexity of each algorithm and their stability. Results will be analyzed and discussed.
P54: Combining a de novo assembly with mappings against multiple reference genomes
Jurgen Nijkamp, Wynand Winterbach, Marcel van den Broek, Jean-Marc Daran, Jack Pronk, Marcel Reinders and Dick de Ridder
Delft University of Technology

In the NGI Kluyver Centre for Genomics of Industrial Fermentation, several evolutionary engineering projects have been formulated. The goal is to enhance strains by selecting for a desired phenotype. It is now feasible to reverse-engineer the resulting yeast strains, by sequencing their genomes and looking for mutations with respect to the sequence of the initial strain. However, no genome sequence is available for the yeast strain that is currently used.

To obtain the DNA sequence of this reference strain, Next Generation Sequencing (NGS) is applied to obtain short-read data. Although read lengths are increasing, de novo assembly of a eukaryotic genome with NGS data is still a challenging task, and the typical result of such an assembly is a set with thousands of contigs. However, several sequenced genomes of close relatives of the target yeast are available, with varying qualities. By mapping the reads against the genomes of these relatives, sets of contigs present in both the target genome and the relatives can be obtained. We developed a graph-based algorithm making use of all available data, by optimally combining de novo and mapped contigs.

The algorithm takes as input multiple sets of contigs (one set per assembly source) and a finished genome of a close relative, which we call the reference genome. Every set has a manually assigned quality score attached to it. The contigs are pair-wise aligned in an all-vs-all fashion among the sets. An overlap graph is then constructed, using the pair-wise alignments that form ‘dove-tails’, i.e. contigs that overlap end-to-end. Nodes in this graph represent contigs, edges represent overlaps. These edges are weighted by assembly quality, contig length and an alignment score, combined using weighted Z-scores [1].

Using a three-step procedure, a chromosome is then assembled by finding an optimal path in the overlap graph. First, the start and end node of the path are determined by aligning the de novo contigs to the target chromosome of the reference genome. The contigs that align closest to the 5’ and 3’ ends of the reference genome will become the start and end node of the path. Second, to ensure that a path exists between the nodes, all contigs are aligned to the reference and edges with a low score (i.e. a penalty) are inserted between nodes that align close to each other. Finally, the optimal path is found by using a tabu search procedure for the highest-scoring path [2]. The assembled chromosome directly follows from the optimal path.

P55: Network analysis of differential expression for the identification of disease-causing genes
Daniela Nitsch, Leon-Charles Tranchevent, Yves Moreau
KU Leuven, ESAT-SCD

Genetic studies that aim to identify the causes of genetic disorders, such as linkage or association studies, identify chromosomal loci involved in a disease or a phenotype of interest. A frequent limitation is that these regions often contain tens or hundreds of genes, offering a serious limitation to further studies. Recently, methods for prioritizing candidate genes have been proposed, but they usually rely heavily on prior knowledge, which prevents them from tackling truly innovative discoveries. A standard practice in genetics consists of checking variations in the expression level of a candidate gene in patient-derived EBV transformed lymphoblastoid cells or fibroblast cells against wild-type cells. Using a gene network, our approach extends this strategy by assessing the level of differential expression in the neighborhood of a candidate gene under the hypothesis that a strong candidate has many differentially expressed neighbors, meaning that it belongs to a disrupted expression module. A network based on the STRING database (von Mering et al. 2007) was built, for which the distances between genes could be computed. Candidate genes located in highly differentially expressed neighborhoods are strong candidates. Different kernel matrices were used as distance measures, e.g. the Exponential Diffusion Kernel (Konder et al. 2002). To improve the enormous computing time, the kernel matrices were approximated by the Cholesky decomposition and the Reduced Eigenvalue decomposition. We score candidates by weighting and aggregating the differential expression of neighbors as a function of distance. Through a randomization procedure, we rank candidates by their p-values. We illustrate our approach on four diseases and successfully prioritize the known disease causing genes, with the key advantage of using unbiased expression data. We further apply the method to polycystic ovarian syndrome (PCOS) with no known disease genes, and suggest a new candidate gene.

P56: Inferring causal relationships using information-theoretic measures
Catharina Olsen, Patrick E. Meyer and Gianluca Bontempi
Université Libre de Bruxelles (ULB)

An important objective of causal discovery is to understand the data generating process underlying observations. This is of particular importance for instance in medical applications where it is usually desired to predict the effect of a drug or to understand the causes of a disease. Moreover, causal modeling provides the means to predict the effects of an intervention when the actual experiment is either too expensive or not feasible for ethical or technical reasons [Guyon et al. 2007]. We propose a novel technique for causal inference based on an information-theoretic measure called interaction information. This is a measure of dependence between three random variables and can be viewed as a multivariate extension of mutual information [Bell 2003].

In particular we will focus on the arc orientation of undirected networks inferred with state-of-the-art techniques. In directed acyclic graphs (DAGs) literature, approaches to arc orientation are generally based on conditional independence tests for inferring causal relations, e.g. the IC-[Pearl 2000] or the PC-algorithm [Spirtes et al. 2001]. Given datasets with a high ratio between the number of variables and the number of samples, these algorithms suffer from certain drawbacks. The two most important ones are: the algorithms’ complexity (testing for conditional independencies with growing, possibly large, conditioning sets) and the low power of the multiple testing procedure, due the small number of samples.

Given an undirected network, our approach computes the interaction information for all candidate triplets and infers colliders if the value is above a certain threshold ($X \rightarrow Y \leftarrow Z$ would be oriented into $X \rightarrow Y \leftarrow Z$ given that $X$ and $Z$ are not connected). Given that the networks under consideration are usually sparse, this approach is able to deal with high numbers of variables while at the same time avoiding high number of conditional independence tests. The remaining arcs are oriented using rules which aim at ensuring the acyclic nature of the resulting graph [Meek 1995].

The experimental validation will consist in assessing the performance of our approach against a number of state-of-the-art algorithms which are based on conditional independence tests. In a first step, the required undirected networks are inferred using well-known network inference algorithms like ARACNE [Margolin et al. 2006], CLR [Faith et al. 2007] or MRNET [Meyer et al. 2007]. The benchmark includes both synthetically generated datasets and real-world datasets (gene expressions (LOCANET) and protein-signaling networks (CYTO)) used for some recent competitions (Dream 2007 and NIPS 2008).
Most genome scale reconstructed cellular networks merely represent the sum of the function annotations (e.g. catalyzes reaction A, interacts with component X). As a result, these networks often poorly reflect the intricacies of cellular behavior. An obvious way to improve the network predictions, is to include the relation to regulatory elements, i.e. the environmental dependencies for the presence of gene products. Therefore, we formulated sequence-based strategies to identify the regulatory connections for any given transcription factor with high specificity [see also 1,2]. Basically, first a combination of sequence and context comparison, phylogeny, and analysis of published knowledge is used to link a DNA-motif and an inducing substrate (=input) to the selected transcription factor. And then, a genome-wide search for the motif is used to establish the corresponding regulon (=output). This reconstruction process can be facilitated by the integrated visualization of the sequence and function information. For that purpose we have considerably extended the capabilities of our web-based Microbial Genome Viewer (MGV). Most new functionalities in version 2.0 [3] relate to inter- as well as intra-species genome context comparisons for uploaded (ranked) lists, like transcriptome data, phylogenetic trees or putative regulons. Lists with selected genes or upstream regions can also be exported to be subjected to further analysis. Simultaneously, we have formulated a straightforward similarity motif-scoring procedure that provides ranked lists of putative binding sites for any given DNA motif. The added value of MGV 2.0 and the new scoring scheme is illustrated by the successful reconstruction of an important regulatory mechanism in Gram positive bacteria; transcription anti-termination by the BglG-family of regulators.

High-throughput screening of protein–protein interactions opens new systems biology perspectives for the comprehensive understanding of cell physiology in normal and pathological conditions. In this context, yeast two-hybrid system appears as a promising approach to efficiently reconstruct protein interaction networks at the proteome-wide scale. This protein interaction screening method generates a large amount of raw sequence data, i.e. the ISTs (Interaction Sequence Tags), which urgently need appropriate tools for their systematic and standardised analysis. We develop pISTil, a bioinformatics pipeline combined with a user-friendly web-interface: (i) to establish a standardised system to analyse and to annotate ISTs generated by two-hybrid technologies with high performance and flexibility and (ii) to provide high-quality protein–protein interaction datasets for systems-level approach. This pipeline has been validated on a large dataset comprising more than 11,000 ISTs. As a case study, a detailed analysis of ISTs obtained from yeast two-hybrid screens of Hepatitis C Virus proteins against human cDNA libraries is also provided. We have developed pISTil, an open source pipeline made of a collection of several applications governed by a Perl script. The pISTil pipeline is intended to laboratories, with IT-expertise in system administration, scripting and database management, willing to automatically process large amount of ISTs data for accurate reconstruction of protein interaction networks in a systems biology perspective. pISTil is publicly available for download at http://sourceforge.net/projects/pistil.
P59: Identification of Developmental Toxicants by their Interference with Cardiomyocyte Differentiation of Embryonic Stem Cells Studied by Transcriptomics

Dorien A. M. van Dartel, Jeroen L. A. Pennings, Frederik J. van Schooten, Aldert H. Piersma

RIVM, Bilthoven, The Netherlands and University of Maastricht, Maastricht, The Netherlands

The mouse embryonic stem cell test (EST) was designed to predict embryotoxicity based on the inhibition of the differentiation of embryonic stem cells (ESC) into beating cardiomyocytes. We hypothesize that the sensitivity of EST may be increased by the analysis of differentiation-related gene expression changes. Therefore, we studied the gene expression changes in ESC cultures during differentiation into cardiomyocytes and their modulation by monobutyl phthalate (MBP) or 6-aminonicotinamide (6-AN). ESC aggregates were formed and cultured in suspension from day 3 onwards. At day 5 the aggregates were seeded on tissue culture plastics, and differentiation was assessed by counting contracting cardiomyocyte foci at day 10. Aggregates were exposed from day 3 onwards to MBP, 6-AN or solvent control. RNA of non-exposed cultures was collected after 0, 24, 48, 72, and 96 hours of suspension culture, RNA of MBP- and 6-AN-exposed cultures were isolated after 24 and 96 hours of exposure. Samples were hybridized to 22k spotted mRNA microarrays, and 1355 differentially expressed genes were identified. Further functional analysis showed that these genes were mainly involved in either loss of pluripotency or gain of differentiation. Principal component analysis (PCA) showed that differentiation during the culture period could be well-described using the first and second principal axis of the PCA. Replicates within different experimental groups clustered together, and different groups appeared in chronological order on this axis, which can therefore be regarded as a representation of the differentiation track. Embryotoxicant-exposed cultures appeared to lag behind on the defined differentiation track or were driven off course. This resulted in significant differences in principal component coordinates between exposed samples and the time-matched controls on the differentiation track. These results show that the differentiation track of differentiating ESC can be described using a defined set of genes. Furthermore, modulation of cardiomyocyte differentiation by MBP and 6-AN could be demonstrated using this gene set.
Large parts of genomes are being transcribed atop of the transcription of coding sequences. A specific case of such transcription is that of sequences in the same location of coding sequences but on the opposite strand, a phenomenon called natural antisense transcription. So far in this matter very little attention has been given to prokaryotes. Among other things, prokaryotes discern themselves from eukaryotes in their compact genome structure. This results in a tight packing of protein-coding sequences, with the consequence that any non-coding sequence is bound to end up opposite of a coding sequence. Additionally, many of the coding sequences are organized in operon structures. In operons, multiple genes are transcribed in one single stretch and processed separately afterwards. We present our findings on genome-wide natural antisense transcription based on the transcriptional patterns of the Staphylococcus aureus mssa476 genome during a standard growth curve with a custom made semi-tiled, validated transcription array. Transcriptional units were defined by correlating the expression patterns of significantly expressed neighboring probes and including information of Genbank gene annotations. We focused on transcript pairs that (partially-) overlap and typified their behavior in terms of expression patterns and correlations towards each other. We found many more anti sense transcription pairs than generally found in eukaryotes. There was a marked difference in correlation between pairs with different overlapping orientation (5′3′ or embedded). With a clustering approach, we found that non-coding partners as a group differ in their overall expression pattern from coding partners of a pair. About 500 transcriptional units (mainly operons) were found to have multiple natural anti sense transcripts. These multiple anti sense transcripts did not necessarily comply in terms of correlation to the (operon-) sense. This may in some cases be a regulatory mechanism for translation of genes within an operon, a subject for further research.
P61: Efficient search in molecular graph space to recognize mass spectra

Jan Ramon
Department of Computer Science, K.U.Leuven

In several experimental procedures in bio-informatics, mass spectrography is an essential step. It is a challenging problem to extract as much information as possible from the resulting mass spectrogram. Ideally, one starts from a pure substance and desires to determine from the fragment peaks in the spectrogram the structure of the original molecule.

One possible strategy is to search in the space of all molecules for molecules that could have generated the observed spectrum. This is a non-trivial computational problem, as the space of all molecules is huge. In this work, we represent the molecules with graphs and we investigate the possibility to use recent and highly efficient graph pattern listing techniques to address this problem.

In order to decide whether a molecule could have generated the observed spectrum, one has to take into account two factors. First, there is the a priori likelihood of the molecule. E.g. when analyzing proteins, sequences that are equal or very similar to parts of the genome are more likely than unrelated sequences. Second, it should be taken into account how likely the observed spectrum is, given the proposed molecule. Usually, one hopes that the molecule produces a sufficiently diverse set of fragments so that the masses of many subgraphs of the graph to find are observed.

We will focus here on the listing part, i.e. the task to list all molecular graphs in a particular constrained space efficiently and without isomorphic duplicates. In the fields of data mining and graph theory, this problem has been studied in some depth. A rather generic result is given by Ramon and Nijssen, who show that one can list all elements of a monotone graph class with polynomial delay; i.e. such that the time needed between any two solutions is bounded by a polynomial in the input size. A graph class is monotone if it holds for all elements of the class that all subgraphs are also elements. This is typically a good property in the mass spectrography application where all subgraphs of a molecular graph are potentially observed fragments. It allows for focusing on the part of the search space with the highest likelihood according to a prior model of the fragmentation probabilities.

We investigate further possibilities to constrain or direct the search. One can show that it is possible to efficiently list molecules satisfying constraints such as a given total mass, a given total number of atoms of the different types, and a given set of possible degrees of nodes (valencies of atom types). However, listing problems that combine such constraints with the requirement that solutions should contain (or not contain) certain subgraphs (or subgraphs of certain weights) may not be solvable in output-polynomial time.

In ongoing work, we also consider the easier case of proteins, which have a linear structure (even though post-translational modifications can generate significant side chains). We'd also like to get a deeper understanding of the fragmentation probabilities as this would allow a better prioritizing in our search methods.

We’d like to thank Eduardo Costa, Hendrik Blockeel, Siegfried Nijssen, Tias Guns, Walter Luyten, Gerben Menschaert, Dries Van Dyck, Dirk Valkenborg, Thomas Gaertner and Tamas Horvath for their interest and suggestions. Jan Ramon is supported by the K.U.Leuven and the European Research Council.
A challenging and complex ecosystem of microbiota is found in the human gut, where the estimation of unique species exceeds 1000. Current developments in proteomics techniques have made high-throughput shotgun analysis of proteomes possible. In principle this even allows for the analysis of complex mixtures of proteins such as derived from a gut community of organisms, in an approach termed “metaproteomics”. However, the size and diversity of the complex potential proteomes that are revealed by sequencing such communities make the interpretation of peptide spectra by peptide spectrum matches (PSMs) inefficient due to the decrease of sensitivity as the search space grows. Most commonly, PSMs are evaluated by software (such as OMSSA) that fit theoretical spectra from peptides in a search space to query spectra. When using a translated metagenome as search space, the sensitivity rapidly decreases as the size of the translated metagenome increases. Using a “synthetic metagenome” constructed from a set of genome sequences from known gut inhabitants as a starting point, we propose a novel method that uses only a specific selection of real metagenomes as search space. In brief, hits with the translated synthetic metagenome allow retrieval of homologous sequences from metagenomes, which effectively reduces the search space. These subsets of the metagenomes can then be translated and used to retrieve more peptides. In later steps, this will increase the accuracy of identification of proteins since number of unique peptides is increased. Peptide spectrum matching using a synthetic metagenome already yields more unique peptide hits than a “real” metagenome, and therefore the proposed method is likely to further increase the number unique peptide hits.
P63: Querying the Nutritional Phenotype Database dbNP - A study querying tool for the new globally developed systems biology database for nutrigenomics

Jahn-Takeshi Saito*, Adem Bilican†, Chris Evelo†

*Department of Bioinformatics (BiGCaT) and Department of Knowledge Engineering (DKE), Maastricht University, The Netherlands
†Department of Bioinformatics (BiGCaT), Maastricht University, The Netherlands

The challenge of modern nutrition and health research is to identify food-based strategies promoting optimal health and well-being throughout the lifespan. This research is complex because it exploits a multitude of bioactive compounds and interactive processes. Whereas nutrition research can profit enormously from the technological revolution in high throughput molecular biology approaches (i.e. -omics), it has discipline-specific requirements for analytical and bioinformatics procedures. In addition to actual measurements, description of subjects and foods or diets is central to describing the nutritional phenotype.

The Nutritional Phenotype database (dbNP) aims at becoming a complete systems biology study database currently being implemented. When fully developed, dbNP will be a research and collaboration tool, and a publicly available data and knowledge repository. The implementation of dbNP will maximize benefits to the research community by enabling integration and interrogation of data from multiple studies, different research groups, different countries and -omics levels. The dbNP is designed to facilitate storage of pre-processed -omics data, as well as study descriptive and phenotype data and to enable the combination of this information at different levels (e.g. link phenotype, genotype, food intake, information on study design and -omics measurements).

The biological information stored in dbNP is integrated in a global modular database structure containing two kinds of modules. (1) Database modules contain raw and preprocessed data from genetics, transcriptomics, proteomics, biomarkers, metabolomics, functional assays, food intake and food composition. (2) Functional modules offer services within the dbNP workflow, process data, and interface with the user. The querying tool of dbNP is the main interface for nutrition researchers. The implementation of the query tool takes the modularity into account by bootstrapping from core modules in an initial phase and to expanding queries across further modules in a later phase of the project. The current version is integrated with the Generic Study Capturing Framework which is a functional module for accessing the study database module.

Study data is evaluated by storing, annotating and prepossessing of multiple layers of study data. Querying studies can take place at and across five levels. (1) meta data, (2) a single layer of data of one study, (3) a single layer of data of multiple studies, (4) multiple layers of data of a single study, (5) multiple layers of data of multiple studies. The querying tool adopts to the user’s workflow with different interfaces. Currently, three workflows are furnished. (1) In the comparison view, the user already operates on a study; the tool suggests a view in which the study’s data is already preselected for comparison. (2) In the meta-data view, studies can be queried by their meta-data. This supports the user in selecting broad overviews. (3) The complex view enables queries for all possible fields. The three views query in all five levels given above.

The current test version of the query tool is hosted on a server of the dbNP Consortium; it is implemented as an open source web application powered by Grails.
P64: Systematic investigation of the glucose signalling pathways in Saccharomyces cerevisiae through mRNA expression profiling
Katrin Sameith*, Eva Apweiler*, Marian Groot Koerkamp, Dik van Leenen, Patrick Kemmeren, Frank C.P. Holstege
Dept. of Physiological Chemistry, University Medical Center Utrecht, Universiteitsweg 100, 3584 CG Utrecht, The Netherlands
(* equal contribution)

DNA microarrays provide a platform to compare gene expression levels between particular cell populations on a genome-wide basis. The global differences in gene expression between single gene mutants and wild-types go beyond indicating which processes the perturbed genes are associated with. They themselves represent detailed molecular phenotypes. We exploit the power of expression profiling by systematically characterising and unwiring glucose responsive pathways in Saccharomyces cerevisiae. Physiological response to glucose levels involves a number of different systems such as the Ras/PKA, GPR1/GPA2 and SNF1 pathways. Using highly standardised, uniform conditions we generated gene expression profiles of over 100 deletion mutants of genes encoding all established and some putative components of these pathways. Surprisingly, the profiled components do not group according to the pathways they belong to. Instead, they can be associated to one of three groups: ‘positively’ or ‘negatively’ influencing processes occurring under high glucose conditions, or ‘inactive/redundant’ under these conditions. Through the construction of a transcriptional perturbation network we show that pathway components involved in the system of glucose signalling hardly regulate each other, but tightly regulate the expression of carbohydrate metabolic enzymes. Superimposing the perturbation network with the information on the association between different expression profiles suggests the presence of feedback in glucose response. We hypothesize that the transcription of enzymes, involved in the production of the reserve carbohydrates trehalose and glycogen, is adapted according to exogenous glucose levels, which in turn leads to fine-tuning the glucose signalling machinery. In our current efforts we aim to increase the level of detail in the transcriptional network by pinpointing transcription factors that regulate glucose responsive genes.
P65: Structure based GPCR pharmacophore construction using sequence information

M. Sanders*, J. Klomp†, S. Verhoeven†, C. de Graaf‡, J. de Vlieg*†

*Computational Drug Discovery Group, Radboud University, Nijmegen, Netherlands
†Department of Molecular Design and Informatics, Schering Plough Research Institute, Oss, Netherlands
‡Leiden/Amsterdam Center for Drug Research, Division of Medicinal Chemistry, Faculty of Science, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

G-protein coupled receptors (GPCRs) represent the largest superfamily of proteins in the human genome and are responsible for the transduction of an endogenous signal into an intracellular response in many different physiological pathways. Therefore they are effective targets to address in various diseases and of major interest of pharmaceutical companies. Like many other membrane proteins, GPCRs are hard to crystallize, which complicates the drug design process. A method was developed to automatically generate structure based pharmacophores using statistics from a large multiple sequence alignment of the 7 transmembrane domains of GPCRs. These pharmacophores can be used to predict poses of known active compounds and provide valuable information for compound optimization. Other applications in the near future are experimental design and pharmacophore guided compound library design.
P66: Non sparse kernel fusion and its applications in genomic data integration
Shi Yu, Tillmann Falck, Anneleen Daemen, Johan Suykens, Bart De Moor, Yves Moreau
Signals, Identification, System Theory and Automation Department of Electrical Engineering,
Katholieke Universiteit Leuven Heverlee, B3001, Belgium

This paper introduces the notion of optimizing different norms in the dual problem of support vector machine. The selection of norms leads to sparse or non-sparse solutions of multiple kernel learning, which may be employed for specific types of problems. In biomedical applications, non-sparse solution is more advantageous than sparse solution to thoroughly combine the redundant and complementary information in heterogeneous data sources. We discussed several L2-norm based convex optimization problems in a unified view. In particular, we propose the novel L2-norm kernel fusion in least squares support vector machine. The L2-norm classifier is compared with other algorithms in two real applications and large scale numerical experiments. It is shown that L2-norm LSSVM is an efficient and promising algorithm for genome-wide data fusion applications.
P67: Gene prioritization and clustering by multi-view text mining
Shi Yu, Leon-Charles Tranchevent, Bart De Moor, Yves Moreau
Signals, Identification, System Theory and Automation Department of Electrical Engineering,
Katholieke Universiteit Leuven Heverlee, B3001, Belgium

Background: Text mining has become a useful tool for biologists trying to understand the genetics of diseases. In particular, it can help identify the most interesting candidate genes for a disease for further experimental analysis. Many text mining approaches have been introduced, but the effect of disease-gene identification varies in different text mining models. Thus, the idea of incorporating more text mining models may be beneficial to obtain more refined and accurate knowledge. However, how to effectively combine these models still remains a challenging question in machine learning. In particular, it is a non-trivial issue to guarantee that the integrated model performs better than the best individual model.

Results: We present a multi-view approach to retrieve biomedical knowledge using different controlled vocabularies. These controlled vocabularies are constructed on the basis of nine well-known bio-ontologies and are applied to index the vast amounts of gene-based free-text information available in the MEDLINE repository. The text mining result specified by a vocabulary is considered as a view and the obtained multiple views are integrated by multi-source learning algorithms. We investigate the effect of integration in two fundamental computational disease gene identification tasks: gene prioritization and gene clustering. The performance of the proposed approach is systematically evaluated and compared on real benchmark data sets. In both tasks, the multi-view approach demonstrates significantly better performance than other comparing methods.

Conclusions: In practical research, the relevance of specific vocabulary pertaining to the task is usually unknown. In such case, multi-view text mining is a superior and promising strategy for text-based disease gene identification.
P68: Rationalizing the chemical space of protein-protein interaction inhibitors
Olivier Sperandio, Christelle H Reynès, Anne-Claude Camproux, Bruno O Villoutreix
Inserm UMR-S973, Université Paris Diderot

Protein-protein interactions (PPIs) represent one of the next major class of therapeutic targets although they are intricate to tackle with standard approaches mainly due to the inadequacy of today’s chemical libraries. However, the emergence of a growing number of experimentally validated PPI inhibitors (PPII) now allows drug designers to use chemoinformatics and machine-learning technologies to unravel the nature of the chemical space covered by the reported compounds. Key characteristics of PPII can then be revealed and highlight the importance of specific shapes and/or aromatic bonds thereby allowing the design of PPII’s enriched focused libraries and therefore of cost effective screening strategies.
P69: The Ectocarpus genome: evolution of a multicellular stramenopile

L. Sterck*, C. Martens*, E. Bonnet*, P. Rouzé*, Y. Van de Peer* and M. Cock†

*Department of Plant Systems Biology, Flanders Institute for Biotechnology (VIB), 9052 Gent, Belgium; Department of Molecular Genetics, Ghent University, 9052 Gent, Belgium.
†UPMC Univ Paris 6, UMR 7139 Végétaux marins et Biomolécules, Station Biologique, F 29682, Roscoff, France

Brown algae (Phaeophyceae) are complex photosynthetic organisms with a very different evolutionary history to green plants, to which they are only distantly related. These seaweeds are the dominant species in rocky coastal ecosystems, where they are responsible for more than 75% of the primary productivity, and they exhibit many interesting adaptations to these, often harsh, environments. The brown algae are also important because they are one of only a very small number of eukaryotic lineages that have evolved complex multicellularity. We report the 200 Mbp genome sequence of the filamentous seaweed Ectocarpus siliculosus (Dillwyn) Lyngbye, a model organism for the brown algae. Some major findings of this study include: 1) The unusual architecture of the genome which combines features found in small compact genome (many alternating genes, very few duplications) with features typical of large genomes (many introns - the genome contains a high percentage of intron sequence than any other genome we have looked at - and very long UTRs). 2) Probably the most important results are those related to the evolution of complex multicellularity in the brown algae. Several approaches have been used to identify genes that are linked with this process and the results of these analyses are very interesting. For example, we have discovered a family of transmembrane receptor kinases that have evolved independently of the receptor kinase families that play key roles in animal and plant development. Moreover, we have evidence that the targets of the Ectocarpus microRNAs include many genes that have evolved since the split from diatoms, suggesting that these microRNAs also may have played a key role in the evolution of complex multicellularity in this group.
P70: CHDWiki : An online collaborative and interactive data repository dedicated to congenital heart defects

S. Brohee*, R. Barriot*, J. Breckpot†, B. Thienpont†, S. Van Vooren*§, B. Coessens*§, L.-C. Tranchevent*, P. Van Loo*, M. Gewilling¶, K. Devriendt†, Y. Moreau*

*Bioinformatics, ESAT-SISTA, Katholieke Universiteit Leuven. Belgium
†Laboratoire de Microbiologie et génétique moléculaires. Université Paul Sabatier. Toulouse. France
‡Center for human genetics. Katholieke Universiteit Leuven. Belgium
§Cartagenia, Leuven, Belgium (http://www.cartagenia.com)
¶Department of Pediatric cardiology. University hospital Leuven. Belgium

How to efficiently integrate the daily practice of molecular biologists, geneticists and clinicians with the emerging computational strategies from systems biology is still much of an open question. Based on a collaborative Wiki, we developed a user friendly knowledge base and gene prioritization portal aimed at mapping genes involved in congenital heart defects (CHDs) and untangling their relations with corresponding human phenotypes. As classical wikis can only deal with text-formatted data from one source and do not support additional tools, we implemented an extension allowing to import data and tools from various online repositories. Our improved wiki system combines thus the text edition and user management facilities of the well known Wikipedia and the strictness of specialized biological databases. Moreover, this portal is not only an evolving community repository of current knowledge on the genetic basis of congenital heart defects but it is also a collaborative environment for the study of candidate genes potentially implicated in CHD - in particular by integrating recent strategies for the statistical prioritization of candidates genes. It serves the broad community that is facing CHDs, ranging from the pediatric cardiologist and clinical geneticist to the basic investigator of cardiogenesis. Of broad interest to the biological community, we argue that specialized collaborative knowledge and analysis portals will play a significant role in systems biology studies of numerous biological data and complexes.
P71: Detection of novel 3’ untranslated region extensions with 3’ expression microarrays
L. Thorrez, L. Tranchevent, H. J. Chang, Y. Moreau and F. Schuit
Center for computational systems biology, Katholieke Universiteit Leuven

The 3’ untranslated regions (UTRs) of transcripts are not well characterized for many genes and often extend beyond the annotated regions. Since Affymetrix 3’ expression arrays were designed based on expressed sequence tags, many probesets map to intergenic regions downstream of genes. We used expression information from these probesets to predict transcript extension beyond currently known boundaries. Based on our dataset encompassing expression in 22 different murine tissues, we identified 845 genes with predicted 3’UTR extensions. These extensions have a similar conservation as known 3’ UTRs, which is distinctly higher than intergenic regions. We verified 8 of the predictions by PCR and found all of the predicted regions to be expressed. The method can be extended to other 3’ expression microarray platforms as we demonstrate with human data. Additional confirming evidence was obtained from public paired end read data. We show that many genes have 3’ UTR regions extending beyond currently known gene regions and provide a method to identify such regions based on microarray expression data. Since 3’ UTR contain microRNA binding sites and other stability determining regions, identification of the full length 3’ UTR is important to elucidate posttranscriptional regulation.
P72: Using tiling microarrays to predict transcription start sites: application to Lactobacillus plantarum WCFS1

Tilman Todt, Michiel Wels, Roland J. Siezen, Sacha A. F. T. van Hijum
(1) Radboud University Nijmegen Medical Centre, Centre for Molecular and Biomolecular Informatics, the Netherlands
(2) Kluyver Centre for Genomics of Industrial Fermentation, the Netherlands

Prediction of transcription start sites (TSS) in prokaryotes allows to accurately determine upstream cis regulatory promoter sequences. Microarray tiling arrays offer a cost-effective alternative to transcriptome sequencing to determine approximate TSS locations for a large complement of genes in a given organism. Our goal is to predict genome-wide TSS by combining existing gene annotation information of L. plantarum WCFS1 with whole-genome tiling microarray expression data. In order to calibrate our algorithm, a training gene set was compiled using the following criteria: (1) gene shows expression signal in tiling array data, (2) predicted promoter location is upstream (up to 1 kb) of the gene, and (3) a predicted transcription start site is situated between predicted promoter location and gene. In this gene set of 620 genes, we found that 53% of the predicted transcription start sites are in close proximity (up to 30 nucleotides) to predicted promoter locations. For the other TSS, predicted promoter locations were further upstream or were not predicted yet. We are currently adapting our method to predict genome-wide TSS and to identify new promoter sequences based on these approximate TSS locations. Current results indicate that the resolution of tiling arrays limits determining exact TSS locations. Transcriptome sequencing would be a more suited means to this end.
P73: How to get the best from your array CGH data?
Léon-Charles Tranchevent*, Stein Aerts†§, Peter Van Loo*‡§, Bert Coessens*, Bassem A. Hassan†§ and Yves Moreau*

*Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, Leuven, Belgium
†Laboratory of Neurogenetics, Department of Molecular and Developmental Genetics, VIB, Leuven, Belgium
‡Human Genome Laboratory, Department of Molecular and Developmental Genetics, VIB, Leuven, Belgium
§Department of Human Genetics, Katholieke Universiteit Leuven School of Medicine, Leuven, Belgium

The development of advanced high-throughput technologies such as array CGH has led to many breakthroughs in human genetics by deciphering disease-loci associations. One typical feature of array CGH is that the loci found to be altered usually contain dozens of genes of which only a few are really causative for the disease / phenotype of interest. Endeavour, a web resource for the prioritization of genes, indicates which genes are the most promising candidates [1]. It relies on evidence that suggest that functionally related genes often cause similar phenotypes [2]. Endeavour uses various data sources that describe the function of the genes, their expression profiles, their genetic interactors, and their regulation processes. Our approach has been successfully validated by mean of a cross-validation on a set of 29 human genetic disorders. Furthermore, we have experimentally validated it with a detailed study on DiGeorge syndrome [1] and on neural development in Drosophila [3]. In conclusion, Endeavour is a gene prioritization resource that was extensively validated and is publicly available at http://www.esat.kuleuven.be/endeavour.

P74: Performance criteria and robustness in biological bistable models at the example of the apoptotic process.
Laura Trotta, Eric Bullinger and Rodolphe Sepulchre
Systems and Modelling, Université de Liège

Apoptosis is the programmed cell death used by multicellular organisms to remove damaged or potentially harmful cells and involved in cellular processes such as development of embryos, neurogenesis or regulation of the immune system (Hengartner, 2000). Defective apoptosis can lead to severe pathologies: while repressed apoptosis is found in most of cancers, enhanced apoptosis is present in developmental defects, autoimmune and neurodegenerative diseases.

Two ODE models of the extrinsic pathway of apoptosis (Eissing et al., 2004; Schliemann et al., 2007) have been proposed. These models present the apoptotic process as a system which can switch between two states, the survival and the death, in function of a pro-apoptotic input signal. Moreover, these models reproduce the main features of the extrinsic apoptotic pathway: the resistance of the survival state to external perturbations, the transition between the survival and the death for a sufficiently strong pro-apoptotic stimulus and the time-delay between the pro-apoptotic input signal and the effective activation of effector caspases.

The systems proposed by T. Eissing and M. Schliemann are bistable systems which means that they have two stable equilibrium points. Both models present, in addition, one saddle node. In order to understand switch mechanism in apoptosis, performance and robustness of bistable systems were investigated. Performance was defined as the capacity of the system to present a typical response to a signal input while robustness was related to its capacity to preserve this response in face of perturbations. It was shown that the main properties of the apoptotic switch can be expressed as performance specifications of particular bistable systems.

Although there have been several studies of the models of apoptosis, especially of the one proposed by T. Eissing et al., the saddle nodes have barely been analysed. However, this node has a special characteristic in the models of Eissing and Schliemann: it presents only one positive eigen value which is, in modulus, the smallest one of all the eigen values of the node. We called the bistable models having this characteristic “Retarding Switch Models” (RSM).

Sensitivity analyses and parametric perturbations were performed on two different RSM, the two-dimensional model of Griffith (Griffith, 1971) and the eight-dimensional model of apoptosis proposed by T. Eissing (Eissing et al., 2004). Our results suggest that the performance of RSM is influenced by properties of their saddle node equilibrium point.

In biology, bistable models have been used to describe physiological decision-making processes not only in apoptosis but also in cell cycle progression, development and differentiation (Tyson et al., 2008; Yan et al., 2009). Further investigation of these models could reveal if they present a similar structure and the part of RSM systems in biological switches.
P75: Training a text miner to summarize various bio-molecular events found in research articles
Sofie Van Landeghem∗†, Yvan Saeys∗†, Bernard De Baets†, Yves Van de Peer∗†

∗Department of Plant Systems Biology, VIB, 9052 Gent, Belgium
†Department of Molecular Genetics, Ghent University, 9052 Gent, Belgium
‡Department of Applied Mathematics, Biometrics and Process Control, Ghent University, 9000 Gent, Belgium

As new genetic studies are carried out every day, we are being overwhelmed with innovative and exciting results which pile up in biological databases and literature. Data mining efforts have become a necessity to find relevant information when interpreting new biological experiments, and text mining in particular is proving to be increasingly important as the literature base keeps growing exponentially.

Building upon previous work on extracting protein-protein interactions from text, we constructed a supervised learning framework that can detect and characterize various bio-molecular events from research articles using binary classifiers and parallel processing. Amongst other types, the resulting network displays events involving phosphorylation, binding, gene expression and localization. Whole events can play a role in new events expressing regulatory control, allowing for a nested structure of predictions. They can model a true interaction graph more accurately than ever before, offering new opportunities for practical applications such as hypothesis generation, database curation and knowledge discovery.

Our framework combines natural language processing techniques for the syntactic and semantic analysis of sentences, with a supervised learning component consisting of rich feature vectors and binary SVMs. Our methods achieve a competitive performance of 37.30 recall, 53.69 precision and 44.02 F1-score.

Finally, we have developed a rule-based system which can detect speculation and negation constructs from text, generating useful clues on the reliability of certain extracted facts and incorporating explicit negative statements into automatically created summaries. While preliminary results are encouraging, we plan on extending this work in the future.
P76: PLAZA: a comparative genomics resource to study gene and genome evolution in plants
Sebastian Proost+, Michiel Van Bel+, Lieven Sterck, Kenny Billiau, Thomas Van Parys, Yves Van de Peer and Klaas Vandepoele
1 Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Ghent, Belgium.
2 Department of Molecular Genetics, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium
(+ contributed equally)

The number of sequenced genomes of representatives within the green lineage is rapidly increasing. Consequently, comparative sequence analysis has significantly altered our view on the complexity of genome organization, gene function and regulatory pathways. To explore all this genome information, a centralized infrastructure is required where all data generated by different sequencing initiatives is integrated and combined with advanced methods for data mining. Here, we describe PLAZA, an online platform for plant comparative genomics (http://bioinformatics.psb.ugent.be/plaza/).

This resource integrates structural and functional annotation of published plant genomes together with a large set of interactive tools to study gene function and gene and genome evolution. Pre-computed data sets cover homologous gene families, multiple sequence alignments, phylogenetic trees, intra-species whole-genome dotplots and genomic colinearity between species. Through the integration of high confidence Gene Ontology annotations and tree-based orthology between related species, thousands of genes lacking any functional description were functionally annotated. Advanced query systems, as well as multiple interactive visualization tools, are available through a user-friendly and intuitive web interface. In addition, detailed documentation and tutorials introduce the different tools while the workbench provides an efficient means to analyze user-defined gene sets through PLAZA’s interface. In conclusion, PLAZA provides a comprehensible and up-to-date research environment to aid researchers in the exploration of genome information within the green plant lineage.
P77: Project HOPE: Providing the last piece of the puzzle...
Hanka Venselaar
CMBI, NCMLS, Radboud University Nijmegen

Recent developments in ultra-high throughput sequencing have led to a rapid increase in the detection of disease-related mutations. A significant part of these mutations affects the three-dimensional structure of the protein. Knowledge of protein conformations is necessary to link the newly found genotypes to an effect on structural level. Unfortunately, the gap between solved protein structures and known sequences continues to grow. Even though homology modeling can be used to predict the structure of significant a part of the sequences, majority remains without a known structure. To obtain more information about proteins we have to rely on sequence-based predictions and information deposited in databases. Existing online web-servers often present their conclusions in a manner that is hard to understand for a non-bioinformatician and do not focus on the actual structural effect of the mutation. Project HOPE is a web-server for automatic point-mutant analysis that can be used by anyone in the (bio)medical field. HOPE predicts the structural effect of a mutation in a way that is easy-to-understand for anyone in the (bio)medical field. The HOPE web server is available on http://www.cmbi.ru.nl/hope

Methods

The core of Project HOPE consists of a module that collects information from a series of Distributed Annotation Servers (DAS)-servers. This information includes predictions for transmembrane domains, secondary structure, accessibility and phosphorylation sites. The information is combined with other structural features known in the uniprot database and properties of the wild-type and mutant residue. A decision tree is used to reach an informed conclusion. The interface of Project HOPE is a website where the user can upload the sequence and mutation. Following analysis the server will provide a prediction of the structural effects of the mutation.

Results

Our approach was tested in numerous collaborative projects with researchers from (bio)medical departments. Our contribution consisted mainly of mutation studies (non-sense and missense mutations) and suggestions for experimental design. With our findings we provided the last piece of the molecular puzzle. Examples of these projects can be found at http://www.cmbi.ru.nl/~hvensela

Conclusion and Future prospective

The examples show that a structural analysis provides essential insight in the effect of a mutation. So far, HOPE’s predictions are sequence based, even when the sequence has a solved 3D-structure or possible modeling template. Mutant studies that use the structural coordinates are expected to improve our results. Therefore, we plan to extend Project HOPE with a module that can perform analysis on the solved structure or homology model. For this purpose we will develop a series of WHAT IF-web-services for structural calculations. Using structural information can be beneficial for research and, therefore, should be accessible and understandable for everyone in the biomedical field.
P78: Data integration in gene regulatory networks through composite network motif clustering
Vanessa Vermeirssen, Tom Michoel, Yves Van de Peer
Laboratory for Bioinformatics and Evolutionary Genomics, VIB Department of Plant Systems Biology, Ghent University, Belgium

Differential gene expression is a tightly controlled process that governs development, function and pathology of metazoan organisms. Several molecular interactions, e.g. protein-DNA interactions between transcription factors and target genes, protein-protein interactions between transcription factors, closely work together in order to establish proper gene expression in space and time. Biological networks have mainly focused on the relationships between one or two types of molecular interactions. In order to get a systems level understanding of how different molecular interactions interrelate to form a coordinated response in gene regulation, we studied composite network motifs in integrated networks containing protein-protein, transcription regulatory, protein-DNA, miRNA-mRNA, sequence homology and genetic interactions of the worm C. elegans. Through a computationally efficient and mathematically rigorous method, we identified dense clusters of several composite network motifs in this integrated C. elegans network. We discuss the biological function of these composite network motifs in the context of eukaryotic gene regulation. We conclude that composite network motif clustering is a useful data integration method to unravel the topological organization of gene regulation in metazoan organisms.
P79: Analyzing the Human Specificity of Pre-eclampsia
S. M. Vidanagamachchi, A. R. Weerasinghe, V. H. W. Dissanayake & R. Wijesinghe
University of Colombo School of Computing, Sri Lanka
University of Colombo School of Computing, Sri Lanka
Human Genetics Unit, Faculty of Medicine, University of Colombo, Sri Lanka
University of Colombo School of Computing, Sri Lanka

Pre-eclampsia is a life threatening, multi-system disorder, which is the de novo occurrence of both hypertension and proteinuria in pregnancy. It may affect both the mother and the fetus. It causes more than 50,000 maternal and fetal deaths annually all around the world. The main objective of this research is to analyze human specificity of pre-eclampsia.

Pre-eclampsia affects about 5% of Sri Lankan women during their pregnancy. It is estimated to develop most commonly during first-time pregnancies and it is seen only among humans. Sri Lanka is a country where a wide range of studies in improving maternal health have been carried out and continues to do so. According to a research carried out by a Sri Lankan research group, pre-eclampsia is reported to be one of the major pregnancy problems among Sri Lankan women. Their research shows that the increased risk of pre-eclampsia is associated with the variations of maternal Epidermal Growth Factor (EGF) gene. Additionally, they have found that variations of the EGF gene are also associated with the weight of the babies. Women with pre-eclampsia have a higher tendency to deliver low birth-weight babies. However, polymorphisms in the EGF gene associated with pre-eclampsia in Sri Lankan women are not the only reason associated with the low birth-weight. As the EGF gene is a member of a family of seven genes that act through a common EGF receptor, the interactions at the feto-maternal interface are very complex. These seven genes of the family include EGF, TGFA, AR, HB-EGF, Epigen, BTC and EPR.

The main goal that we intend to achieve is to compare the humans with other organisms in order to find out the reasons for human specificity of pre-eclampsia. The design process we followed is discussed next.

Hence we followed a procedure which sequentially utilizes several Bioinformatics techniques. Initially, gene and protein data of humans for the whole EGF family and the receptor from the GeneBank database were collected. Consequently, Basic Local Alignment Search Tool (BLAST) was applied to each of the seven genes and the receptor of the human EGF family to find organisms that are closely related to humans. Afterwards, organisms with the minimum E-value that reduces the possibility of random hits were selected. Next, protein sequences of selected organisms as well as of humans for each seven ligands and the receptor were collected. Afterwards, Multiple Sequence Alignments for all collected protein sequences using ClustalW tool were performed. Finally, if there were no clues present, a structural comparison using DaliLite tool was carried out.

We have selected four organisms that are closely related to humans: Chimpanzee, Monkey, Mouse and Boar to compare with humans. According to the results, we had to perform structural comparisons to find an acceptable clue for our objective. Therefore, we can conclude that structural difference which exists among humans and other organisms may be the reason behind the fact; pre-eclampsia appearing only in humans.
P80: Finding related pathways with text mining
Andra Waagmeester, Thomas Kelder and Chris Evelo
Maastricht University, Department Bioinformatics - BiGCaT, P.O.Box 616, NL-6200 MD Maastricht, the Netherlands

Pathways are abstract representations of interactions within the cell. Pathway knowledge captured in diagrams or models is instrumental in the analysis of large volumes of high-throughput studies, such as Microarray studies. This requires accurate and up-to-date pathway representations. Often a pathway is represented as an independent cascade of processes. This is usually because they are constructed around a specific context or research question. Connections with other pathways are either left out for clarity or because they are yet unknown to a pathway curator. Finding related pathways can be interesting for various reasons. One example is for example extending pathways with text mining. A typical text mining work requires corpora, which are collections of text. The results can be optimized if related pathways are clustered, since various pathway repositories such as WikiPathways and Kegg collect references relevant to the pathways under scrutiny. By clustering pathways, larger volumes of corpora can be gained. PubMed provides a method to capture relevant literature. We have applied this method available through the PubMed Eutils API, to construct methods to find relations between known pathways in either WikiPathways and KEGG. In our method Pathway A and Pathway B relates with each other if they share references suggested by the relevant literature suggestion function of PubMed. With our method we were able to identify various relations between pathways in wikipathways and kegg.
P81: Machine learning algorithms for coupling FAME and 16S rRNA data in bacterial species identification

Dries Van Audenhove, Willem Waegeman, Bram Slabbinck, Peter Dawyndt, Bernard De Baets
Ghent University

Fatty acid methyl ester (FAME) and 16S rRNA data are both used as descriptors for discriminating bacterial species. From a biological point of view, it makes sense to combine these two types of data to improve bacterial species identification, since FAME profiles and 16S rRNA are considered as biologically unrelated. However, by computing similarity measures between different bacterial species, one can observe that the similarity matrices obtained with both data sources contain certain dependencies.

In this work, we apply two state-of-the-art machine learning algorithms in order to further explore these dependencies. Firstly, we use learning approaches for biological network inference, in which the similarity measures obtained on 16S rRNA are treated as output variables (i.e. edges in a graph), while FAME profiles serve as input variables (i.e. nodes in a graph). Subsequently, we change the role of both types of data, so that the goal now consists of predicting FAME profiles from 16S similarities. In this second approach, a supervised classification algorithm that can handle partial class memberships is fitted to the data, in which 16S similarities are plugged into the method as a kernel matrix. The results of both approaches illustrate that statistical relationships exist between FAME and 16S rRNA. Using machine learning algorithms, we can determine in addition the peaks in the FAME profiles that cause these relationships.
P82: A functional-genomics fermentation platform to identify and optimize industrial-relevant properties of Lactobacillus plantarum

Michiel Wels∗†§, Peter Bron∗†, Anne Wiersma∗†§, Maria Marco∗†¶ and Michiel Kleerebezem∗†‡§

∗TIFN, †NIZO food research B.V., Kernhemseweg 2, 6718ZB Ede, the Netherlands ‡WUR, Wageningen, the Netherlands §KLuyver Centre for Industrial Fermentations, Netherlands ¶Present address UC Davis, Davis CA, USA

Probiotics constitute an important growth market for the food industry. However, further development of the probiotics market is presently constrained by a lack of knowledge on probiotic cellular components important for health-promoting effects. We have developed a functional-genomics fermentation platform approach for the identification and optimization of expression of specific probiotic functionality parameters. The platform employs Lactobacillus plantarum WCFS1 as an extensively studied model microbe for which advanced molecular tools are available. In the first set of fermentations, L. plantarum was grown in chemically defined medium according to a combinatorial fermentation scheme that included variations in medium composition and mild stress conditions (NaCl, pH, oxygen, amino acids and temperature). The molecular characteristics of the bacteria in samples harvested from these fermentations were investigated by transcriptome analysis, while in parallel their functiona l parameters were assessed using specific probiotic functiona lity assays, but also flavor profile changes that might influence product taste. Advanced bioinformatics analyses were performed to correlate transcriptomics results with the functionality parameters obtained, enabling the identification of genes involved in specific functions that are relevant for performance. The function of the identified genes and their relevance for performance will be further studied by genetic engineering (KO, overexpression). In addition, this approach will allow optimization strategies for the improvement of pre-selected target genes through specific modulation of fermentation conditions. Although developed with the use of a model organism, this approach is applicable to other bacteria, and the correlation of desired functional properties to “omics” datasets may assist identification of the underlying molecular mechanism, which opens avenues towards the design of fermentation strategies geared to improve the functional properties of bacterial (health-impact) cultures.
P83: Visualizing gene expression in transcription regulation and metabolic networks

M. A. Westenberg* and R. Bourqui†

*Eindhoven University of Technology, Eindhoven, The Netherlands
†University Bordeaux 1, Bordeaux, France

GENeVis was the first application for visualization of networks that supports overlaying time series data with associated statistical data on the nodes. This approach was extended to enable visualization of genome expression and network dynamics in both regulatory networks and metabolic pathways [1]. The main limitation of this approach is that metabolic pathways are considered as disjoint processes. Most metabolites (i.e. compounds or reactions), however, are shared among several pathways. Therefore, to investigate an organism’s metabolism, it is relevant to study all the pathways simultaneously. This requires visualization of the whole metabolic network, for which we have developed a specialized layout algorithm previously [2]. The MetaViz layout algorithm has two main steps: a clustering step and a rendering step. The clustering step computes a set of independent pathways, and it detects particular topological structures, such as cycles and cascades of reactions. Two pathways are considered independent if they do not share any reaction and/or compound. The clustering process can be constrained by the user, who can provide a list of focus pathways. The clustering algorithm then tries to respect a proximity constraint for these pathways, i.e., it will not split up the pathway across multiple clusters if possible. The rendering step draws the clustered graph and the clusters computed in the previous step, while respecting as much as possible the biological drawing conventions.

We now present an application that combines our previous work on GENeVis [1] and MetaViz [2]. In addition to visualization of data from time series experiments in the context of a regulatory network and a metabolic network, our application supports identification and visualization of affected subnetworks based on the expression data. The nodes of a subnetwork are highlighted in both the regulatory network and metabolic network. This approach can reveal quickly whether two independent regulatory subnetworks are connected at the metabolic level by a metabolite, implying a functional relation, or, conversely, whether two independent metabolic subnetworks are connected at the regulatory level by a shared regulator, implying a transcriptional relation. The technical details of the approach can be found in [3].

P84: Conservation of transcription factor binding sites in plants and its application to reduce false positive target genes

Elisabeth Wischnitzki, Yves Van de Peer, Klaas Vandepoele
UGent/VIB

Transcription factors regulate the transcription of genes by binding to certain specific binding sites in their promoter with just a few amino acids. Therefore these binding sites are generally small motifs composed of 6-12 nucleotides on average and they mostly include degenerated sites. On the contrary the presence of such a motif does not necessarily represent a functional binding site. Especially its small size and degenerated nature can lead to a high rate of false positive instances in a genome wide search. Therefore the individual instances have to be evaluated further to identify the most significant candidates. The growing number of sequenced genomes provides the unique opportunity to study the evolutionary conservation of motifs over different plant species. Our approach focuses on this evolutionary conservation as validation criterion. It is based on the assumption that a functional binding site is more likely to be conserved in promoter sequences of orthologous genes than a false positive motif instance. To estimate the evolutionary conservation the candidate motifs are evaluated per orthologous group using a background model built in an iterative process to obtain an empirical estimated p-value. Additionally the motifs are tested for a possible bias within the examined promoter regions (e.g. AT rich sequences). An iterative procedure of nucleotide shuffling and motif detection estimates how the nucleotide composition affects the motif detection results. The approach has been tested successfully in different settings and the results show that this is a powerful approach to detect evolutionary conserved binding sites.
P85: Fusion of sequence data and microarray data, a systematic approach toward cross-species comparison

Peyman Zarrineh*, Carolina Fierro†, Bart De Moor*, Kathleen Marchal†

*Department of Electrical engineering, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium

†Department of Microbial and Molecular systems, Katholieke Universiteit Leuven, Kasteelpark Arenberg 20, 3001 Leuven, Belgium

Cross-species comparison is essential to reveal the evolutionary relations, and to generalize the available information of well-studied organisms across other organisms for which laboratory data is limited. We developed a new methodology for cross-species comparison. This methodology combines microarray data and orthology data in an elaborated manner. We applied our methodology to compare networks in different bacterial species. Many conserved bioprocesses with different degree of conservation were found in distinct bacteria by the developed methodology.
Cladosporium fulvum is a biotrophic fungal pathogen that causes leaf mold on tomato. It is a member of the Mycosphaerellaceae family of the Dothideomycete class of fungi, one of the largest and most important classes of plant pathogenic fungi. During infection of its host, C. fulvum secretes a number of small proteins into the apoplast of infected leaves, which are collectively called effectors. These effector proteins play a vital role during pathogenesis by assisting the fungus in establishing disease. So far, ten effector proteins have been characterised from C. fulvum that in general show no or limited sequence similarity to other proteins present in public databases. In this study we try to identify additional effector proteins involved in the infection process. Currently, the genome sequences of at least six Dothideomycete species are available, including those of Mycosphaerella fijiensis, a devastating pathogen of banana, and Mycosphaerella graminicola, a pathogen of wheat that are both phylogenetically close related to C. fulvum. Recently, the genome of C. fulvum has also been sequenced. Here, using a combination of computational methods the secretomes for these three Dothideomycete species were determined and compared to each order in order to gain insight into the proteins that are specific for the infection process of C. fulvum. Protein-coding genes in C. fulvum were predicted using GeneMark-ES. The proteomes of M. fijiensis and M. graminicola were obtained from the Joint Genome Institute (JGI) genome portal.

We then defined a secreted protein as a protein containing a predicted signal peptide (SignalP), without putative transmembrane domains outside the secretion signal (TMHMM) or GPI-anchored regions (predGPI). All secreted proteins were functionally annotated by integrating BLAST similarities and predicted InterPro domains using Blast2GO. About half of the 1,274 predicted secreted proteins in C. fulvum have orthologues in the two other species, 289 have an ortholog in one of the two species, whereas 216 appear to be specific to C. fulvum, including chitin-binding proteins and hydrophobins. Furthermore, several protein families have been identified in the C. fulvum genome which are prevalent in the secretome, including serine-type endopeptidases, peroxidases, and fungal hydrophobins. In order to identify protein families an all-against-all similarity search was performed (BLASTP). Protein families were extracted with the MCL algorithm (inflation factor of 2.5) after a performance benchmark test with various inflation factors. Further research will focus on the evolution of protein families in C. fulvum compared to related species which will provide information on gene gain, gene loss, and protein family expansions in these species.
P87: Biodiversity, functional analysis and dynamics of a complex microbial consortium containing mainly Lactic Acid Bacteria.

V.C.L. de Jager, R.J. Siezen, M. Kleerebezem, S.A.F.T. van Hijum
NBIC, TIFN, WUR, CMBI

Lactic acid bacteria (LAB) form the major part of the complex microbial consortium (CMC) in an industrial cheese starter culture. Their natural diversity is an important reservoir of industrially relevant phenotypes. The challenge is to comprehensively describe the inherent strain diversity and functional dynamics of this complex microbial system. A workflow is presented to extract an accurate microbial taxonomy from random shotgun pyro-sequence data of the CMC meta-genome. The same data is used to provide insight in gene functions and their diversity present in the meta-genome. Profiling the CMC over time provides insight in the microbial processes involved in a model cheese manufacturing process.

Keywords: metagenomics, biodiversity, genotype-phenotype, industrial microbiota, functional analysis, lactic acid bacteria, life-science grid, snp/indel detection.
P88: Computerized analysis of experimental pictures on zebrafish
M. Menu†‡, J. Aceto†‡, P. Geurts∗‡, R. Marée∗‡, L. Wehenkel∗‡ and M. Muller†‡
∗ Department of EE and CS, † Molecular Biology and Genetic Engineering Unit, ‡ GIGA-Research, University of Liège, Belgium

The zebrafish is increasingly popular with biology laboratories for its ease of use and scientific advantages. We used this model to observe the effects of various substances, such as parathyroid hormone (PTH) or vitamin D3, on formation of cranial cartilage and bone. The purpose of this work was to create a computer program which would be able to analyze pictures of zebrafish where the head cartilage had been revealed by alcian blue staining. We wanted this program to differentiate a reference group from another one which had been given a specific substance and to identify the parts of the skeleton that changed. This objective was reached through a computerized quantization of zebrafish cartilage development and measurement of various length and angles between specific elements.

The idea applied here was to generate from the original color photograph a black-and-white picture where only the skeleton appears (in white). The method used is specific to zebrafish and is based on the fact that almost only the skeleton appears in blue on the original picture. Unfortunately, we had to face two problems. On the one hand, fish eyes mask part of the skeleton and on the other hand, some elements appear in blue although they are not part of the skeleton. A fairly complex technique allowed us to recover hidden parts and to eliminate superfluous ones.

With this new image, we were able to quantify the development of the skeleton and perform morphometric measurements. In the end, from the different measures obtained on the two groups of zebrafish (a control group and another one submitted to a tested molecule), our program makes a statistical analysis to identify significant differences.

The results of the statistical study are consistent with what we expected: our program concluded that vitamin D3 enhances fish skeleton development whereas it is the opposite with PTH. Moreover, we obtained these results with high rates of reliability. Besides, the program allowed us to observe some unexpected effects. Indeed, we noticed that vitamin D3 and PTH also affect the length of some parts of the skeleton. In conclusion, this computerized approach presents two big advantages compared to an analysis made by hand: A gain of objectivity and a lot of saved time!
P89: Characterization of spatio-temporal organization of slow waves during human NREM sleep
Jessica Schrouff, Yves Leclercq, Ariane Foret, Laura Mascetti, Pierre Maquet, Christophe Phillips
Cyclotron Research Centre, University of Liège, Belgium

Sleep is a behavior commonly observed in a large number of animal species. However, neuroscientists still poorly understand the meaning of this loss of consciousness absolutely needed for life. In the present work, we established different methods to characterize the Slow Wave Sleep most recognizable patterns: the Slow Waves (SWs). Since the anatomical structure of white matter tracts that connect various brain regions is not random and thus must constraint the propagation of waves (Hagmann et al., 2008), our basic hypothesis was that large white matter bundles would bias the propagation of SW along specific patterns, which could be identified in homogeneous clusters of waves. To investigate our hypothesis, SWs were detected automatically on the three first periods of SWS using an algorithm based on Massimini et al., 2004. They were then clustered using a two steps procedure involving a hierarchical clustering based on delay maps and a k-means clustering based on the SWs potential in a given time interval around the maximum power of the SW negative peak. To compute the relevance of the final clusters, a mathematical criterion was implemented as well as a visual check. Results of the multisubjects study showed that only bad quality and small clusters could be obtained, suggesting that there is no particular organization of SWs across the night and enforcing the hypothesis that SWs are local phenomena, each one decreasing the homeostatic pressure in only one specific area.
P90: Determinants of interaction specificity in the plant MADS transcription factor network
Aalt-Jan Van Dijk, Roeland van Ham, Richard Immink, Gerco Angenent
PRI, Wageningen University And Research Centre

We obtain sequence-level determinants of protein interaction specificity for the Arabidopsis MADS proteins, which are involved in a wide range of important developmental processes (e.g. floral organ formation). Our predictions were experimentally validated using site-specific mutagenesis and yeast-two-hybrid screening. Not only loss-of-interactions was observed, but also gain-of-interactions.
P91: Improving the prediction of protein-protein interactions by combining different biological sources

Herman van Haagen, Peter-Bram ’t Hoen, Antoine de Morree, Barend Mons, Martijn Schuemie
LUMC Leiden, Erasmus MC Rotterdam

The prediction of biological related entities such as protein-protein interactions (PPIs) is a common study in systems biology. The goal is to predict new pairs of proteins that have a high probability of interacting, or are related in another way such as being part of the same complex or pathway. We investigated if scores derived from individual data sources can be combined and that the combined score will increase precision and recall (e.g. reduces the number of false positives) compared to each individual data source. Our combined predictor includes six databases: textmining tools on PubMed abstracts, two co-expression databases namely COXPRESdb and Gene atlas, the Gene Ontology, ESTs from TiGER, and interacting domains taken from InterPro and DOMINE. Our combined predicted is benchmarked against the STRING database. Five combining approaches were evaluated: The product rule from STRING, classifier combiners calculated for each class (class of protein-protein interactions and the class of random protein pairs), maximum likelihood of ROC curves, combining p-values with Fisher’s method, and rank based combiniers. An AuC of 0.984 was obtained (compared to STRING AuC of 0.939) by using a weighted rank combiner where weights are taken from the AuC values of the individual data sources. We set a cutoff value of 50 false positives and retrieved 21,966 true positives for the weighted rank combiner compared to 15,341 for STRING. Recently we are investigating how those combiners perform in our protein prioritizer dubbed Nermal. We have two case studies: the DMD gene when defected causes Duchenne muscular dystrophy and the gene Calpain-3, when defected causes Limb girdle muscular dystrophy type 2.
P92: ROAST: robust and sensitive SNP/InDel detection in high-throughput sequence data

S. A. F. T. van Hijum∗†§, V. C. L. de Jager∗§, B. Renckens∗†, R. J. Siezen∗†§

∗Radboud University Medical Centre, Center for Molecular and Biomolecular Informatics, Nijmegen Center for Molecular Life Sciences, P.O. Box 9101, 6500 HB Nijmegen, the Netherlands
†NIZO food research, P.O. Box 20, 6710 BA Ede, the Netherlands
‡TI Food and Nutrition, P.O. Box 557, 6700 AN Wageningen, the Netherlands
§Netherlands Bioinformatics Centre, 260 NBIC, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands

Detecting single nucleotide polymorphisms (SNPs) and small insertions or deletions (InDels) in high throughput sequence data, such as Solexa (Illumina) and SOLiD (ABi), serves multiple purposes: (i) resequencing of a known organism, (ii) comparing the sequence of a natural derivative to its sequenced counterpart, and (iii) identification of functional mutations in organism derivatives compared to its parent. Current algorithms merely produce a list of SNPs/InDels for a given sequence and can only use above-mentioned short read sequence data. ROAST is a stand-alone tool that can handle both short read and pyrosequencing data (454, Roche) to determine SNP/InDels. We introduce a number of novel criteria that provide more complete information to judge whether a putative SNP/InDel is actually present or not. Our software runs on any Linux distribution with a minimum of requirements and very rapidly provides the user with a short-list of SNPs/InDels identified in a short-sequence or pyrosequencing data set compared to a reference sequence. Compared to existing tools, SNP/InDel detection by ROAST is very sensitive and specific.
P93: Deleteome - A systems-level approach towards understanding cellular processes
Sander van Hooff, Linda Bakker, Katrin Sameith, Philip Lijnzaad, Patrick Kemmeren, Frank Holstege
Holstege Lab, Department of Physiological Chemistry, UMC Utrecht

Regulation of gene expression is pivotal for most biological processes. Through an intricate interplay between the transcription machinery and signal transduction pathways, cells respond to external stimuli and internal processes. With the advent of genome-wide technologies it is now possible to study transcription regulation at a systems-level scale. Here, we are collecting a large number of gene deletion expression profiles to systematically characterize the role of transcription regulation and its interplay with other biological processes. Over 900 deletion mutants have been profiled so far, including proteins involved in signal transduction, ubiquitination, general transcription machinery and glucose sensing. To manage this large scale effort we have developed an integrated approach for obtaining and analyzing these expression profiles. This includes a completely automated microarray facility handling all steps from total RNA extraction to RNA amplification and labeling; a dedicated platform for the storage, analysis and sharing of microarray data as well as a large variety of software tools for downstream analysis and visualization. Through this integrated approach many interesting biological phenomena can be observed, such as signal transduction cascades, protein complex compositions, enzyme-substrate pairs and crosstalk between various biological processes.
P94: Network Centered Data Integration applied to Ageing Studies
E. B. van den Akker, B. T. Heijmans, M. Beekman, J. N. Kok, M. J. T. Reinders & P. E. Slagboom

* Molecular Epidemiology, Leiden University Medical Centre, Leiden, The Netherlands
Algorithms, Leiden Institute of Advanced Computer Science, Faculty of Mathematics and
Natural Sciences, University Leiden, Leiden, The Netherlands
Information and Communication Theory Group — Bioinformatics, Faculty of Electrical
Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The
Netherlands

Ageing is a complex process associated with a progressive loss of function which is thought to be
driven by several functionally unrelated pathways. Studies in model organisms indicate the pres-
ence of mechanisms capable of attenuating the biological rate of ageing and may provide protection
from multiple diseases. The Leiden Longevity Study (LLS) was designed to identify these so called
longevity assurance mechanisms in humans. This study enables the search for markers reflecting
variations in the progress of ageing between subjects of long-lived families and controls. Already,
candidate markers have emerged from this study design by the analysis of data on different levels
of molecular information (genome, transcriptome, metabolome and glycome). Separately, these
data-sets provide a limited description of the regulation of longevity assurance mechanisms. To
obtain a comprehensive description, an integrated cross-platform analysis combined with existing
knowledge from publically available sources is required. The goal of this project is to develop the
required bioinformatical and statistical framework. Our initial approach is that we try to identify
relevant gene-sets using gene expression data, while taking the network topology and SNP gene
interactions into account. We do this by performing a weighted gene set enrichment test in which
the weights of the genes are modulated according to a score summarizing a gene’s position in its
network and additional supportive experimental evidence. Attenuations in gene set rankings are
validated by employing literature mining algorithms.
P95: Logic gene network design: a CAD tool based on modularity and standardization

Bastiaan van den Berg
Delft UT, The Delft Bioinformatics Lab

Synthetic biology aims at building biological systems for useful purposes. Relatively simple gene networks have been engineered, but the design process is limited. Many papers advocate the use of engineering concepts like standardization and modular design to simplify the design process and enable the design of more complex systems. Currently, there are no tools available that implement both concepts in a practical way.

We have developed a software tool to show how standardization and modular design can be used for the design of logic gene networks, gene networks that implement a logic function. Using modular gene network templates, a user can simply connect logic gates to build a logic network. The software can translate a template into so called devices, using standard biological parts from an artificial bipartisan database. Next, the software can turn each device into a model, run a stochastic simulation, and evaluate the performance of a device based on the simulation result.

We have designed three logic gate templates and used them to build two logic gene networks: a demultiplexer and a D-latch. The software tool was used to turn the templates into devices and to evaluate the performance of the devices. The results show that the devices are evaluated correctly. Furthermore, the results show that for the design of a gene network our method can be used to indicate which biological parts are preferred at what location in the gene network.
P96: Mutator a fully automated method to scan and extract mutations from full text articles

Tom van den Bergh∗, Remko Kuipers†, Henk-Jan Joosten‡, Peter J. Schaap∗

∗Laboratory of Microbiology, Wageningen University, Wageningen, The Netherlands
†Centre of Molecular and Biomolecular Informatics, Radboud University, Nijmegen, The Netherlands
‡Bio-Prodict, Wageningen, The Netherlands

Genetic disorders are often caused by non-synonymous nucleotide changes in one or more genes associated with the disease. Specific amino acid changes, however, can lead to large variability of phenotypic expression. For many genetic disorders this results in an increasing amount of publications describing phenotype associated mutations in disorder-related genes. Keeping up with this stream of publications is essential for molecular diagnostics and translational research purposes but often impossible due to time constraints: there are simply too many articles to read. To help solve this problem, we have created Mutator a fully automated method to scan and extract mutations from full text articles. Extracted mutations are automatically cross-referenced to sequence data and a scoring method is applied to distinguish true mutations from false-positives. Fabry disease, a monogenetic gene disorder of the hGLA gene was used as a test case. Compared with existing Fabry mutation data-sets available at the Human genome Mutation Database (HGMD) and Swiss-Prot databases. Mutator extracted 30% additional mutations from the literature.
# List of Authors

<table>
<thead>
<tr>
<th>Authors</th>
<th>Page(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abeel, T.</td>
<td>25, 37</td>
</tr>
<tr>
<td>Abeln, S.</td>
<td>38</td>
</tr>
<tr>
<td>Aceto, J.</td>
<td>124</td>
</tr>
<tr>
<td>Adams, D. J.</td>
<td>21</td>
</tr>
<tr>
<td>Adriæns, M. E.</td>
<td>39, 58</td>
</tr>
<tr>
<td>Aerts, S.</td>
<td>109</td>
</tr>
<tr>
<td>Aguilar-Ruiz, J. S.</td>
<td>23</td>
</tr>
<tr>
<td>Akker, E. B. v. d.</td>
<td>130</td>
</tr>
<tr>
<td>Alard, P.</td>
<td>40</td>
</tr>
<tr>
<td>Alkema, W.</td>
<td>66</td>
</tr>
<tr>
<td>Amoutzias, G.</td>
<td>70</td>
</tr>
<tr>
<td>Angenent, G.</td>
<td>126</td>
</tr>
<tr>
<td>Apweiler*, E.</td>
<td>100</td>
</tr>
<tr>
<td>Apweiler, E.</td>
<td>76</td>
</tr>
<tr>
<td>Arts, P.</td>
<td>19</td>
</tr>
<tr>
<td>Audenhove, D. V.</td>
<td>117</td>
</tr>
<tr>
<td>Azuaje, F.</td>
<td>23</td>
</tr>
<tr>
<td>Babu, M. M.</td>
<td>26</td>
</tr>
<tr>
<td>Baets, B. D.</td>
<td>111, 117</td>
</tr>
<tr>
<td>Baeyens, L.</td>
<td>43</td>
</tr>
<tr>
<td>Bakker, L.</td>
<td>129</td>
</tr>
<tr>
<td>Bakker, L. V.</td>
<td>76</td>
</tr>
<tr>
<td>Balev, S.</td>
<td>67</td>
</tr>
<tr>
<td>Bareke, E.</td>
<td>44, 54</td>
</tr>
<tr>
<td>Barriot, R.</td>
<td>106</td>
</tr>
<tr>
<td>Baurain, D.</td>
<td>45, 65</td>
</tr>
<tr>
<td>Bayjanov, J.</td>
<td>46</td>
</tr>
<tr>
<td>Becker, J.</td>
<td>47</td>
</tr>
<tr>
<td>Beeckman, M.</td>
<td>130</td>
</tr>
<tr>
<td>Been, M. d.</td>
<td>93</td>
</tr>
<tr>
<td>Bel+, M. V.</td>
<td>112</td>
</tr>
<tr>
<td>Belshaw, N. J.</td>
<td>39</td>
</tr>
<tr>
<td>Bensonhuc, J. J.</td>
<td>76</td>
</tr>
<tr>
<td>Berg, B. v. d.</td>
<td>131</td>
</tr>
<tr>
<td>Berger, F.</td>
<td>54</td>
</tr>
<tr>
<td>Bergh, T. v. d.</td>
<td>132</td>
</tr>
<tr>
<td>Berkum, N. L. v.</td>
<td>76</td>
</tr>
<tr>
<td>Bilican, A.</td>
<td>99</td>
</tr>
<tr>
<td>Billiau, K.</td>
<td>112</td>
</tr>
<tr>
<td>Blockeel, H.</td>
<td>30, 62</td>
</tr>
<tr>
<td>Bodt, S. D.</td>
<td>53</td>
</tr>
<tr>
<td>Boel, E.</td>
<td>96</td>
</tr>
<tr>
<td>Bonder, M. J.</td>
<td>84</td>
</tr>
<tr>
<td>Bonnet, E.</td>
<td>105</td>
</tr>
<tr>
<td>Bontempi, G.</td>
<td>92</td>
</tr>
<tr>
<td>Bosnacki, D.</td>
<td>81</td>
</tr>
<tr>
<td>Bot, J.</td>
<td>21, 48</td>
</tr>
<tr>
<td>Botta, V.</td>
<td>49</td>
</tr>
<tr>
<td>Bourqui, R.</td>
<td>119</td>
</tr>
<tr>
<td>Bouwmeester, D.</td>
<td>82</td>
</tr>
<tr>
<td>Brabers, N.</td>
<td>76</td>
</tr>
<tr>
<td>Breckpot, J.</td>
<td>106</td>
</tr>
<tr>
<td>Breit, T. M.</td>
<td>96</td>
</tr>
<tr>
<td>Broek, M. v. d.</td>
<td>90</td>
</tr>
<tr>
<td>Brohée, S.</td>
<td>106</td>
</tr>
<tr>
<td>Bron, P.</td>
<td>118</td>
</tr>
<tr>
<td>Brouwer, R. W.</td>
<td>28</td>
</tr>
<tr>
<td>Bruning, O.</td>
<td>96</td>
</tr>
<tr>
<td>Brunner, H.</td>
<td>19</td>
</tr>
<tr>
<td>Buckley, M.</td>
<td>19</td>
</tr>
<tr>
<td>Bulcke, T. V. d.</td>
<td>72</td>
</tr>
<tr>
<td>Bullinger, E.</td>
<td>17, 110</td>
</tr>
<tr>
<td>Burg, H. v. d.</td>
<td>122</td>
</tr>
<tr>
<td>Cabane, C.</td>
<td>79</td>
</tr>
<tr>
<td>Calusinska, M.</td>
<td>50</td>
</tr>
<tr>
<td>Camproux, A.</td>
<td>104</td>
</tr>
<tr>
<td>Carvajal, D.</td>
<td>53</td>
</tr>
<tr>
<td>Cattaert, T.</td>
<td>14</td>
</tr>
<tr>
<td>Chang, H. J.</td>
<td>107</td>
</tr>
<tr>
<td>Charlier, C.</td>
<td>45</td>
</tr>
<tr>
<td>Charlouteaux, B.</td>
<td>71</td>
</tr>
<tr>
<td>Chassey, B. d.</td>
<td>24, 94</td>
</tr>
<tr>
<td>Cheng, J.</td>
<td>51</td>
</tr>
<tr>
<td>Claeya, M.</td>
<td>27</td>
</tr>
<tr>
<td>Cock, M.</td>
<td>105</td>
</tr>
<tr>
<td>Coessens, B.</td>
<td>106, 109</td>
</tr>
<tr>
<td>Conklin, B.</td>
<td>59</td>
</tr>
<tr>
<td>Coenen, M. L. J.</td>
<td>58</td>
</tr>
<tr>
<td>Coort, S. L.</td>
<td>52</td>
</tr>
<tr>
<td>Coort, S. L. M.</td>
<td>58</td>
</tr>
<tr>
<td>Coppieters, W.</td>
<td>45</td>
</tr>
<tr>
<td>Costa, E.</td>
<td>30</td>
</tr>
<tr>
<td>Couder, L.</td>
<td>88</td>
</tr>
<tr>
<td>Cruyce, J. V. d.</td>
<td>53</td>
</tr>
<tr>
<td>Daemen, A.</td>
<td>15, 18, 102</td>
</tr>
<tr>
<td>Daran, J.</td>
<td>90</td>
</tr>
<tr>
<td>Dartel, D. A. M. v.</td>
<td>95</td>
</tr>
<tr>
<td>Dateema, E.</td>
<td>122</td>
</tr>
<tr>
<td>Dawyndt, P.</td>
<td>117</td>
</tr>
<tr>
<td>Deelen, P.</td>
<td>56</td>
</tr>
<tr>
<td>Delft, J. v.</td>
<td>68</td>
</tr>
<tr>
<td>Demoulin, J. B.</td>
<td>16</td>
</tr>
<tr>
<td>Depiereux, E.</td>
<td>44, 54</td>
</tr>
<tr>
<td>Desmet, J.</td>
<td>40</td>
</tr>
<tr>
<td>Desmet, R.</td>
<td>60</td>
</tr>
<tr>
<td>Deutekom, H. W. M. v.</td>
<td>76</td>
</tr>
<tr>
<td>Devaux, Y.</td>
<td>23</td>
</tr>
<tr>
<td>Devriendt, K.</td>
<td>106</td>
</tr>
<tr>
<td>Dhaene, T.</td>
<td>89</td>
</tr>
<tr>
<td>Dijk, A. V.</td>
<td>126</td>
</tr>
<tr>
<td>Duarte, L.</td>
<td>57</td>
</tr>
</tbody>
</table>
Dupont, P., 25
E.Postnikov, 80
Eijssen, L. M. T., 58
Engelen, K., 60, 72
Ernst, D., 64
Essaghir, A., 16
Evelo, C., 42, 59, 68, 99, 116
Evelo, C. T., 39, 52
Evelo, C. T. A., 58
Falck, T., 102
Fawcett, J., 61
Fehrmann, R., 56, 84
Fey, D., 17
Fiedler, D., 20
Fierens, D., 62
Fierro, C., 60, 121
Fleuren, W., 63
Fluit, A. C., 96
Fonteneau, R., 64
Foret, A., 125
Fourré, F., 65
Francke, C., 93
Franke, L., 56, 84
Frenchy, P. J., 18
Frenkel, D., 38
Frigerio, J., 79
Frijters, R., 66
Fu, J., 56
Gaci, O., 67
Gaj, S., 58, 68
Galagan, J., 37
Garcia, R. A., 20
Gautier, C., 24
Georges, M., 45
Geurts, P., 26, 47, 49, 69, 73, 83, 124
Gevaert, O., 15, 18
Gewilling, M., 106
Giesbertz, P., 68
Gilsassen, C., 19
Graaf, C. d., 101
Gravendeel, L. A. M., 18

Haagen, H. v., 127
Ham, R. v., 122, 126
Hanspers, K., 59
Hassan, B. A., 109
He, Y., 70
Heijmans, B. T., 130
Helden, J. v., 16
Helleputte, T., 25
Hertog, B. D., 54
Heuze, F., 71
Hiard, S., 45
Hijum, S. A. F. T. v., 108, 128

Hilbers, P. A. J., 81
Hoen, P. ’., 127
Hoenkamp, E., 42
Hoischen, A., 19
Hollander, M. d., 122
Hollunder, J., 53
Holstege, F., 129
Holstege, F. C., 100
Holstege, F. C. P., 20, 76
Holstege., F., 82
Hooft, S. v., 82, 129
Houdt, R. V., 87
Hugenholz, J., 28
Hulsman, M., 48
Hurles, M., 31
Huynen, M., 57
Huynh-Thu, V., 73

Jager, V. C. L. d., 128
Jager, V. d., 123
Janssen, P., 87
Jennen, D., 68
John, J. M. M., 14
Jonker, M. J., 96
Jonkers, J., 21
Joosten, H., 132
Joosten, H. J., 78
Joshi, A., 41

Kallin, A., 16
Kelder, T., 52, 59, 68, 116
Kemmeren, P., 20, 76, 82, 100, 129
Ketany, M., 22
Khutko, V., 88
Kleerebezem, M., 118, 123
Kleinjans, J., 68
Klijn, C., 21
Klomp, J., 101
Knockaert, L., 89
Knoops, L., 16
Ko, C., 20
Koerkamp, M. G., 100
Kok, J. N., 130
Kolmeder, C., 98
Konings, P., 51
Kool, J., 75
Kormelink), T. G., 93
Kremer, A., 79
Krogan, N. J., 20, 76
<table>
<thead>
<tr>
<th>Authors</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kuipers, O. P.</td>
<td>28</td>
</tr>
<tr>
<td>Kuipers, R.</td>
<td>132</td>
</tr>
<tr>
<td>Kuipers, R. K. P.</td>
<td>78</td>
</tr>
<tr>
<td>Landeghem, S. V.</td>
<td>111</td>
</tr>
<tr>
<td>Lavrova, A. I.</td>
<td>80</td>
</tr>
<tr>
<td>Leclercq, Y.</td>
<td>125</td>
</tr>
<tr>
<td>Leenen, D. v.</td>
<td>76, 82, 100</td>
</tr>
<tr>
<td>Leeuw, W. C. d.</td>
<td>96</td>
</tr>
<tr>
<td>Lemmens, K.</td>
<td>55, 60, 72</td>
</tr>
<tr>
<td>Leys, N.</td>
<td>87</td>
</tr>
<tr>
<td>Ligtenberg, W. P. A.</td>
<td>81</td>
</tr>
<tr>
<td>Lijnzaad, P.</td>
<td>76, 82, 129</td>
</tr>
<tr>
<td>Lin, Y.</td>
<td>79</td>
</tr>
<tr>
<td>Loo, P. V.</td>
<td>106, 109</td>
</tr>
<tr>
<td>Lotteau, V.</td>
<td>24, 94</td>
</tr>
<tr>
<td>Maair, T.</td>
<td>80</td>
</tr>
<tr>
<td>Manevitz, L.</td>
<td>22</td>
</tr>
<tr>
<td>Maquet, P.</td>
<td>125</td>
</tr>
<tr>
<td>Maree, R.</td>
<td>83, 124</td>
</tr>
<tr>
<td>Marchal, K.</td>
<td>27, 55, 60, 72, 74, 121</td>
</tr>
<tr>
<td>Marco, M.</td>
<td>118</td>
</tr>
<tr>
<td>Margaritis, T.</td>
<td>82</td>
</tr>
<tr>
<td>Martens, C.</td>
<td>85, 105</td>
</tr>
<tr>
<td>Mascetti, L.</td>
<td>125</td>
</tr>
<tr>
<td>Mathers, J. C.</td>
<td>39</td>
</tr>
<tr>
<td>Meerman, G. t.</td>
<td>56, 84</td>
</tr>
<tr>
<td>Mensink, R. P.</td>
<td>58</td>
</tr>
<tr>
<td>Menu, M.</td>
<td>124</td>
</tr>
<tr>
<td>Mergeay, M.</td>
<td>87</td>
</tr>
<tr>
<td>Meulder, B. D.</td>
<td>44, 54</td>
</tr>
<tr>
<td>Meuleman, W.</td>
<td>86</td>
</tr>
<tr>
<td>Meyer, G.</td>
<td>69</td>
</tr>
<tr>
<td>Meyer, P. E.</td>
<td>92</td>
</tr>
<tr>
<td>Meyniel, L.</td>
<td>24, 94</td>
</tr>
<tr>
<td>Meysman, P.</td>
<td>60</td>
</tr>
<tr>
<td>Michel, T.</td>
<td>41, 114</td>
</tr>
<tr>
<td>Miles, A. J.</td>
<td>20</td>
</tr>
<tr>
<td>Mons, B.</td>
<td>127</td>
</tr>
<tr>
<td>Monseurs, P.</td>
<td>87</td>
</tr>
<tr>
<td>Moor, B. D.</td>
<td>15, 18, 27, 29, 72, 102, 103, 121</td>
</tr>
<tr>
<td>Moors, H.</td>
<td>87</td>
</tr>
<tr>
<td>Moreau, P. K. a. Y.</td>
<td>77</td>
</tr>
<tr>
<td>Moreau, Y.</td>
<td>51, 91, 102, 103, 106, 107, 109</td>
</tr>
<tr>
<td>Morree, A. d.</td>
<td>127</td>
</tr>
<tr>
<td>Movahedi, S.</td>
<td>53</td>
</tr>
<tr>
<td>Muller, A.</td>
<td>23, 88</td>
</tr>
<tr>
<td>Muller, M.</td>
<td>124</td>
</tr>
<tr>
<td>Murphy, S.</td>
<td>64</td>
</tr>
<tr>
<td>Navratil, V.</td>
<td>24, 94</td>
</tr>
<tr>
<td>Nazarov, P. V.</td>
<td>23, 88</td>
</tr>
<tr>
<td>Nepomuceno, I.</td>
<td>23</td>
</tr>
<tr>
<td>Nguyen, H. M.</td>
<td>89</td>
</tr>
<tr>
<td>Nijkamp, J.</td>
<td>90</td>
</tr>
<tr>
<td>Ning, Z.</td>
<td>31</td>
</tr>
<tr>
<td>Nitsch, D.</td>
<td>91</td>
</tr>
<tr>
<td>Olsen, C.</td>
<td>92</td>
</tr>
<tr>
<td>Overmaars, L.</td>
<td>93</td>
</tr>
<tr>
<td>Palmeira, L.</td>
<td>24</td>
</tr>
<tr>
<td>Parys, T. V.</td>
<td>37, 112</td>
</tr>
<tr>
<td>Peer, Y. V. d.</td>
<td>25, 37, 41, 61, 70, 79, 85, 105, 111, 112, 114, 120</td>
</tr>
<tr>
<td>Peckmans, K.</td>
<td>29</td>
</tr>
<tr>
<td>Pellet, J.</td>
<td>24, 94</td>
</tr>
<tr>
<td>Penel, S.</td>
<td>24</td>
</tr>
<tr>
<td>Pennings, J. L. A.</td>
<td>95</td>
</tr>
<tr>
<td>Peric-Hupkes, D.</td>
<td>86</td>
</tr>
<tr>
<td>Philips, T.</td>
<td>29</td>
</tr>
<tr>
<td>Phillips, C.</td>
<td>125</td>
</tr>
<tr>
<td>Pico, A.</td>
<td>59</td>
</tr>
<tr>
<td>Pierre, M.</td>
<td>44, 54</td>
</tr>
<tr>
<td>Piersma, A. H.</td>
<td>95</td>
</tr>
<tr>
<td>Plas, R. V. d.</td>
<td>29</td>
</tr>
<tr>
<td>Plomion, C.</td>
<td>79</td>
</tr>
<tr>
<td>Pronk, J.</td>
<td>90</td>
</tr>
<tr>
<td>Pronk, T. E.</td>
<td>96</td>
</tr>
<tr>
<td>Proost+, S.</td>
<td>112</td>
</tr>
<tr>
<td>Qiang, F.</td>
<td>60</td>
</tr>
<tr>
<td>Rabourdin-Combe, C.</td>
<td>24, 94</td>
</tr>
<tr>
<td>Ramon, J.</td>
<td>97</td>
</tr>
<tr>
<td>Reinders, M.</td>
<td>21, 48, 75, 86, 90</td>
</tr>
<tr>
<td>Renckens, B.</td>
<td>128</td>
</tr>
<tr>
<td>Reynès, C. H.</td>
<td>104</td>
</tr>
<tr>
<td>Ridder, D. d.</td>
<td>90</td>
</tr>
<tr>
<td>Ridder, J. d.</td>
<td>75</td>
</tr>
<tr>
<td>Romano, A.</td>
<td>39</td>
</tr>
<tr>
<td>Rooi, J. J. D.</td>
<td>18</td>
</tr>
<tr>
<td>Rooijers, K.</td>
<td>98</td>
</tr>
<tr>
<td>Rouzé, P.</td>
<td>79, 105</td>
</tr>
<tr>
<td>Saeyt, Y.</td>
<td>25, 43, 89, 111</td>
</tr>
<tr>
<td>Saito, J.</td>
<td>99</td>
</tr>
<tr>
<td>Sameith*, K.</td>
<td>100</td>
</tr>
<tr>
<td>Sameith, K.</td>
<td>129</td>
</tr>
<tr>
<td>Sanchez, A.</td>
<td>27</td>
</tr>
<tr>
<td>Sanders, M.</td>
<td>101</td>
</tr>
<tr>
<td>Sauter, T.</td>
<td>65</td>
</tr>
<tr>
<td>Schaap, P.</td>
<td>98</td>
</tr>
<tr>
<td>Schaap, P. J.</td>
<td>78, 132</td>
</tr>
<tr>
<td>Schaal, R. v.</td>
<td>66</td>
</tr>
<tr>
<td>Scheffer, H.</td>
<td>19</td>
</tr>
<tr>
<td>Schooten, F. J. v.</td>
<td>95</td>
</tr>
<tr>
<td>Schrouff, J.</td>
<td>125</td>
</tr>
<tr>
<td>Schrynenmackers, M.</td>
<td>26</td>
</tr>
<tr>
<td>Schuermie, M.</td>
<td>127</td>
</tr>
<tr>
<td>Schnit, F.</td>
<td>107</td>
</tr>
<tr>
<td>Sepulchre, R.</td>
<td>69, 110</td>
</tr>
</tbody>
</table>
Shimansky-Geier, L., 80
Showe, L. C., 22
Showe, M. K., 22
Siezen, R., 46, 123
Siezen, R. J., 28, 93, 108, 128
Signoretto, M., 15
Slabbinck, B., 117
Slagboom, M. J. T. R. a. P. E., 130
Smeets, R., 66
Smet, R. D., 55
Smits, N. J. P., 96
Spek, P. J. v. d., 18
Sperandio, O., 104
Steen, K. V., 14
Steensel, B. v., 86
Sterck, L., 105, 112
Stergiopoulos, I., 122
Storms, V., 27
Stubbs, A. P., 18
Sun, H., 55, 72
Suykens, J., 102
Suykens, J. A., 15
Tafforeau, L., 94
Taghipour, N., 62
Theunissen, D. H., 28
Thienpont, B., 106
Thijss, I., 55, 60, 74
Thorrez, L., 107
Toet, T., 108
Toffalini, F., 16
Tranchevent, L., 91, 103, 106, 107, 109
Trotta, L., 110
Uren, A., 75
Vallar, L., 23, 88
Vandepeeke, K., 112, 120
Vanneste, E., 51
Veltman, J., 19
Verschae, H., 30
Venselaar, H., 113
Verhoeven, S., 101
Vermeesch, J., 51
Vermeirssen, V., 114
Verstuyf, A., 105
Vidalain, P. O., 94
Vidanagamachchi, S. M., 115
Villoutreix, B. O., 104
Vlieg, J. d., 63, 66, 101
Vliet, W. A. v. d., 19
Voet, T., 51
Vooren, S. V., 106
Vriend, G., 78
Vugt, M. v., 66
Waagmeester, A., 42, 116