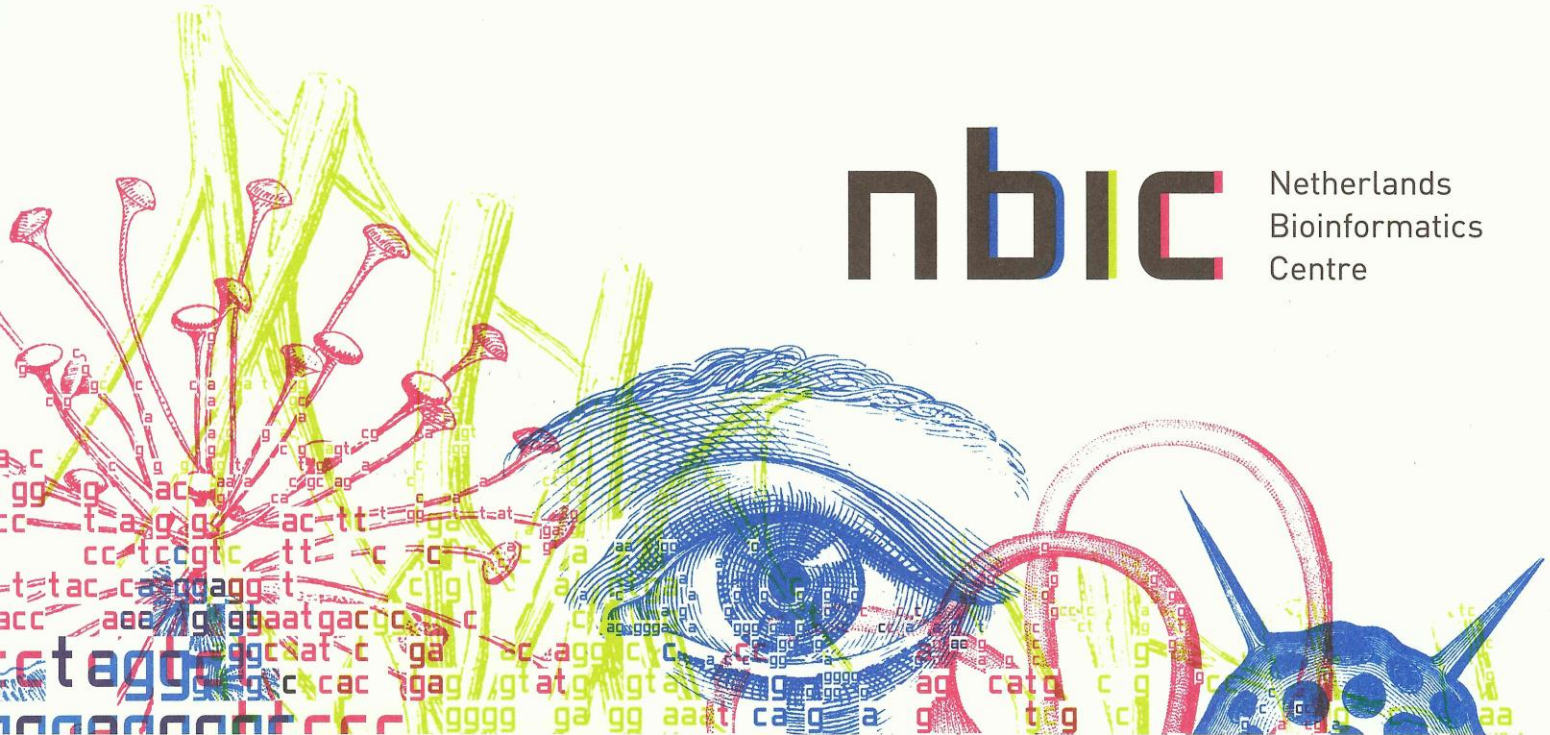
Netherlands Bioinformatics Conference (NBIC2012)

April 24 & 25

Abstracts Posters & Application showcase



nbic Netherlands Bioinformatics Centre

Index

Application showcase abstracts

List of participants:

1. Rick de Reuver, Radboud University Nijmegen Medical Centre (UMC St Radboud)
   VIPtool : a tool for prioritizing pathogenic variants in large annotated sets

2. Egon Willighagen, Maastricht University
   Bioclipse-OpenTox: interactive predictive toxicology

3. Juha Karjalainen, Department of Genetics, University Medical Center Groningen
   A gene co-regulation network based on 80,000 samples allows for accurate prediction of gene function

4. Kees Burger/Christine Chichester, Netherlands Bioinformatics Centre
   A new look for the ConceptWiki

5. Niek Bosch, e-Biogrid/SARA
   Life Science Grid Portal

6. Kostas Karasavvas, Netherlands Bioinformatics Centre
   Running Taverna workflows from the web

7. George Beyelas, UMC Groningen
   Just-enough workflow management system to run bioinformatics analyses on different computational back-ends, such as clusters and grids.

8. Mattias de Hollander, Netherlands Institute of Ecology (NIOO-KNAW)
   Ongoing efforts of a Galaxy solution for the BiGGrid/SARA HPC Cloud

9. Martijn Vermaat, Leiden University Medical Center
   Generating and checking complex variant descriptions with Mutalyzer 2

10. Marnix Medema, University of Groningen
    antiSMASH: genomic analysis of secondary metabolism

11. Pierre-Yves Chibon, Wageningen UR
    Marker2sequence: From QTLs to potential candidate genes.

12. Jos Lunenberg, Genalice B.V.
    Genalice: A new groundbreaking Systems Biology Data Processing and Correlation Analysis Platform

13. Farzad Fereidouni, Utrecht University
    Spectral phasor ImageJ Plugin

1. VIPtool : a tool for prioritizing pathogenic variants in large annotated sets

Rick de Reuver, Yannick Smits, Nienke Wieskamp, Marcel Nelen, Joris Veltman, Christian Gilissen

Department of Human Genetics, Nijmegen Centre for Molecular Life Sciences and Institute for Genetic and Metabolic Disorders, Radboud University Nijmegen Medical Centre, PO Box 9101, 6500 HB Nijmegen, The Netherlands.

Whole exome sequencing (WES) has proven to be a successful approach in identifying novel causative genes. As the technologies in sequencing progress, they become more accessible and broader applicable. The department of human genetics of the Radboud University Nijmegen Medical Center recently started using WES as a method in genetic diagnostics. WES results in large text files, holding over 40,000 annotated variants for a single individual. Interpretation of these large sets is challenging and the ability of prioritizing variant by non bioinformaticians is becoming more essential.
The Variant Interface Prioritizing tool (VIPtool) is a web-based shell that leaves the annotated variant data intact, but allows easy manipulation based on predefined filters and views. The input is a tabular separated file, containing annotated variants and general information about the dataset as a whole (e.g. quality control values). VIPtool only shows variants that meet the requirements stated by the filters or a predefined view. In a diagnostic setting a static disease gene filter allows to assess only those variants which occur in genes known to be involved in the patients indication. Users can inspect the meta information based on predefined criteria. Values that not meet the requirements are labeled to provide a visual indication of the data quality and data validity. The java applet itself has draggable columns with the option to (un)hide them.
VIPtool provides a safe, easy to use and uniform approach in prioritizing large sets of variants, making it ideal for both research and diagnostic purposes. The current implementation allows users to go from a variant list to the most likely pathogenic variant in two easy steps: 1, load input file, 2. select a view. This version is currently being tested by the different theme groups of our diagnostics department. The reports about the results and usability are encouraging.

---

2. Bioclipse-OpenTox: interactive predictive toxicology

Egon Willighagen, *Maastricht University*

Computational predictive toxicology draws knowledge from many independent sources, providing a rich support tool to assess a wide variety of toxicological properties. A key example would be for it to complement alternative testing methods. The integration of Bioclipse and OpenTox permits toxicity prediction based on the analysis of chemical structures, and visualization the substructure contributions to the toxicity prediction. This demo will focus on the application of Bioclipse in predictive toxicology, and show the interactive, visualization, and cloud computing features. There is an additional option for attendees to learn about how they can add their own predictive models.

Willighagen, E.L., Jeliazkova, N., Hardy, B., Grafström, R.C., Nov. 2011. Computational toxicology using the OpenTox application programming interface and bioclipse. BMC Research Notes 4 (487).

URL: www.biomedcentral.com/1756-0500/4/487

3. A gene co-regulation network based on 80,000 samples allows for accurate prediction of gene function

Juha Karjalainen, Rudolf Fehrmann, Gerard te Meerman, Harm-Jan Westra, Cisca Wijmenga, Lude Franke

Department of Genetics, University Medical Center Groningen and University of Groningen, Groningen, The Netherlands

High-throughput DNA microarray technology now provides us with a detailed view of the human transcriptome under different biological conditions. The increasing amount of publicly available microarray data helps in identifying and predicting genes that contribute to the same biological processes. To create a gene co-regulatory model of the human transcriptome, enabling the prediction of gene function, we analyzed 55,000 human, 17,000 mouse and 6,000 rat Affymetrix microarrays from the Gene Expression Omnibus. We created an integrated three-species gene network with 20,000 unique human genes and developed a principal component based statistical algorithm to predict functions for individual genes.
We benchmarked the algorithm against several pathway databases (including Gene Ontology, KEGG, BioCarta and Reactome) and observed that gene function could be generally predicted very well. For over 75% of all 20,000 genes we could predict at least one significant pathway association, function or protein localization. Furthermore, predictions could be made for more than half of the 5,000 genes that currently lack any known function. These results indicate that through the integration of many gene expression arrays biological knowledge can be obtained, even for those genes for which currently nothing is known.
To allow researchers and general public access to the results, we have created a website on which predictions and coregulation visualizations can be viewed based on genes of interest. An API for data access and further analysis features are in the works.

---

4. A new look for the ConceptWiki

Christine Chichester, Kees Burger, Hailiang Mei, David van Enckevort, Freek de Bruijn, Rob Hooft, Barend Mons

Netherlands Bioinformatics Centre

ConceptWiki is a collaborative knowledge resource for the life sciences, which is based on the general wiki idea of community annotation but employs specifically developed technology to enable distinct functionality. Inspired by Concept Web Alliance (CWA) grassroots movement and receiving support from the IMI Open PHACTS project, the rationale behind ConceptWiki is to provide a platform for the scientific community to collect and communicate basic semantic information about biomedical and chemical concepts. To support community annotation in an early stage, the ConceptWiki is preloaded with data from several authoritative sources on chemical compounds, proteins, organisms, diseases, and other biomedical concepts. Many of these concepts have been mapped between sources to remove redundancy and provide one uniformed view of a concept present in many sources.
The Open PHACTS Open Pharmacological Space (OPS) system is the first to fully exploit the potential of the updated ConceptWiki. Several APIs have been developed that provide a mapping between scientific textual representations of concepts and database and ontology identifiers.

These data are used to hook into the OPS system to retrieve query results. Using the ConceptWiki, it's possible, for instance, to map the term 'p53' to the specific human TP53 protein in several databases without the need to know the exact database numbers.
In the application showcase, we will demonstrate the new interface of the ConceptWiki to show the editing capabilities currently available.

---

5. Life Science Grid Portal

Niek Bosch

e-Biogrid/SARA

The Life Science Grid Portal enables its users to run several applications as parallel jobs. It contains several applications, such as BlastP, BlastN, BWA and Maq. Other applications can be added on request. The portal provides an easy interface to select parameters and run these applications. Multiple jobs can be submitted in parallel without the need to wait for previous jobs to finish. When many comparable jobs need to be done, this parallelization reduces the overall computation time.
It is also possible to store frequently used databases on the portal. These are then easily accessible as a parameter for corresponding applications. This way, databases can be used by your jobs without the need to transfer them to the portal every time an application is invoked. The use of Grid Storage provides enough storage space for big databases of several gigabytes.
The Life Science Grid portal provides a REST interface over the HTTP protocol, so all calls on the portal can be done using any HTTP client or client library, such as Curl, Curllib, wget or a regular browser. These are readily available and often installed on Linux systems by default. This provides an easy API, enabling programmers to easily submit jobs from the command line, scripts and other programs, without the need for any knowledge on the underlying (Grid) infrastructure.

---

6. **Executing Taverna Workflows from the web with no dependencies for the end user**

Konstantinos Karasavvas and Marco Roos

Netherlands Bioinformatics Centre

The Taverna Web Interface Form Generator is a web application that allows a user to run arbitrary Taverna workflows through a web page.

The bioinformatician finds the required Taverna workflow in the myExperiment repository and makes a note of that workflow's identifier. Appending that identifier to our web application's URL constucts the URL that can be shared with the end user. The latter visits the URL and is presented with that workflow's web form interface, where he can configure and execute it. Behind the scenes, the web application acquires the workflow details from myExperiment and constructs the appropriate user interface using templates. The user interface acts as an intermediary between the user and the Taverna server (using the REST API).

The generator takes advantage of all available information in myExperiment to create an easy to understand interface for the user. For example, the workflow's description is displayed together

with tooltips for inputs' descriptions as well as examples inputs. The users will have all the information they need, i.e. they need not be familiar with the workflow. Of course, that means that the generated interface is only as good as the available information that the workflow submitter provided. If information is missing in the repository it will also be missing from the interface. A link of the original workflow submission from myExperiment is provided if further information is required.

---

7. Using grid middleware as a computational resource for large-scale NGS and imputation in the eBioGrid project

George Byelas[1], Martijn Dijkstra[1], Alexandros Kanterakis[1], Pieter Neerincx[1], Freerk van Dijk[1], Tom Visser[2], Jan Bot[3,5], Irene Nooren[4,5] and Morris Swertz[1]

[1] *UMC Groningen*
[2] *SARA*
[3] *TU Delft*
[4] *University of Amsterdam*
[5] *e-BioGrid*

Specifying computation protocols for NGS and imputation analysis and running these at a large scale is a complicated task, due to many steps involved, long runtimes, heterogeneous clusters and large files. This process becomes error-prone when treating hundreds of samples, such as in genomic (diagnostic) facilities, if done without an integrated workflow framework and data management system.
From NGS and imputation projects we learnt that bioinformaticians don't want high-level workflow management systems with graphical user-interfaces, but prefer low-level shell scripts they can fully control instead. In addition researchers like to concentrate on the biology of their analyses and have automatic mechanisms taking care of complex job submission and monitoring details for grid computing.
Here we will demonstrate a 'just-enough' workflow declaration and execution system to address these needs. It is build on top of the MOLGENIS framework [1, 2, 3] for data tracking and uses templates to generate analysis scripts, This lightweight environment has been developed in collaboration with bioinformaticians working on NGS and imputation projects using grid or cluster middleware in particular in the nation-wide 'Genome of the Netherlands' project (700TB of data; >200.000 compute hours).

References:
[1] Byelas HV, Swertz MA, (2012), Introducing data provenance and error handling for NGS workflows within the MOLGENIS computational framework, in proceedings of the BIOSTEC BIOINFORMATICS-2012
[2] Swertz MA, Parkinson H, (2010), The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button, in BMC Bioinformatics. 2010 Dec 21:11 Suppl. 12:S12
[3] Swertz MA, Jansen RC, (2007) Beyond standardization: dynamic software infrastructures for systems genetics, in Nature Reviews Genetics 8

---

8. Ongoing efforts of a Galaxy solution for the BiGGrid/SARA HPC Cloud

Mattias de Hollander[1], Hailiang Mei[2], David van Enckevort[2] and Eiko Kuramae[1]

[1] *NIOO-KNAW: Netherlands Institute of Ecology*
[2] *NBIC: Netherlands Bioinformatics Centre*

The rapid evolution of next generation sequencing technologies (NGS) together with decreasing costs are creating a challenge to store and analyze the vast amount of sequencing data that are generated by experimental biologists. Configuring suitable data analysis software and having access to readily available computation and storage resources are the two major bottlenecks faced by many research groups. In this on-going collaboration, NIOO and NBIC BioAssist program are building Galaxy@HPCcloud solution to jointly tackle these two challenges.
This collaboration is motivated by two major developments in the Dutch Life Science community: the launch of High Performance Cloud at BigGrid/SARA and the increasing adoption of Galaxy as a common data analysis platform across different research groups. "Cloud computing" enables on demand access to the needed hardware resource for a certain period of time, while Galaxy bundles a set of advanced software programs dealing with various biological datasets and provides an easy and standard interface. This solution was demonstrated successfully by running NIOO Galaxy at the beta version of the HPCcloud using the Galaxy Cloudman system.
The goal of this joint project is two-fold. First, we will develop the Galaxy@HPCcloud solution together by expertise sharing and tackle the common problems like resource auto-scaling and persistent data storage. Second, we plan to make Galaxy@HPCcloud available to researchers in a dual way. One is to provide Galaxy@HPCcloud images to setup your own instances similar to the NIOO project in the beta cloud system. The other is to install the current NBIC Galaxy server (http://galaxy.nbic.nl) at the HPC cloud. This new NBIC Galaxy@HPCcloud will provide a more complete list of tools that can be used for Genomics and Proteomics data analysis and become a demonstration place for the tools and pipelines developed by all NBIC BioAssist developers.

---

9. Generating complex descriptions of sequence variants using HGVS nomenclature based on sequence comparison

J.F.J. Laros, M. Vermaat, J.T. den Dunnen, P.E.M. Taschner

Center for Human and Clinical Genetics, Leiden University Medical Center, The Netherlands

Descriptions of sequence variants can be checked and corrected with the Mutalyzer sequence variation nomenclature checker (https://mutalyzer.nl/) to prevent mistakes and uncertainties which might contribute to undesired errors in clinical diagnosis. Construction of variant descriptions accepted by Mutalyzer requires comparison of the reference sequence and the variant sequence and basic knowledge of the Human Genome Variation Society sequence variant nomenclature recommendations (http://www.hgvs.org/mutnomen/). With the advert of sophisticated variant callers (e.g., Pindel) and the rise of long read sequencers (e.g., PacBio), the chance of finding a complex variant increases and so does the need to describe these variants. An algorithm performing the sequence comparison would help users to describe complex variants.
The algorithm closely follows the human approach to describe a variant. It will first find the "area of change", and then finds the largest overlap between the original area and the area in the observed sequence. This process is repeated until the smallest description is found.

This algorithm ensures that the same description will be generated every time researchers observe this variant. Furthermore, no knowledge of the HGVS nomenclature is required to generate this description. This not only helps clinicians to generate the correct description, but its implementation also allows automation of the description process.
We have incorporated this algorithm in the Mutalyzer suite under the name Description Extractor (https://mutalyzer.nl/descriptionExtract).
Funded in part by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 200754 - the GEN2PHEN project.

---

10. Computational Genomics and Synthetic Biology Implementation of Microbial Secondary Metabolite Biosynthesis Pathways

M.H. Medema [1,2], P. Cimermancic[3], K. Blin[4], M.T. Alam[2], A. Trefzer[5], M. van den Berg[5], U. Müller[5], W. Heijne[5], R.A. Bovenberg [5,6], T. Weber[4], M.A. Fischbach[3], R. Breitling [2,7] & E. Takano [1]

[1] *Department of Microbial Physiology and*
[2] *Groningen Bioinformatics Centre, University of Groningen, Groningen, The Netherlands*
[3] *Department of Bioengineering and Therapeutic Sciences, UCSF, San Francisco, USA*
[4] *Mikrobiologie/Biotechnologie, Interfakultäres Institut für Mikrobiologie und Infektionsmedizin, Eberhard Karls Universität Tübingen, Tübingen, Germany*
[5] *DSM Biotechnology Center, DSM Food Specialties B.V., Delft, The Netherlands.*
[6] *Centre for Synthetic Biology, University of Groningen, Groningen, The Netherlands.*
[7] *Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom*

Microbial secondary metabolites are a major source of pharmaceuticals with antibiotic, cholesterol-lowering and antitumor activities. The genes encoding their biosynthetic pathways are usually clustered tightly together in microbial genomes. During the past decade, genome sequencing has revealed that microbial genomes may contain large numbers of such gene clusters, the products of which are often unknown and are not produced under typical laboratory conditions (e.g., [1, 2]).
In order to rapidly identify these gene clusters and prioritize them for experimental characterization, while keeping up with the ongoing acceleration of genome sequencing, we designed the computational pipeline antiSMASH [3]. It detects gene clusters using a curated library of hidden Markov models for signature genes specific to known gene cluster types, and predicts the chemical structure of encoded compounds based on several available substrate specificity predictors of the underlying enzyme modules. Moreover, antiSMASH compares the identified clusters to a comprehensive database of gene clusters from known nucleotide sequences, through the implementation of a simple algorithm that combines single BlastP queries and synteny information to identify homologous genomic regions.
In an effort to globally identify all prokaryotic gene clusters in the current databases, we combined antiSMASH with a more general algorithm that distinguishes biosynthetic gene cluster regions from other genomic regions based on their underlying Pfam domain frequencies. Analysis of the ±10,000 gene clusters identified shows that there are still vast numbers of unexplored types of biosynthetic pathways, and indicates that there has been large bias in the gene clusters selected for experimental characterization, both from a phylogenetic and from a biosynthetic point of view.
The birds-eye view on secondary metabolite biosynthesis offered by our computational approach now paves the way for new strategies to find novel drugs by a targeted exploitation of all available biochemical diversity using plug-and-play synthetic biology methodologies [4-6].

References
1. Medema et al. (2010) Genome Biol. Evol 2: 212-224.
2. Medema et al. (2011) Microb. Biotechnol. 4: 300-305.
3. Medema et al. (2011) Nucl. Acids. Res. 39: W339-W346.
4. Medema et al. (2011) Nature Rev. Microbiol. 9: 131-137.
5. Medema et al. (2011) Methods Enzymol. 497: 485-502.
6. Medema et al. (2012) Nature Rev. Microbiol. 10: 191-202.

---

11. Marker2sequence: From QTLs to potential candidate genes

Pierre-Yves Chibon, Richard Visser, Richard Finkers

Wageningen UR Plant Breeding

Quantitative Trait Loci (QTL) mapping is the process of determining the association between a trait and genetic marker data using statistical approaches. This technique is used by plant breeders on a regular basis within their research. However, QTL analysis by itself is not useful for determining which gene is the potential candidate explaining the observed phenotype. For this, a link between the genetic markers and the genome sequence should be established. Nowadays, this link is becoming available for more and more genomes due to the explosion in the use of Net Generation Sequencing (NGS) approaches. Within our research, we want to integrate QTL with the genome information and investigate the genomic region for potential candidate genes. For this we created the tool caller Marker2sequence. It is a web-based tool integrating the genome functional annotation with GO terms, proteins, pathway and literature information. Through this integration the related annotation for a gene can be presented and a short list of potential candidates genes can be generated for a given QTL.
This tool is thus helpful for both biologists and breeders to understand the underlying genomic regulation of an observed phenotype.
URL: https://www.eu-sol.wur.nl/marker2seq/

---

12. Genalice: A new groundbreaking Systems Biology Data Processing and Correlation Analysis Platform

Jos Lunenberg, Bert Reijmerink, Hans Karten

Genalice B.V.

Mission:
With the ambition to become the “Google DNA”, Genalice paves the way, with its vital enabling technology, to transform cancer from a terminal to a chronic and eventually preventable disease. Genalice’s primary goal is saving lives and improving the Quality of Life of patients suffering from cancer.

Vision:
Cancer originates from specific defects in DNA. In recent years, technologies have been developed to provide faster, cheaper and more detailed identification of such defects. Yet there is a gap to be

bridged between identification and patient benefits. A tsunami of complex data has to be analyzed and correlated now and in the future to translate this important new molecular information into valuable predictions on individual patient's disease development and treatment responses. The only way to get this massive job done and to get this vital information in a usable format to a treating physician is using an intelligent data-crunching engina as hidden in the hart of the Genalice solution. Major gains can be expected in different areas of the care cycle; earlier diagnostics and thus important time benefit, reduced insecurity, improved (personalized) treatment, and better tumor mutation/treatment success tracking. Better care and more precise/timely care cycles can only be achieved with state of the art technology.

Product:
Genalice provides a high performance DNA interpretation, analysis and correlation platform. Deployed via (private) cloud. It combines the robustness of a world-class database with the performance and flexibility of custom written software. Technology used for the platform is generally applicable; not only in the field of white, red and green biology, but in any field where small footprint of date, high-speed data interpretation/transformation and multiple domain pattern- or profile-correlation is required.

Technology:
Transformation of all types of diagnostic data into a uniform and small footprint format. Open, high speed interpretation engine to identify common patterns in large datasets. Pattern correlation engine (biomarker find). Weighted (fuzzy) pattern lookup (biomarker check).

---

13. Spectral phasor ImageJ Plugin

F.Fereidouni, G.A. Blab

Department of Molecular Biophysics, Debye Institute, Utrecht University

ImageJ is a set of tools for viewing and manipulation of images which can be extended by independent program components or "plugins". Recently we have developed a plugins which is designed to analyze (hyper-) spectral images by phasor approach. This provides a graphical representation of spectra and rapid (real time) semi-blind spectral unmixing of up to three components in the image.

URL: www.staff.science.uu.nl/~ferei101

Poster abstracts

1

Deep-sequencing of TALENs targeted embryonic stem cells to estimate their efficacy in genome editing

Seyed Yahya Anvar[1,2,*], Cor Breukel[1], Marcel H.A.M. Veltrop[1], Jaap W.F. van der Heijden[1], Jeroen F.J. Laros[1,2], Johan T. den Dunnen[1,2], Annemieke M. Aartsma-Rus[1], Sjef Verbeek[1]

[1] Center for Human and Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands.
[2] Leiden Genome Technology Center, Leiden University Medical Center, Leiden, the Netherlands.

Engineering of biological systems that recapitulate human genetic disorders relies on efficient manipulation of the genome. Recently, transcription activator-like effector nucleases (TALENs) have shown promising potential in site-specific genome editing. Their modular structure enables the design and simple construction of TALENs that can specifically recognize virtually any DNA sequence. Upon delivery into embryonic stem cells (ESCs), TALENs initiate a double strand break that is repaired by non-homologous end-joining, introducing a large variety of mutations. Since this method lacks a selection procedure the applicability depends largely on its efficacy.

Here we focused on altering the hDMD transgene, introduced into the genome of mouse ESCs. So far, editing by homologous recombination of hDMD transgene has failed. Thus, we engineered a TALENs-pair targeting intron 52 of the hDMD gene. The utility of the assembled TALENs-pair was determined by measuring the variation within the targeted sequence of the hDMD transgene among TALENs-transfected mouse ESCs. The 135bp locus of hDMD was PCR-amplified and sequenced for 100,000 TALENs transfected and non-transfected cells using IonTorrent semiconductor sequencing. The targeted locus was covered ≥450,000x. In TALENs-transfected ESCs, the rate of editing events, mainly small deletions and insertions, was 4-fold higher (~11%) than in non-transfected cells. Furthermore, we assembled a list of the most frequently occurred structural variations and cleavage sites to facilitate follow-up functional studies.

Our data endorse the use of TALENs for modifying the hDMD transgene in ESCs. The TALENs efficacy in genome editing can be further estimated using the Pacific Biosciences Single Molecule Real-Time sequencing.

2

PRISM (Protein Interactions by Structural Matching)

Attila Gürsoy, Özlem Keskin, Pelin Atıcı

Koc University

Protein-protein interactions occur when two or more proteins bind together to carry out their biological function. Prediction of protein-protein interactions at the structural level on the entire set of chromosomes expressed by a genome is a current and important problem since it allows prediction of protein function, helps investigation of pathways in the cell, modeling of protein complex structures and for gaining insights into various biochemical processes.

However, determining such a large scale problem is very hard from the experimental view and is very computationally costly from the computational view. There is a gap between the available data relating to cellular networks which suggest which proteins interact and this gap is widening. Computational methods can guess the protein-protein interactions at different levels. However, even if we do not consider the computational costs, if we have lack of biochemical information about the interaction sites, it is very difficult to predict the native interaction because there are many energetically-favorable ways for proteins to interact.

We used an alternative strategy which is to use structural similarity to an interface of a known protein complex. In the light of information that the protein pairs with different structures and functions can associate via similar interface architectures, using these interface structures can produce promising models for protein complexes even in the absence of global sequence or fold similarity. PRISM (Protein Interactions by Structural Matching) is a novel algorithm for finding interactions between the proteins based on this idea.

Here, we present a system PRISM (Protein Interactions by Structural Matching) which employs a novel prediction algorithm for protein-protein interactions. It predicts interactions and binding residues between target proteins by using structural and evolutionary similarity to known template interfaces. We define interfaces as interacting residues and nearby residues, respectively. The target proteins are the structures of protein chains in a target cellular pathway.

Our PRISM web server can be used to explore protein interfaces and predict protein-protein interactions by using this method. It constructs a database of pre-calculated protein-protein interaction predictions and every instance of similarity matchings can be searched for using the appropriate PRISM web page. Also, visualization of these template and target structures are generated using Jmol plug-in in our web server.

3

MicroWeb: making BD Pathway Data more Accessible

Coos Baakman

CMBI, UMC St. Radboud

The BD Pathway 855 System is a cell-imaging system that can automatically generate images and meta-data for large sets of samples like 96-well plates.

Working with the raw output of the microscope can be troublesome because it consists of many directories and unintelligible text files. We present MicroWeb, a web-application that is based on all modern concepts of eScience and that can present the data in a more user-friendly way.

MicroWeb translates 855-language to human language and thus can be seen as a very literal form of translational science. It has been adapted to the needs of the researchers.

4

Detecting recurrent gene mutations in pathway context using multi-scale graph diffusion

Sepideh Babaei [1,2], Marc Hulsman1, Marcel Reinders [1,2], Jeroen de Ridder [1,2]

[1] *Delft Bioinformatics Lab, Delft University of Technology, The Netherlands.*
[2] *Netherlands Bioinformatics Center*

Cancer is a complex disease caused by a myriad of changes in the DNA sequences that deregulate genes in their vicinity. It is well known that deregulation of different genes represent alternative routes for acquisition of the same cancer hallmark and, moreover, some cancer hallmarks require mutation of multiple genes. Determining the pathways that are deregulated in cancer, and using these to determine which genes are likely to play a role in the disease, is therefore of vital importance.

In this study, we aim to identify pathways of frequently mutated genes by exploiting their functional neighborhood encoded in the protein-protein interaction (PPI) network. To do this, we introduce a multi-scale kernel diffusion framework to discover recurrent gene mutations in the context of the PPI network neighborhood. We apply this novel methodology to a large collection of murine retroviral insertional mutagenesis data. The diffusion kernels method captures local topology of the interaction network and allows mutation scores to diffuse across the PPI network. The diffusion strength plays the role of scale parameter, determining the size of the network neighborhood that is taken into account. A permutation approach is employed to determine significance of a gene, within the context of the interaction network. As a result, recurrence of mutation is detected when the gene and/or genes in its neighborhood harbor frequent mutations. We demonstrate that statistically significant genes that are found with our approach are organized in connected components and are strongly enriched for cancer related pathways across the scales. Moreover, many genes that did not harbor sufficient mutations to be called significant individually are found to be significant in the context of their network neighborhood. A substantial portion of these are well-known cancer genes. Importantly, the putative cancer genes detected in this study were found to be significant at different diffusion scales, indicating the need of a multi-scale analysis. Taken together, these results demonstrate the importance of defining recurrent mutations while taking into account the pathway context.

5

Random Forest based data analysis of ~omics data: a novel method for detecting sub-classifications governed by interacting variables.

Lennart Backus, Jos Boekhorst, Sacha A. F. T. van Hijum

Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Centre

Within the field of biology, machine learning methods are increasingly used for classification problems. Random Forest (RF) allows the differentiation of predefined classification groups of samples based on variables. Such classifications are for example used in (i) genome wide association studies (GWAS) where (combinations of) SNPs (the variables) are linked to diseases or (ii) gene (variable) - trait matching. RF creates a model consisting of decision trees (usually 500+) that each have been trained with random subsets of samples and variables. A trained classification model can subsequently be used to classify new samples based on variables (e.g., SNP profiles). RF is very suited for the analysis of ~omics data as it is non-parametric and difficult to over fit. It effectively handles so-called large p, small n datasets, datasets with few samples (e.g., dozens) and many variables (e.g., thousands).

The RF method generally is used to generate variable importance scores, which give an estimation of the prediction impact for that variable. Based on the, often few, variables important for a given classification, their relation to the biological question can be determined in follow-up studies. With many ~omics datasets, however this approach does not allow extracting all relevant information from a dataset: the same phenotype can be caused by different combinations of variables (e.g., interacting SNPs). Typically, a researcher indicates the main classes: e.g., diseased vs. healthy. Possible subclasses of samples (e.g., disease caused by different SNPs) cannot be known beforehand and would ideally be uncovered from the data. To date, there are no methods that allow identifying subclasses of samples (e.g., subjects with the same disease but due to different interacting SNPs) and subsequently the interacting variables per subclass (e.g., interacting SNPs).

To detect these sub-classifications we developed a method to rank variable combinations based on the information gain when they co-occur in the same branch of trained classification trees. The minimum number of (interacting) variables that still allow optimal classification is used to cluster samples per classification into sub-classifications. The (interacting) variables important for the sub-classifications (groups of samples with the same phenotype) are determined and can be used for further establishing biomarkers for each group as well as further biological interpretation. This work was based on an artificial dataset that simulates GWAS data. As we our methodology uses only output from the random forest, it can be broadly applied to any ~omics dataset suitable for classification, such as GWAS, proteomics, metabolomics, microbial community data, as well as other datasets such as clinical data.

6

Structural homology in Solanaceae: analysis of genomic regions in support of synteny studies in tomato and potato

Joachim W. Bargsten [1,4,5], Dóra Szinay[2], José van de Belt[2], Richard G.F. Visser [3,4], Yuling Bai[4], Hans de Jong[2], Sander A. Peters[1]

[1] Plant Research International, Business Unit of Bioscience, cluster Applied Bioinformatics, Plant Research International, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands.
[2] Laboratory of Genetics, Wageningen University and Research Centre, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands.
[3] Wageningen University and Research Centre, Department of Plant Sciences, 6700 AJ Wageningen, The Netherlands,
[4] Laboratory of Plant Breeding, Wageningen University and Research Centre, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands,
[5] Netherlands Bioinformatics Centre (NBIC), Nijmegen, The Netherlands.

Crop plants such as potato and tomato suffer severe inbreeding depression. Susceptibility to pathogens and sensitiveness to abiotic stress are common consequences. One way to overcome these shortcomings is the introduction of favourable traits from wild varieties. To introduce alien chromosomal sequence from wild varieties into a crop plant, identification of structural rearrangements can facilitate the breeding process. We investigated the structural rearrangements on the long arm of chromosome 2 between the closely related crop plants potato and tomato. The selected chromosomal segment contains several genes related to abiotic stress. Molecular organization and collinearity junctions were delineated by using multi-color BAC FISH analysis and comparative sequence alignment. We found large-scale rearrangements including inversions and segmental translocations, which were not reported in previous comparative studies. Comparative linkage maps suggest these rearrangements are tomato lineage specific. On a microsyntenic level, gene segments are mostly conserved. However, the gene collinearity (linkage) is often interrupted by small scale rearrangements. These results suggest frequent recombination events during the evolution of potato and tomato as well as high abundance of small scale rearrangements, i.e. transpositions mediated by mobile elements in the genome. The unravelled rearrangements cannot be identified solely on known genetic markers. Thus, not only a high resolution of genetic markers, but also experimentally supported comparative genomics is essential for successful inter-species breeding.

This work is supported by a TTI Green Genetics grant 2CC037RP and financial aid from Rijk Zwaan, Syngenta AG and Monsanto.

7

How do you recognise your distant relatives? Application of profile HMM for aligning distant proteins

Punto Bawono, Sanne Abeln, and Jaap Heringa

The Centre of Integrative Bioinformatics (IBIVU), Vrije Universiteit Amsterdam

Performing multiple sequence alignment (MSA) on distantly related protein sequences (without 3D structures) is still a very challenging task in bioinformatics. One of the major difficulties here is to correctly align the functional/structural motifs. In this work, a new technique is introduced to improve the sensitivity of MSA of distant protein sequences. This method employs profile HMM as representation of the alignment instead of the standard (Gribskov) profile. Since the the quality of the seed alignment is of paramount importance for a profile HMM, a structural alignment of input sequences, with an available 3D structure or one obtained via homology search, will be used as the seed alignment for the profile HMM.

The evolutionary diverged Flavodoxin and Cupredoxin families are used as test cases for this method. A relatively distant set of flavodoxin FMN-binding redox proteins is used as the first test case, to this set we added a signal transduction che-Y protein which is known to adopt the same α/β flavodoxin fold but with very low sequence identity compared to the other sequences in this set. The second test set: Cupredoxin family consists of sub families that are notoriously difficult to align. Correct gap treatment is essential factor in aligning the cupredoxin set.

8

Grid and cloud computing for high throughput assembly and annotation of (meta)genome sequences

Jumamurat R. Bayjanov [*,§], Victor de Jager [*,#,§], Sacha A.F.T. van Hijum [*,§,†,‡]

[*] *Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Centre, PO Box 9101, Nijmegen, the Netherlands*
[§] *Netherlands Bioinformatics Centre, 260 NBIC, P.O. Box 9101, 6500 HB, Nijmegen, the Netherlands*
[#] *WUR, Laboratory of Microbiology, Dreijenplein 10, Buildingnumber 316, 6703 HB Wageningen, the Netherlands*
[†] *TI Food and Nutrition, P.O. Box 557, 6700 AN Wageningen, the Netherlands*
[‡] *NIZO food research, Kluyver Centre for Genomics of Industrial Fermentation, P.O. Box 20, 6710 BA Ede, The Netherlands*

Metagenomics is the study of consortia of micro-organisms that co-exist in many environments, e.g., sea, soil, gut, skin. In these studies, next-generation-sequencing technology is used to generate millions of sequence reads from these consortia. The quality control, handling, assembly of reads into contigs, and functional annotation of either contigs or reads of these large datasets requires significant computing resources.

In this proposal, we aim to create Grid and Cloud computing solutions for metagenomic datasets within the NBiC E-BioScience taskforce. These solutions will be used to analyze datasets that are generated in public-private partnerships such as TI Food and Nutrition, the Kluyver Center for Genomics of Industrial Fermentation, NIZO food research, and academic groups. The E-BioScience metagenomics project will closely collaborate with the Metagenomics taskforce that is currently being setup at NBiC, SARA, BigGrid, and the NBiC next-generation-sequencing platform. Project objective is to provide publicity on use-cases of Grid and Cloud computing applied to metagenomic datasets.

9

PhenoLink – a web-tool for linking phenotype to ~omics data for bacteria: application to gene-trait matching for Lactobacillus plantarum strains

Jumamurat R Bayjanov [*,§], Douwe Molenaar [‡,#], Vesela Tzeneva [†,‡], Roland J Siezen [*,§,†,‡], Sacha A F T van Hijum [*,§,†,‡]

[*] Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Centre, PO Box 9101, Nijmegen, the Netherlands
[§] Netherlands Bioinformatics Centre, 260 NBIC, P.O. Box 9101, 6500 HB Nijmegen, the Netherlands
[†] TI Food and Nutrition, P.O. Box 557, 6700 AN Wageningen, The Netherlands
[‡] Kluyver Centre for Genomics of Industrial Fermentation, NIZO food research, P.O. Box 20, 6710 BA Ede, The Netherlands
[#] Systems Bioinformatics IBIVU, Free University of Amsterdam, 1081HV, Amsterdam, the Netherlands

Background
Linking phenotypes to high-throughput molecular biology information generated by ~omics technologies allows revealing cellular mechanisms underlying an organism's phenotype. ~Omics datasets are often very large and noisy with many features (e.g., genes, metabolite abundances). Thus, associating phenotypes to ~omics data requires an approach that is robust to noise and can handle large and diverse data sets.

Results
We developed a web-tool PhenoLink (http://bamics2.cmbi.ru.nl/websoftware/phenolink/) that links phenotype to ~omics data sets using well-established as well new techniques.
PhenoLink imputes missing values and preprocesses input data (i) to decrease inherent noise in the data and (ii) to counterbalance pitfalls of the Random Forest algorithm, on which feature (e.g., gene) selection is based. Preprocessed data is used in feature (e.g., gene) selection to identify relations to phenotypes. We applied PhenoLink to identify gene-phenotype relations based on the presence/absence of 2847 genes in 42 Lactobacillus plantarum strains and phenotypic measurements of these strains in several experimental conditions, including growth on sugars and nitrogen-dioxide production. Genes were ranked based on their importance (predictive value) to correctly predict the phenotype of a given strain. In addition to known gene to phenotype relations we also found novel relations.

Conclusions
PhenoLink is an easily accessible web-tool to facilitate identifying relations from large and often noisy phenotype and ~omics datasets. Visualization of links to phenotypes offered in PhenoLink allows prioritizing links, finding relations between features, finding relations between phenotypes, and identifying outliers in phenotype data. PhenoLink can be used to uncover phenotype links to a multitude of ~omics data, e.g., gene presence/absence (determined by e.g.: CGH or next-generation sequencing), gene expression (determined by e.g.: microarrays or RNA-seq), or metabolite abundance (determined by e.g.: GC-MS).

10

Detection of pathogenic copy number variations in shallow next generation sequencing data

Daphne van Beek

TU Delft, VUMC

Structural variations in the chromosome have proven to be an important factor in the development of various diseases, such as intellectual disability. The change in the copy number in certain regions of the DNA can be part of the natural variation among humans but can also be pathogenic.
The goal of this research is to determine if next generation sequencing (NGS) techniques can replace the detection of copy number variations (CNVs) using aCGH arrays in a clinical setting and, if this is the case, determine what coverage is sufficient to detect CNVs of a certain size.

NGS data (Illumina HiSeq2000) of three patients with a pathogenic CNV are available. aCGH array data is also available for these patients for comparison with CNV detections made with NGS data. Alignment and filtering of the reads (to reduce the formation of read-towers) is performed and multiple CNV detection tools are tested.

The formation of read-towers influences CNV detection algorithms that are based on read depth. Removing duplicates and using a filter that removes stacks of reads further improves the preprocessing of the data.

11

Using grid middleware as a computational resource for large-scale NGS and imputation in the eBioGrid project

George Byelas[1], Martijn Dijkstra[1], Alexandros Kanterakis[1], Pieter Neerincx[1], Freerk van Dijk[1], Tom Visser[2], Jan Bot [3,5], Irene Nooren [4,5] and Morris Swertz[1]

[1] *UMC Grongingen*
[2] *SARA*
[3] *Delft University of Technology*
[4] *University of Amsterdam*
[5] *e-BioGrid*

Specifying computation protocols for NGS and imputation analysis and running these at a large scale is a complicated task, due to many steps involved, long runtimes, heterogeneous clusters and large files. This process becomes error-prone when treating hundreds of samples, such as in genomic (diagnostic) facilities, if done without an integrated workflow framework and data management system.

From NGS and imputation projects we learnt that bioinformaticians don't want high-level workflow management systems with graphical user-interfaces, but prefer low-level shell scripts they can fully control instead. In addition researchers like to concentrate on the biology of their analyses and have automatic mechanisms taking care of complex job submission and monitoring details for grid computing.

Here we will demonstrate a 'just-enough' workflow declaration and execution system to address these needs. It is build on top of the MOLGENIS framework [1, 2, 3] for data tracking and uses templates to generate analysis scripts, This lightweight environment has been developed in collaboration with bioinformaticians working on NGS and imputation projects using grid or cluster middleware in particular in the nation-wide 'Genome of the Netherlands' project (700TB of data; >200.000 compute hours).

References:
1. Byelas HV, Swertz MA, (2012), Introducing data provenance and error handling for NGS workflows within the MOLGENIS computational framework, in proceedings of the BIOSTEC BIOINFORMATICS-2012
2. Swertz MA, Parkinson H, (2010), The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button, in BMC Bioinformatics. 2010 Dec 21:11 Suppl. 12:S12
3. Swertz MA, Jansen RC, (2007) Beyond standardization: dynamic software infrastructures for systems genetics, in Nature Reviews Genetics 8

12

VASP3 - Central Storage and Analysis Pipeline for Transcriptomics Data

Maarten A. Bijl[1], Henk-Jan van den Ham[1], Marinus J.C. Eijkemans[2], Albert D.M.E. Osterhaus[1], Arno C. Andeweg[1]

[1] *ErasmusMC, Dept of Virology, Rotterdam, the Netherlands*
[2] *Julius Center for Health Sciences and Primary care, UMC Utrecht*

VASP3 is a software tool specifically designed to facilitate the uniform analysis of -omics data in a multi-centre research project. It centralizes the storage and bio-informatics analysis of multiple related microarray data sets in a "low-threshold" environment, bringing together researchers from various disciplines. Furthermore, VASP3 offers web based transparent data analysis, prevents data redundancy, offers secure and private data access and contributes to experiment reproducibility. VASP3 is a web application written in Java/ICEfaces which serves as a user-friendly graphical shell around the R statistical programming language. It's a deliberately lean platform that guides its users through a logical sequence of analysis steps found to be used frequently during pre-VASP analysis ventures. In a nutshell these are:
- data preprocessing, including variation removal, background correction, and normalization
- quality control of both raw and normalized data
- identification of differential gene expression on gene level as well as gene set (pathway) level
- data visualization

Raw data as well as processed data, at any intermediate analysis step, are stored in a central location and the analysis history is fully logged. Users are encouraged to comment on their own analyses. The comments serve as an electronic lab journal. These VASP3 key features highly contribute to experiment reproducibility. Data analysis is facilitated by several data visualization approaches implemented in VASP3 (e.g., boxplots, MA-plots, PCA plots, heatmaps, Venn diagrams). For any post-VASP analysis, data can be exported in different formats or linked to commercial software like, for example, Ingenuity IPA. VASP3 is hosted on a public application server and is therefore accessible for all partners in a project, independent of geographical location. However, data sets within the system are owned privately unless explicitly made available by the owner of the data.

Because only privileged users can connect to the system, solely through encrypted SSL tunnels, VASP3 offers safe and secure data storage. The intuitive GUI and extensive help function take away the need to learn programming in R, enabling researchers with varying backgrounds to work together on the same project all in one place. Because of its centralized nature, remote accessibility and data sharing options, VASP3 is an ideal –omics data analysis tool for collaboration between partners in a large consortium like VIRGO*.

The emphasis has been put on the processing of transcriptomics microarray data. Currently, tools for the analysis of Next Generation Sequencing data (RNA-Seq) are being implemented. In the future, utilities for proteomics and metabolomics data analysis may also be added. Until now only R/Bioconductor generated scripts can be handled by VASP3, but virtually any programming/scripting language (BioJava, Python, etc.) can be used, depending on which is best for the research question at hand. Although developed for large research projects, the application can be used within smaller scaled projects or even by individual users. We anticipate to release the VASP3 pipeline as open source software.

*This study was supported by the Virgo consortium, funded by the Dutch government project number FES0908, and by the Netherlands Genomics Initiative (NGI) project numbers BSIK 03012 and 050-060-452.

13

Ready-made IT platform for diagnostic analysis of medical data

A. Michie, T. Binsl and R. Nanninga

Crosslinks B.V., Willemskade 18C, 3016DL Rotterdam

As a consequence of the new 'Omics Age', research organizations as well as biotech and diagnostic companies are nowadays able to develop increasingly complex diagnostic and prognostic techniques. However, the certification of complex diagnostics with authorities such as the FDA remains a major obstacle for deploying or developing a diagnostic product quickly and efficiently.

One reason for this is the need of a validated IT environment and the fact that many parties lack the knowledge, time or money to develop their own IT infrastructure in a validated way. Therefore, we have built our ready-made IT platform vDAP for the diagnostic analysis of medical data allowing research organizations and companies to concentrate on their core competencies – which is developing diagnostic products and not validated IT infrastructure. Because the bulk of the platform is standardized, rapid customization for new diseases, devices and data types can be performed allowing customers a faster time-to-market yet with lower business and regulatory risks. In addition, centralized processing in a secure environment offers protection of IP and simplifies the collection of data for analysis or providing regulatory information.

Using an example application for bacterial diagnostics - developed with our partner IS-Diagnostics - we demonstrate how vDAP enables fast, secure and reliable diagnostics. In addition, we show how the use of TIBCO Spotfire empowers IS-Diagnostics to get more out of their data to accelerate research and product development.

14

Optimal and robust regulation of gene expression

Evert Bosdriesz, Douwe Molenaar, Bas Teusink and Frank Bruggeman

NBIC, VU Amsterdam, CWI

Unicellular organisms operate in changing environments. They are subject to strong selective pressure on growth rate maximization which requires the maximization of specific rate of catalysis in biosynthetic modules and ultimately of the entire self-replicating machinery. This goal is attained by synthesizing the biosynthetic machinery in proportions that are optimal under the given environmental conditions. This raises the question what kind of regulatory network can reliably tune gene expression levels towards their optimum. Here we present a simple feedback mechanism for the regulation of gene expression that maximizes the flux per total amount of enzyme in a pathway. We show that this mechanism is able to attain optimal expression levels over a wide range of external conditions, and that the ability to do so is extremely robust. These properties arise because, for nearly optimal enzyme concentration, small fluctuations in the enzyme concentration lead to extremely large fluctuations in the metabolite levels, making these a very reliable regulatory signal. As an example, we study the regulation of the main biosynthetic machinery of Escherichia coli in silico. Because of its simplicity and robustness, we expect that this feedback motif is more commonly used in gene expression regulation.

15

BiG Grid infrastructure: Facilitating the computational needs for current bioinformatics applications

Jan J. BOTb,d,e, Irene M. NOORENa,b,d, Machiel JANSENc, Coen Schrijversc, Joost N. KOKb,d and Timo M. BREITa,d

a Microarray Department & Integrative Bioinformatics Units, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands
b Leiden Institute of Advanced Computer Science, Leiden University, Niels Bohrweg 1, 2333 CN, Leiden, The Netherlands
c SARA, Science Park 140, 1098 XG Amsterdam, The Netherlands
d Netherlands Bioinformatics Centre (NBIC), Geert Grooteplein 28, 6525 GA Nijmegen, The Netherlands
e Delft Bioinformatics Laboratory, Delft University of Technology, Mekelweg 4, 2628CD, Delft

The BigGrid computational resources are evolving to better support the computational needs of the life sciences. Typically, large data sets such as genome and imaging data are being produced that require analysis with sufficient computational infrastructure. Here, we describe the most recent changes to the infrastructure to cope with these demands and give examples of how it can be used.

Replacement of Life Science Grid clusters
The oldest generation of Life Science clusters, 7 in total, will be replaced. The new hardware has been chosen to better suit the needs of bioinformaticians: the amount of scratch space and memory has been increased to be able to deal with the demands of e.g. the latest NGS pipelines and the number of cores has per machine have been increased to better support multi-threaded applications. Two new institutes are partipating: the VU (group Boomsma) and Hubrecht (group Cuppen).

Cloud facility
Based on pilots with bioinformatics applications, the HPC Cloud has reached production status at the beginning of 2012. This new facility allows users to upload their own virtual machine which they can use to install their own software. This flexible approach ensures easy access to scalable hardware.

Hadoop cluster
Hadoop allows users to more easily analyze large volumes of data. The new Hadoop cluster installed at SARA gives researchers in the Netherlands access to a compute environment based on this new technology and provides a completely new data processing paradigm.

Data ingest services
Currently, the transport of large volumes of data is often done using hard disks. This causes significant delays as the data needs to be extracted from the transport drives and copied to centralized storage. With the new data ingest server, located at SARA, many drives can be connected and copied simultaneously to grid or cloud storage, reducing the overhead of the copy process.

Lightpath connections
To further speed up collaborations between national and international partners a lightpath network is currently being rolled out. This network connects most of the Life Sience Grid clusters and aims to connect both the BGI and Complete Genomics.

Thanks to. BiG Grid (led by partners NCF, Nikhef and NBIC).

16

Structured data analysis and storage to facilitate systems biology analysis

Jildau Bouwman(1), Margriet Hendriks(2), Ben van Ommen(1) et al.

(1)TNO Quality of Life, Zeist, The Netherlands;
(2)UMC Utrecht, The Netherlands

In recent years, nutritional studies question the effects of a treatment, more often, in a holistic way. In these studies the effects of many molecules (transcripts, proteins, or metabolites) are studied in multiple organs. Often all data necessary to answer a biological question cannot be collected in one single study, which makes between study comparisons essential. Moreover, different measurement platforms are used for the analysis of different molecules (e.g. transcripts and metabolites). In order to relate the effects seen on the different platforms, knowledge on the relation between the different molecules is required. These developments require a new strategy for analysis and storage of data. First, for between platform comparison the relations between data sets should be clear (e.g. a transcript codes for an enzyme that is involved in the catalysis of a specific metabolite). Second, for between organ comparison the identity of the measured molecules should be standardized and the relation between the dataset should be stored. Third, for between study comparisons processing and analysis of data should be comparable and all study metadata should be available.

NuGO and NMC (Netherlands Metabolomics Centre) have joined forces to resolve these issues. The NMC- Data Support Platform (NMC-DSP) is developing a metabolomics pipeline to preprocess, analyze and store metabolomics data. NuGO is developing dbNP (nutritional phenotype database) to store and retrieve study data and clean omics-data. Their common interest is the storage of study metadata of studies with a complex study design. In collaboration a web-based, open-source study capture tool has been developed. The NuGO PPS (proof of principle) studies are stored in this application (test.dbnp.org).

Preprocessing, analysis, identification and biological interpretation tools, built within the NMC, and other metabolomics tools will be made available via a web interface using the galaxy application (www.nmcdsp.org).

17

Sampling the Sweet Spot

Oskar Bruning[1,2], Wendy Rodenburg[3], Martijs J. Jonker[1,2], Han Rauwerda[1,2], Annemieke de Vries[3], Timo M. Breit[1,2]

[1] MicroArray Department and Integrative Bioinformatics Unit (MAD-IBU); Swammerdam Institute for Life Sciences; Faculty of Science; University of Amsterdam (UvA); Amsterdam, The Netherlands
[2] Netherlands Bioinformatics Centre (NBIC); Nijmegen, The Netherlands
[3] Laboratory for Health Protection Research (GBO); National Institute of Public Health and the Environment (RIVM); Bilthoven, The Netherlands

Background:
The design of –omics experiments has to be optimized by taking into account statistical and biological principles of experimental design. Statistical principles for experimental design are well known, e.g. orthogonality, factorial designs, and replication. Biological principles are less established, but are related to a key aspect of genetic responses: they occur in specific timescales and in relation to changes in their environment and cell state. Traditionally, perturbation strength and sampling times are often determined using other endpoints than gene expression (such as apoptosis).

Results:
We show that by using gene expression as endpoint the results of perturbation and longitudinal studies are less noisy and better interpretable. As an example our own study on the effects of UV-radiation in time on WT and p53-mutant mice was used (1). This study used a high dose of UV-radiation, which was based on the phenotypic endpoint apoptosis. The results of the analysis of this experiment left us with a large amount of highly significant, differentially expressed genes (DEG's). However, these high numbers of DEG's made the data largely un-interpretable, probably due to many a-specific responses that clouded the relevant specific responses (2). We propose the following protocol for doing small-scale dose and time range finding before running large-scale studies:
1) Take "external" control measurements such as RNA yield, DNA content, cell counts, with the goal to relate gene expression to biomass.
2) Determine relevant gene expression endpoints: i.e. gene sets (e.g. p53 signalling, DNA repair)
3) Use many samples and few replicates (if any)
4) Determine the sweet spot

Conclusion:
Biological principles for experimental design need to be formalized. We show here that this is crucial for interpretable –omics experiments.

References:
1. The absence of Ser389 phosphorylation in p53 affects the basal gene expression level of many p53-dependent genes and alters the biphasic response to UV exposure in mouse embryonic fibroblasts. Bruning et al. Mol Cell Biol. 2008 Mar;28(6):1974-87.
2. Serious complications in gene-expression studies with stress perturbation: An example of UV-exposed p53-mutant mouse embryonic fibroblasts. Bruning et al. Transcription. 2010 Nov;1(3):159-164.

18

Marker2sequence: From QTLs to potential candidate genes

Pierre-Yves Chibon, Richard Visser, Richard Finkers

Wageningen UR Plant Breeding

Quantitative Trait Loci (QTL) mapping is the process of determining the association between a trait and genetic marker data using statistical approaches. This technique is used by plant breeders on a regular basis within their research. However, QTL analysis by itself is not useful for determining which gene is the potential candidate explaining the observed phenotype. For this, a link between the genetic markers and the genome sequence should be established. Nowadays, this link is becoming available for more and more genomes due to the explosion in the use of Net Generation Sequencing (NGS) approaches. Within our research, we want to integrate QTL with the genome information and investigate the genomic region for potential candidate genes. For this we created the tool caller Marker2sequence. It is a web-based tool integrating the genome functional annotation with GO terms, proteins, pathway and literature information. Through this integration the related annotation for a gene can be presented and a short list of potential candidates genes can be generated for a given QTL.
This tool is thus helpful for both biologists and breeders to understand the underlying genomic regulation of an observed phenotype.
URL: https://www.eu-sol.wur.nl/marker2seq/

19

Ongoing efforts of a Galaxy solution for the BiGGrid/SARA HPC Cloud

Mattias de Hollander[1], Hailiang Mei[2], David van Enckevort[2] and Eiko Kuramae[1]

[1] *NIOO-KNAW: Netherlands Institute of Ecology*
[2] *NBIC: Netherlands Bioinformatics Centre*

The rapid evolution of next generation sequencing technologies (NGS) together with decreasing costs are creating a challenge to store and analyze the vast amount of sequencing data that are generated by experimental biologists. Configuring suitable data analysis software and having access to readily available computation and storage resources are the two major bottlenecks faced by many research groups. In this on-going collaboration, NIOO and NBIC BioAssist program are building Galaxy@HPCcloud solution to jointly tackle these two challenges.
This collaboration is motivated by two major developments in the Dutch Life Science community: the launch of High Performance Cloud at BigGrid/SARA and the increasing adoption of Galaxy as a common data analysis platform across different research groups. "Cloud computing" enables on demand access to the needed hardware resource for a certain period of time, while Galaxy bundles a set of advanced software programs dealing with various biological datasets and provides an easy and standard interface. This solution was demonstrated successfully by running NIOO Galaxy at the beta version of the HPCcloud using the Galaxy Cloudman system.
The goal of this joint project is two-fold. First, we will develop the Galaxy@HPCcloud solution together by expertise sharing and tackle the common problems like resource auto-scaling and persistent data storage. Second, we plan to make Galaxy@HPCcloud available to researchers in a dual way. One is to provide Galaxy@HPCcloud images to setup your own instances similar to the NIOO project in the beta cloud system. The other is to install the current NBIC Galaxy server (http://galaxy.nbic.nl) at the HPC cloud. This new NBIC Galaxy@HPCcloud will provide a more complete list of tools that can be used for Genomics and Proteomics data analysis and become a demonstration place for the tools and pipelines developed by all NBIC BioAssist developers.

20

Managing cloud computing for life sciences research via smart interfaces

Wim de Leeuw, Linda Bakker, Han Rauwerda & Timo M. Breit

MicroArray Department & Integrative BioInformatics Unit, University of Amsterdam

In a large class of bioinformatics applications, the processing power required fluctuates strongly and it is not feasible nor needed to keep the maximum processing capability available locally all the time. The SARA HPC-cloud offers compute power on demand in the form of freely configurable virtual machines. In the cloud one can configure a system: number of cores, amount of memory, secondary storage and network of the machine and freely install desired software running on this machine. These machines are stored as images, which can be deployed at a later time.

We have implemented a system which can be used to control deployment of machine images in the cloud bypassing the cloud user interface. Using this system it is easy to setup applications in which cloud resources are transparently used from outside. It consists of a lightweight server, which is capable of starting and stopping machine images, just as is possible through the web-interface. It also keeps track of running machines under its control. Clients can request the starting of machines or request information about running machines. These clients can be used in applications to access cloud resources with minimal user intervention.

We describe two use cases: the first one is about creating an R-cluster in the cloud. In this use case a user can start an R-cluster on the cloud from within an local R session and distribute the calculation work using the normal R-cluster commands over the cloud. The second use case is the back end of the array designer web-application. In the web application, the user can generate a microarray design based on input sequence data and a number of additional parameters. The required resources for generation of the array design are not available on the web-server and in this use case, the work is done in the cloud. For each array design a machine is instantiated on the cloud and stopped when the design is ready.

21

Determining gene regulation mechanisms through multi-scale integrative analysis of genomic signals

J.R. de Ruiter, J. de Ridder, T.A. Knijnenburg

Delft University of Technology, Netherlands Cancer Institute (NKI-AVL)

The identification of the functional elements in the genome and their interactions are key to understanding the processes that take place within a cell. Integrating the wealth of information present in available high-throughput datasets, such as those stemming from Chip-seq and RNA-seq experiments, is crucial to unraveling the complex mechanisms that underlie the regulation of human gene expression.

We present a framework that approaches the prediction of gene expression as a classification problem and predicts gene expression levels from genomic signals measuring the presence of functional elements. The constructed model applies a scale-space analysis to capture the spatial locality (relative to the gene) and scale of the signals by employing a multi-scale segmentation algorithm. This is important since functional elements manifest themselves at various distinct genomic scales. Taken together, our method places special emphasis on the biological interpretability of the model to allow easy derivation of possible mechanistic interactions between functional elements and the spatial scales at which they interact.

Preliminary results show that a classification model is able to predict the expression class of a gene (high/low expression) with a high accuracy from ChIP-seq methylation signals (H3k4, H3k27, H3k36). Analyses of the model have shown that each of these signals has a specific effect on gene expression and that they exhibit distinct preferences for spatial locality and scale. This demonstrates that our approach is able to capture spatial characteristics of genomic signals and use this information to accurately predict gene expression.

22

Effect of population specific imputation reference set using second genotype chip as gold standard

Alexandros Kanterakis, Javier Gutierrez-Achury, Isis Ricaño Ponce, Patrick Deelen, Cisca Wijmenga and Morris Swertz
Members of GoNL consortium and imputation working group

Genomics Coordination Center & Department of Genetics, University Medical Center Groningen & University of Groningen, Groningen, The Netherlands

We here present the first explorative results of using the Genome of the Netherlands (GoNL) as imputation reference. GoNL is a whole genome deep sequencing project of 250 father-mother-child trios. The result is a comprehensive map of common and rare alleles specific to the Dutch population. In particular the presence of trios enables a more accurate phasing process that should be useful to predict low frequency variants in other individuals via imputation [1]. Such imputed data can be useful for fine-mapping regions apart of the possibility of finding new hits in genome wide association studies (GWAS). However, the advantage of using population specific sequenced variants have not been tested before, because of the cost-expensive of an initiative like GoNL and the lack of gold standard real genotypes that allows the direct comparison.

To evaluate the impact of population specific imputation reference we assed 750 Dutch individuals coming from a Celiac disease GWAS that were genotyped using the Hap550 platform and the custom design ImmunoChip. The Hap500 genotypes were imputed using three different reference sets: the GoNL pilot release, 1000 genomes project [2] and HapMap2 [3]. We calculated the quality metrics by comparing the imputed variants to the 140,197 variants that are unique to the ImmunoChip. 65,274 of these variants have a minor allele frequency lower than 0.05 allowing us to also ascertain the quality of the imputation of rare variants. We used the Immunochip platform as a genotyping 'gold standard' for a high number of individuals/markers to measure in a straightforward way the performance of these different reference sets.

1. Marchini, J. and B. Howie, Genotype imputation for genome-wide association studies. Nat Rev Genet, 2010. 11(7): p. 499-511.
2. The 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing. Nature, 2010. 467(7319): p. 1061-73.
3. Frazer, K.A., et al., A second generation human haplotype map of over 3.1 million SNPs. Nature, 2007. 449(7164): p. 851-61

23

Compressed Adjacency Matrices: Untangling Gene Regulatory Networks

K. Dinkla and M.A. Westenberg

Eindhoven University of Technology

Compressed Adjacency Matrices provide an alternative to the infamous hairballs that are generated by applying standard visualization and layout techniques to bacterial gene regulatory networks.

Our approach isolates and emphasizes topological characteristics inherent to GRNs of bacteria. The adjacency matrix of such a network is large but also sparse. By cutting open and folding this matrix onto itself, we achieve a compact and neatly-arranged visualization. This enables quick visual detection of motifs while preserving the context of the network as a whole. Moreover, many of the visual analytics techniques that apply to adjacency matrices can also be used for this compressed variant.

We demonstrate compressed adjacency matrices on the GRNs of Escherichia coli and Bacillus subtilis. Our interactive prototype supports arrangement clustering, highlighting, and filtering.

Link: www.win.tue.nl/~kdinkla/CAM.png

24

Bioinformatical analysis of the GRAS family of proteins

Magdalena Dymecka[1,2], Katarzyna Kokoszyńska[2], Leszek Rychlewski[1], Cordelia Bolle[3], Lucjan S. Wyrwicz[2]

[1] *Bioinfobank Institute, Limanowskiego 24A, 60-744, Poznań, Poland*
[2] *Laboratory of Bioinformatics and Biostatistics, M. Skłodowska-Curie Cancer Center and Institute of Oncology, WK Roentgena 5, 02-781 Warszawa, Poland*
[3] *Ludwig-Maximilians-Universitat, Biozentrum, Department Biologie I, Großhaderner Strasse 2-4, 82152 Planegg-Martinsried*

GRAS proteins, which are a plant-specific family of proteins, play an important role in plant growth and development, as well as take part in processes such as signal transduction, gibberellic acid (GA) signalling and meristem maintenance. The exact molecular mechanism by which these proteins act remains unknown, therefore a detailed bioinformatical analysis was performed.

As described in previous reports, GRAS proteins are thought to belong to the transcription factor family, due to their nuclear localization and the presence of five characteristic sequence motifs. Here we present the results of a detailed bioinformatical study on GRAS proteins. The sequence analysis revealed the presence of two regions - the N-terminal DELLA domain and the C- terminal domain (also known as the GRAS domain). Mapping of the identified functional regions revealed the presence of a novel methyltransferase fold within the GRAS domain . The obtained results allow us to pose new hypotheses on the function and molecular action of these proteins.

25

COMPANION: Comparative genome annotation in prokaryotes: a halt to error propagation?

Thomas H. A. Ederveen[1], Amy de Bruin[2], Brechtje Hoegen[2], Bernadet Renckens[1], Roland J. Siezen [1,4,5], Sacha A. F. T. van Hijum[1,3,4,5]

[1] *Centre for Molecular and Biomolecular Informatics, Radboud University Medical Centre, Nijmegen, the Netherlands.*
[2] *HAN University of Applied Sciences, Nijmegen, the Netherlands.*
[3] *NIZO food research, Ede, the Netherlands.*
[4] *Top Institute Food and Nutrition, Wageningen, the Netherlands.*
[5] *Kluyver Centre for Genomics of Industrial Fermentation, Delft, the Netherlands.*

Background
With next-generation sequencing, genome sequences are generated faster than the capacity to adequately annotate these assembled genome sequences. Automated genome annotation engines facilitate genome annotation. However, these engines suffer from inaccuracy in gene coordinate- and function prediction compared to manually annotated and/or curated efforts by scientists. In public sequence databases, many erroneous gene annotations exist making database error propagation a great concern as annotation engines rely on these databases for their predictions. Many automated annotation pipelines exist, but the choice of what engine to use seems an arbitrary one.

Method
In this work, our goal is to significantly reduce the number of genes that need to be manually checked and curated. To this end, we combine for a given genome sequence the results of multiple automated annotation engines into a consensus gene coordinate- and function prediction. We hypothesize that this consensus prediction will be more accurate compared to the use of solely one individual engine. We present a pipeline for processing data from popular, publicly available automated genome annotation engines for prokaryotes: BASys, IGS, ISGA, RAST and xBASE. New engines can be added relatively straight-forward. It compares coordinate- and function prediction calls made by the various engines and provides the user with weighted gene coordinate- and function predictions. We use rules based on prior-knowledge in addition to majority voting. These rules are based on errors that frequently occur with a given annotation engine for certain genes. They allow favoring certain “trusted” predictions based on prior-knowledge over those provided by majority voting. These “trusted” predictions are currently based on three phylogenetically diverse bacteria with manually curated genome annotations: Moraxella catarrhalis RH4, Lactobacillus plantarum WCFS1 and Lactococcus lactis KF147.

Conclusion
By analysis of annotations from different engines for these three strains we can: (i) (for specific gene sets) identify automated annotation engine specific biases in their method of gene coordinate- and function prediction; (ii) build up a vocabulary of annotation terms that are equivalent across different engines, termed a translation list; and, (iii) by data analysis and pattern recognition, derive specific prior-knowledge based rules that allow circumventing error propagation and if relevant, remove annotation engine specific bias. Our comparative annotation pipeline provides more accurate gene coordinate- and function predictions, leaving the curator with only a subset of genes to be manually checked and/or curated.

26

NatalieWEB: A web server for topology-aware global protein-protein interaction network comparison

Mohammed El-Kebir [1,2,3], Bernd W. Brandt [4], Jaap Heringa [2,3,5] and Gunnar W. Klau [1,3]

[1] *Life Sciences, Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, the Netherlands*
[2] *Centre for Integrative Bioinformatics VU (IBIVU), VU University Amsterdam, De Boelelaan 1081A, 1081 HV Amsterdam, the Netherlands*
[3] *Netherlands Institute for Systems Biology, Amsterdam, the Netherlands*
[4] *Department of Preventive Dentistry, Academic Centre for Dentistry Amsterdam (ACTA), University of Amsterdam and VU University Amsterdam, the Netherlands*
[5] *Netherlands Bioinformatics Centre, Geert Grooteplein 28, 6525 GA Nijmegen, the Netherlands*

We present NatalieWEB, a web server for topology-aware alignment of a specified query protein-protein interaction network to a selected target network. By incorporating both similarity at sequence and network level, NatalieWEB computes alignments that allow for the transfer of functional annotation as well as for the prediction of missing interactions.

We highlight the capabilities of NatalieWEB through a biological case study involving the Wnt signaling pathway by demonstrating that topology-aware network alignment produces better results than traditional comparison solely based on sequence. We also show that NatalieWEB is able to predict putative interactions. The server is available at: www.ibi.vu.nl/programs/nataliewww/.

27

A sequence optimization method for increased gene expression

Alexey A. Gritsenko, Marcel J.T. Reinders, Dick de Ridder

The Delft Bioinformatics Lab, Kluyver Centre for Genomics of Industrial Fermentation, Platform Green Synthetic Biology

Given an amino acid sequence of a protein, multiple degrees of freedom are available to tune efficiency and fidelity of its translation for a specific organism or condition. These degrees of freedom are made available through redundancy of the genetic code, in which multiple synonymous codons may encode the same amino acid. Although synonymous, preferential usage of these codons can influence translation initiation and elongation.

In practice biased usage of the genetic code is harnessed to increase protein levels in heterologous gene expression experiments via a process called codon optimization. Existing codon optimization algorithms employ simplistic models of the process of translation, often resulting zero or lower protein expression compared to the non-optimized version of the gene.

We present on-going work on a sequence optimization method, which given a reference set of highly expressed genes, adjusts codon composition of a target sequence to match local and global codon usage patterns of the reference set. The employed model is extendable and currently takes codon usage, tRNA abundance and tRNA recycling into account.

28

Dockland's got talent! Determining relations between structure quality and docking quality.

B. Hanssen, S.C.B. Jans, G. Vriend

Radboud University

Molecular docking can predict the preferred orientation of a ligand in the active site pocket of a receptor. With more accurate atomic coordinates, one would think that it should be possible to achieve a more reliable docking result.

By performing molecular (self-)docking with different tools (Autodock, FlexX, Fleksy) on original and re-refined PDB-files, several docking complexes were obtained. Of these docking complexes the RMSD (Root Mean Square Deviation) between the docked ligand and the co-crystallised ligand was calculated. From these results unfortunately it can be concluded that the quality of the protein structure is not the limiting factor in molecular docking.

29

Eighteen-fold Performance Increase for Short-Read Sequence Mapping on Genome of the Netherlands (GoNL) data using a Hybrid-Core Architecture

Hans Heideman [1,2], Kirby Collins[2], George Vacek[2], Jan Bot[3,4]

[1] *Corresponding author: hheideman@conveycomputer.com, +31 6 5141 9507*
[2] *Convey Computer Corporation, Richardson, TX, USA.*
[3] *Leiden University, Leiden, The Netherlands.*
[4] *Technical University Delft, Delft, The Netherlands.*

Advances in sequencing technology have significantly increased data generation and require similar computational advances for bioinformatics analysis. Advanced architectures based on reconfigurable computing can reduce application run times from hours to minutes and address problem sizes unattainable with commodity servers. The increased capability also fundamentally improves research quality by allowing more accurate, previously impractical, approaches. This work describes the use of a hybrid-core computing architecture, combining a traditional x86 environment with a reconfigurable co-processor, to solve a data-intensive problem of next-generation sequencing analysis: reference mapping of short-read sequences.

Bioinformatics applications consist of large numbers of relatively simple operations on large randomly accessed data structures. Conventional architectures lack sufficient parallelism in the core processing elements and the memory subsystem to efficiently execute these algorithms. The Convey hybrid-core (HC) architecture incorporates both a highly parallel processing architecture and a highly parallel, randomly accessible memory subsystem. Algorithms also benefit from massively parallel implementations of application-appropriate-data-type operations, which use logic gates more efficiently than commodity servers. In Burrows-Wheeler mapping applications, significant gains are made.

Results are presented comparing the performance of Convey's BWA pipeline with standard BWA. The Convey accelerated BWA pipeline on an HC server can deliver up to 18x the throughput of BWA on a standard x86 system as demonstrated using human genome sequence data from the Genome of The Netherlands (GoNL).

30

Data-driven characterisation of protein signalling networks in cancer

Steven M. Hill [1,2,3], Yiling Lu[4], Jennifer Molina[4], Gordon B. Mills[4], Sach Mukherjee [1,3,2]

[1] *The Netherlands Cancer Institute (NKI), Amsterdam, The Netherlands*
[2] *Centre for Complexity Science, University of Warwick, UK*
[3] *Department of Statistics, University of Warwick, UK*
[4] *Department of Systems Biology, University of Texas M.D. Anderson Cancer Center, Houston, TX*

Signalling networks play a key role in cellular function, and their dysregulation is central to many diseases, including cancer. However, we remain limited in our understanding of cancer-specific changes to signalling networks. Protein signalling involves combinatorial interactions between multiple proteins, ultimately driving downstream cellular effects. To shed light on signalling network topology in a sample of interest requires interrogation of multiple proteins through time and statistical approaches to make inferences regarding network structure. To date, such data-driven characterization of context-specific signalling networks has remained a challenging and open problem. Here, we utilise stochastic models known as dynamic Bayesian networks to make inferences regarding network structure and thereby generate testable hypotheses. We take a Bayesian approach, incorporating existing biological knowledge by means of an informative network prior, weighted objectively by an empirical Bayes approach. Instead of resorting to approximate schemes, such as MCMC, we exploit a connection between variable selection and network inference to enable exact calculation of posterior probabilities of interest. The procedure is computationally efficient and essentially free of user-set parameters. Results on simulated data place the approach favourably relative to other existing network inference approaches. We apply these methods to phospho-proteomic time series data from a breast cancer cell line (MDA-MB-468) and predict novel signalling links, which we independently validate using targeted inhibition. The methods proposed offer a general approach by which to elucidate molecular networks specific to biological context (i.e. cell type, tissue or disease state), including, but not limited to, human cancers.

31

Advanced SKOS-aided Query Builder (ASQ-builder): Building smarter queries for public databases (PubMed, OMIM) using SKOS Vocabularies

Gerbert A. Jansen [1,2], Angela C.M. Luijf[1], Andrew P. Gibson[4], Eelke van der Horst[1], Ronald J.A. Wanders[5], Peter G. Barth[6], Antoine H.C. van Kampen [1,2,3,4]

[1] Bioinformatics Laboratory, department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Centre - University of Amsterdam, Amsterdam, The Netherlands
[2] Netherlands Bioinformatics Centre, Geert Grooteplein 28, 6525 GA Nijmegen, The Netherlands
[3] Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands
[4] Netherlands Consortium for Systems Biology, University of Amsterdam, Amsterdam, The Netherlands
[5] Laboratory Genetic Metabolic Diseases, departments of Clinical Chemistry and Pediatrics, Emma Children's Hospital, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands
[6] Department of Pediatrics, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands

In previous work [1], we have described the creation of a SKOS vocabulary containing clinical terms needed to describe the knowledge domain of clinical aspects of Zellweger Spectrum Disorders (ZSDs), a group of rare inherited metabolic disorders. The concepts contained in this SKOS vocabulary have been gathered from various sources, including a medical textbook [2] and through consulting clinicians. Concepts are richly annotated with alternative labels describing synonyms and abbreviations, definitions, and notes provided by clinical experts. Importantly, all terms have been curated by clinical experts in the field of peroxisomal disorders, ensuring a high quality of this set of clinical terms.

Furthermore, links to external public vocabularies (if available) are included in the vocabulary. We now demonstrate the advantage of the rich annotation in the SKOS vocabulary for building improved queries for the PubMed [3] and OMIM [4] databases.
A standard PubMed or OMIM query usually includes a single name for a certain term and returns those papers that contain this specific term only. In life sciences, especially in the fields of biochemistry, molecular biology, and medicine, a term often has an extensive number of synonyms and abbreviations, some of which more often used than others, while some are no longer used at all. When unknowingly choosing a disused term, a PubMed query returns 'old' publications only, and searching with a seldom used synonym returns very few papers.

To improve the construction of these queries, a web-application, the Advanced SKOS Query Builder (ASQ Builder), has been constructed, that uses a SKOS vocabulary. ASQ-Builder uses the main term from a vocabulary, and suggests and enables the addition of synonyms and abbreviations to this query, based on the altLabels from the SKOS vocabulary. Apart from including these suggestions, the query can be manually adjusted prior to execution. Subsequently, the query is executed against PubMed or OMIM and the results can be retrieved either in the PubMed or OMIM web-interface, or as a flat text file.

We demonstrate that these extended queries yield more complete results, which will be very valuable for anyone trying to find PubMed papers and OMIM entries, as not many users are aware of all synonyms or deprecated terms in the topic of interest.
The application of the Advanced SKOS Query Builder is not limited to a default (preloaded) SKOS vocabulary. Any SKOS vocabulary created with our freely available SKOS Vocabulary Editor [5] can be used for creating smart queries for the research domain of interest.

References

[1] Jansen GA, Gibson AP, Barth SL, Wanders RJA, Barth PG, van Kampen AHC: The Peroxisome Knowledge Base within the BioExpert Framework: Development of a SKOS Vocabulary on Clinical Information on Peroxisomal Disorders. 2011.
[2] Gould SJ, Raymond GV, Valle D: The Peroxisome Biogenesis Disorders. In Metabolic and Molecular Bases of Inherited Disease. 8th edition. New York: McGraw-Hill; 2001.
[3] PubMed - NCBI [http://www.ncbi.nlm.nih.gov/pubmed/].
[4] OMIM Home [http://www.ncbi.nlm.nih.gov/omim].
[5] SKOSVocabularyEditor [https://trac.nbic.nl/skosvocabularyeditor/].

32

PDB_REDO: optimized protein structures for bioinformatics research

Robbie P. Joosten

Netherlands Cancer Institute

Structural bioinformatics studies such as drug docking, homology modeling, biostatistics, etc. rely heavily on experimental X-ray structure models from the Protein Data Bank (PDB). The quality of these experimental models is an important limiting factor of the quality of bioinformatics studies. In other words, we can only do good bioinformatics with good structures. Unfortunately, many structures in the PDB are not as good as they can be because:

- ...they were made with protocols that have become obsolete due to the ongoing development of X-ray crystallographic methods.
- ...they are the result of manual interpretation of the data which is prone to inconsistencies and human errors.

The PDB_REDO databank of optimized structure models deals with these challenges. For each entry a structure model plus its original experimental X-ray data is taken from the PDB and optimized using a fully automated procedure. In the first phase, the model is refined to improve the fit with the experimental X-ray data. Here, the correct balance between protein geometry, temperature factors and the X-ray terms is established. The complexity of the temperature factor model is optimized using a Hamilton test based procedure to avoid overfitting or undermodelling. In the second phase, the model is further enhanced by rebuilding side chains, adding missing atoms, removing 'fake' waters, performing peptide flips and optimizing hydrogen bonding. The resulting model is refined again to get a fully optimized structure model. In the third phase, the new model is validated to check its fit with the experimental data and prior knowledge of protein geometry. Full cross validation is used for models with few experimental data.
Models in PDB_REDO have higher geometric quality (e.g. better Ramachandran plot, fewer bumps), energetically more favorable conformations and better fit with the experimental data over their counterparts in the PDB.

The PDB_REDO databank contains nearly 60,000 entries covering more than 95% of all X-rays structures in the PDB.

Link: www.cmbi.ru.nl/pdb_redo

33

A genetics-based approach to partitioning the natural diversity of an RNA virus family

Chris Lauber[1] and Alexander E. Gorbalenya [1,2,3]

[1] Molecular Virology Laboratory, Department of Medical Microbiology, Leiden University Medical Center, Leiden, The Netherlands
[2] A.N. Belozersky Inst. of Physico-Chemical Biology and 3Faculty of Bioengineering and Bioinformatics, M.V. Lomonosov Moscow State University, Moscow, Russia

The explosive accumulation of genome sequences in virology challenges virus taxonomy by experts who traditionally rely on extensive virus characterization. Here, we present a computational approach to hierarchically classify RNA viruses of a family basing only on genetic divergence. To quantify genetic divergence, we use pairwise evolutionary distances (PEDs) estimated by maximum likelihood inference on a multiple alignment of family-wide conserved proteins. PEDs are calculated for all virus pairs, and the resulting distribution is modeled via a mixture of probability density functions. The model enables the quantitative inference of regions of distance discontinuity in the family-wide PED distribution, which define the levels of hierarchy. For each level, a limit on genetic divergence, below which two viruses join the same group, is objectively selected among a set of candidates by minimizing violations of intragroup PEDs to the limit. In a case study, we apply the procedure to hundreds of genome sequences of viruses from the family Picornaviridae, for which an advanced expert-based taxonomy is available. Furthermore, we (i) evaluate the temporal stability of the classification and important key parameters, (ii) analyze the relation of virus sampling, host range, and observed genetic diversity of taxa, and (iii) provide an evolutionary perspective on known and still unknown natural diversity in the family. Since the use of genetic sequences for storage and transmission of information presents a universal property common to all biological entities, the developed approach might be applicable beyond RNA viruses.

http://f1000.com/13703957

34

Towards Watching the Bottlenecks of Membrane-Protein Biogenesis by Single-Molecule Methods

Da Liu, Jacek Mika, Victor Krasnikov, Lieke van Gijtenbeek, Michiel Punter, Jan Kok, Bert Poolman, Antoine van Oijen

Single-Molecule Biophysics, Zernike Institute for Advanced Materials, 9747AG, Groningen, the Netherlands

35% of the genome encodes for membrane proteins (MPs). Further, 75% of clinically used drugs are targeting membrane proteins. Understanding the functioning of MPs is a field of growing importance, but the production of MPs for basic study and characterization remains a major bottleneck. Overexpression of genetically recombinant proteins can lead to protein malfunction or can even be lethal to cells, especially in the case of overproduction of heterologous MP in bacteria. These bottlenecks have not been well documented. Cutting-edge singe-molecule fluorescence microscopy will be employed to shed light on each individual step in the process that leads to successful incorporation of functional protein into the cellular membranes. The use of photoactivable fluorescent proteins fused to the protein of interests will enable the sub-diffraction limited imaging of intracellular structure. By using photoactivated localization microscopy (PALM), which is a novel optical super-resolution method, we will allow for a spatial characterization of the intracellular distribution of MPs (or their mRNA precursors) at various stages of the transcription, translation, folding, and membrane-insertion processes. Further, by dynamically following the transcription, translation, and membrane insertion of individual molecules, we will be able to obtain information on which kinetic steps are affected when overexpressing a MP. By observing the cellular synthesis machinery at the single-molecule level we will obtain a level of mechanistic insight unobtainable by classical ensemble averaging methods.

http://singlemolecule.nl

35

Fast and accurate calculation of protein-protein interaction: contribution of surface and interface residues

René Pool, Ali May, Jaap Heringa and K. Anton Feenstra

IBIVU Center for Integrative Bioinformatics VU University

In many cellular processes, physical interactions between proteins play crucial roles. For biological functionality, many proteins must organise into protein complexes. On the other hand, biologically non-functional complexes might give rise to a number of pathologies. Existing methods for predicting protein-protein interactions (PPIs) that use statistical learning teachniques or other bioinformatic approaches appear to have limited accuracy. We will therefore try to build up a method based on physical principles using simplified molecular models for computational efficiency, while still maintaining the accuracy of calculating the interactions.

As a first step, we have established that we can accurately calculate potentials of mean force (PMF) using a coarse-grained force field where approximately four atoms are represented by a single sphere (or `super atom'). This yields almost a 100-fold simulation time speedup. As a second step, we will establish that direct interactions from the interface residues dominate the interaction potential. This opens the possibility of limiting the calculations to the interface region only, providing a speedup of another three orders of magnitude.

In order to verify the interface dominance, we have mutated surface and interface residues and studied the influence on the effective interaction of two different protein complexes; a TCR-pMHC and an MP1-p15 scaffolding complex. We performed random mutations on an equal number of residues in either the surface (non-interface), the interface core (buried in complex) or the interface rim (less exposed in complex). For each of these, we measure how favorable the interaction is between two proteins by calculating the PMF using coarse-grained molecular dynamics simulations at different distances. Our results show that mutations on the interface core lead to less attractive interactions than mutations on the interface rim, while non-interface surface mutations hardly effect the PPI.

We are now able to calculate an interaction potential between any two proteins in a matter of hours. Restricting this to the interface region only will allow us to calculate protein interactions on a genomic scale (tens of millions of PPIs) in about a day on a medium size compute cluster.

36

MS PRUNA: Mass Spectrometry Peptide Re-mining Using Network Analysis

E.A. McClellan, B. van Breukelen, A. Kremer, P.J. van der Spek, A.P. Stubbs

Erasmus MC, NBIC, NPC, Utrecht University

Here we describe Net WeAvers (Network Weighted Averages) for analyzing groups of regulated proteins, e.g. as defined by clusters of protein-protein interactions, which have been identified as significantly more or less abundant in a mass spectrometry experiment. Unlike gene set algorithms that generally use subjective cutoffs combined with the hypergeometric test, which is biased towards large sets, or methods that require a large number of user-defined parameters, Net WeAvers is a non-parametric and largely data-driven method that produces objective results.

e.mcclellan@erasmusmc.nl

37

e-Science in bioinformatics: enabling the reuse of methods while unravelling the epigenetic factors of Huntington's Disease

Eleni Mina[1,2], Reinout van Schouwen[1], Kostas Karasavvas[2], Kristina Hettne[1], Peter-Bram 't Hoen[1], Pernette J. Verschure [3,4], Christine Chichester[2], Barend Mons[1,2], Willeke van Roon-Mom[1], Marco Roos [1,2]

[1] *Department of Human and Clinical Genetics, Leiden University Medical Centre, Leiden, The Netherlands*
[2] *Netherlands BioInformatics Centre*
[3] *Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands*
[4] *The Netherlands Institute for Systems Biology, University of Amsterdam, Amsterdam, The Netherlands*

Unravelling the molecular mechanisms underlying disease typically involves years of generating new types of data to fill gaps in our understanding. Research has thus become complex and multidisciplinary. An example is Huntington's Disease (HD). The mutation and the gene are known since 1993, but the downstream mechanisms are still poorly understood and no cure has been found. New methods to analyse already existing data could unravel new disease mechanisms such as the role of changes in chromatin modification in HD. The field of e-Science (enhanced Science) aims to enhance data analysis, but also collaboration between disciplines and standards of reproducibility.

We present an approach where the results of software developers support the collaboration between bioinformaticians and biologists, whose aim is to identify and explore a possible role of chromatin modification in HD. We developed knowledge mining Web Services that can be combined in workflows by bioinformaticians, and tools to automatically turn a workflow into a Galaxy (http://galaxy.psu.edu) tool or Web Application.

We present results of analysing the role of CpG islands, an epigenetic marker, in HD by workflows created in Taverna (http://www.taverna.org.uk) and how this supports our multidisciplinary collaboration. Using our approach, we statistically determined the putative role of CpG islands in transcriptional dysregulation in HD. In addition, we could predict and prioritise the most likely candidate proteins that could interact with mutant HTT, to explain the deregulation of genes with a CpG island in their promotor, such as the brain specific DNA binding protein BAIAP2. We used the workflows to discuss experimental design, while the Galaxy and Web Application tools were used to explore the basic functionality of the workflow taking advantage of a user-friendly work environment for our collaborators.

In conclusion, we applied Web Services, Workflows, and worfklow dissemination tools in order to gain insight into the role of CpG islands in HD. It is a powerful and flexible way to implement data analysis. At the same time, it provides a useful representation for internal discussion and scientific discourse. The simplified interfaces of Galaxy and the Web Application are useful to disseminate our work. The workflows will be reused to investigate the role of other epigenetic markers in HD and other (neurological) diseases. Finally, we participate in the Digital Library project 'Workflow Forever' (http://www.wf4ever-project.org) to further improve the preservation and reproducibility of bioinformatics analyses.

38

Transgenerational analysis of DNA methylation inheritance in Arabidopsis thaliana

Lionel Morgado[1], Maria Colomé-Tatché[1], René Wardenaar[1], Sandra Cortijo[2], Ritsert C. Jansen[1], Vincent Colot[2], Frank Johannes[1]

[1] *Groningen Bioinformatics Centre, University of Groningen, Nijenborgh 7, 9747 AG, Groningen, The Netherlands.*
[2] *Institut de Biologie de l'Ecole Normal Supérieure (IBENS), CNRS UMR8197 – INSERM U1024, 46 rue d'Ulm, 75230 Paris Cedex 05, France.*

Inter-individual differences in DNA methylation states (epialleles) can provide a source of heritable phenotypic variation independent of DNA sequence changes. For this mode of epigenetic inheritance to be relevant for evolution or artificial breeding it is necessary to demonstrate the long-term stability of epialleles in controlled settings. Here we employ a genome-wide approach to study the transgenerational behavior of the methylome of a single Arabidopsis Epigenetic Recombinant Inbred Line (EpiRIL) over seven inbreeding generations. We show that a subset of epialleles segregate in a stable Mendelian fashion and that their rate of approach toward fixation is consistent with genetic inbreeding theory. Given this level of stability, we conclude that DNA methylation states can be subject to selection and may constitute an important component of phenotypic evolution.

39

Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Centre
Compendium database of rodent and human metabolic syndrome related datasets

Umesh Nandal[1,3], A.H.C. van Kampen[1,2,3], P.D. Moerland[1,3]

[1] Bioinformatics Laboratory, Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, PO Box 22700, 1100 DE Amsterdam, the Netherlands
[2] Biosystems Data Analysis, Swammerdam Institute for Life Science, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands
[3] Netherlands Bioinformatics Centre, Geert Grooteplein 28, 6525 GA Nijmegen, the Netherlands

Background:
Animal models have been useful for improving our knowledge of molecular interactions underlying human diseases. However, often animal models fail to mimic human disease adequately. One way to validate the similarity of a model organism to its human counterpart is by using gene set enrichment-based methods to compare pairs of gene expression datasets. This requires not only the integration of expression profiles from heterogeneous resources but also a large number of pairwise comparisons between human and animal model gene expression datasets. Moreover, narrowing down gene expression datasets for a specific domain of study from public repositories such as GEO can be challenging.

Results:
We developed a MySQL database that provides a homogeneous and flexible platform for the systematic collection, storage and retrieval of gene expression data from heterogeneous resources. The download of gene expression datasets from GEO to the database as well as querying the database can be performed using dedicated R-functions. The database and the R-functions have been bundled together into an R package named geoDB. Differential expression and cross-species gene expression analysis using enrichment statistics can be performed using an automated pipeline provided with the geoDB package. Molecular similarity between rodent models and human is investigated by testing for enrichment of a signature from one species in a rank-ordered gene list from a dataset in the other species. The enrichment results are represented in the form of heatmaps and bipartite graphs.

Conclusion:
The geoDB R package provides a flexible platform for loading gene expression data from heterogeneous resources in a local database and performing various types of analyses. The integrated expression data will be used to validate animal models for complex multifactorial human diseases.

40

Small Multiples View Cytoscape plugin for comparing gene expression in a network context

Harm Nijveen[1,2], Job Geerligs[1], Jack A.M. Leunissen[1,2]

[1] *Laboratory of Bioinformatics, Wageningen University,*
[2] *Netherlands Consortium for Systems Biology*

Comparing the expression levels of sets of genes between different conditions or a series of time points is a powerful approach to identify the important factors that play a role in a biological process of interest. The heat map is an often-used visualization method to analyze these expression levels, by using a matrix where each cell represents a gene in a certain condition or time point and the color of the cell shows the corresponding expression level. A shortcoming of heat maps is the lack of additional information on the displayed entities: the matrix structure of a heat map dictates a simple one dimensional organization of the genes, whereas a two dimensional network structure would provide more information about the relationship of the different genes.

We are developing a plugin for the popular network visualization tool Cytoscape that helps the biologist to visualize the expressions levels of genes or abundance of metabolites in a network context (gene regulation network, metabolic pathway, etc.) and compare these levels between different conditions or time points. The same network is repeated in a grid of images, one for each individual condition, with the expression levels for the different genes mapped onto the network per condition using a heat map-like coloring for the nodes. This visualization strategy that was popularized by Edward Tufte is known as “Small Multiples”. The additional information of the network structure can help to better interpret the results of an experiment or it could be used to validate and perhaps improve the used network.

41

e-BioGrid: Building a Dutch e-science infrastructure for life science research

Irene M. Nooren [a,b,d], Han Rauwerda [a,d], Machiel Jansen [c], Jan J. Bot [b,d,e], Joost N. Kok [b,d] and Timo M. Breit [a,d]

[a] *Microarray Department & Integrative Bioinformatics Units, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands*
[b] *Leiden Institute of Advanced Computer Science, Leiden University, Niels Bohrweg 1, 2333 CN, Leiden, The Netherlands*
[c] *SARA, Science Park 140, 1098 XG Amsterdam, The Netherlands*
[d] *Netherlands Bioinformatics Centre (NBIC), Geert Grooteplein 28, 6525 GA Nijmegen, The Netherlands*
[e] *Delft Bioinformatics Laboratory, Delft University of Technology, Mekelweg 4, 2628CD, Delft*

Life science and health research today often involves data-intensive experimentation. However, local ICT infrastructure is increasingly insufficient to deal with the associated growing computational demands. This poses a multi-disciplinary challenge in which e-science methods, tools and resources need to be combined into a functional e-bioscience infrastructure for life science research. To meet this challenge in the Netherlands, an e-BioGrid initiative was started within the national BiG Grid program. It consists of a multidisciplinary central e-Core plus an e-science support team and it operates through selected large project with decentralized staff or through continuously submitted ad-hoc projects. The e-BioGrid projects are organized via seven Technology Areas that comprise most life science technologies and align the initiative with ongoing NBIC and bioinformatics programs. Here an overview will be presented of the e-BioGrid initiative and progress in the infrastructure developments. The e-BioGrid initiative so far has led to a growing e-bioscience community as well as life sciences e-research user community. Also, there are some valuable lessons learned and several data analysis applications are configured on the BiG Grid infrastructure using either cluster, grid or cloud computing.

42

TraIT: an integrated translational research IT platform

J.H. Obbink, Jan-Willem Boiten, Jeroen A.M. Beliën, Gerrit A. Meijer

Philips Research, Centre for Translational Molecular Medicine, VU University Medical Center

TraIT (Translational Research IT) aims to establish a significant improvement of the IT infrastructure for translational research in the Netherlands, and to share and secure the data generated in the translational research projects sponsored by CTMM. Ultimately this project is meant to be the basis of an IT infrastructure that facilitates translational research logistics, data management, data integration, and data analysis at a national level, as well as being the Dutch hub in international networks. The project aims to adopt and adapt existing point solutions rather than embarking on major software development projects. The main challenge will be to unite the current isolated solutions into one interoperable IT platform for translational research.

It is TraIT's challenge specifically to:

- Provide a sustainable and scalable translational research information infrastructure that allows industry and public organizations to share and disseminate data, analyses and understanding. This should be driven by community-endorsed open data and technical standards.
- Provide accessible, high-quality powerful tools to allow scientists to interact and explore a unified translational medicine research space, i.e. the TraIT tools should be interoperable with other systems, but above all also intra-operability between the individual TraIT work packages.
- Provide accessible, high-quality powerful tools to allow translational researchers to manage the business processes of translational research.
- Build as much as possible upon existing solutions and develop for unmet needs only.
- Do this in a user-oriented and process-driven fashion.

All key organizations within translational research in The Netherlands are committed to the TraIT project. This community building aspect of the project is crucial for its eventual success: implementation of a common data sharing platform for translational research integrated across all relevant data generating disciplines (clinical, imaging, biobanking, and "omics-type" experimental data).

The TraIT project is a joint initiative between CTMM, KWF Dutch Cancer Society, Dutch Heart Foundation, Netherlands eScience Center, the Netherlands Federation of University Medical Centers (NFU), the Netherlands Bioinformatics Centre (NBIC), and the Parelsnoer Initiative (PSI).

43

Are REPs genetic insulators that enable differential regulation of gene expression in bacteria?

Lex Overmars[1,2,4], Roland Siezen[1,2,4] and Christof Francke[1,2,3,4]

[1] *Radboud University Medical Centre, Centre for Molecular and Biomolecular Informatics, Nijmegen*
[2] *TI Food and Nutrition, PO Box 557, 6700AN Wageningen*
[3] *Kluyver Centre for Genomics of Industrial Fermentation, PO Box 5057, 2600GA Delft*
[4] *Netherlands Bioinformatics Centre, 260 NBIC, PO Box 9101, 6500HB Nijmegen*

Background: Repetitive Extragenic Palindromic elements (REPs) are short palindromic sequences that were first identified in E.coli and closely related enteric bacteria. Recently, REPs were identified in more diverse bacterial taxa. REPs exhibit some remarkable characteristics. They (i) are almost exclusively found in the intergenic space, in which they are often arranged in repeats called BIMEs; (ii) occur in high abundance, up to 500-1000 REPs in some species, occupying a substantial portion of the intergenic space; (iii) are highly conserved within a genome
Various functional links have been proposed in literature. For instance, it was shown that REPs play a role in stabilizing mRNA, i.e. that REP-containing transcripts are less prone to degradation. Specific REPs were also shown to act as binding sequence for either DNA polymerase or DNA gyrase. However none of the related literature provides a common functional denominator for the complete set found, let alone a satisfying mechanism of action. We therefore decided to investigate the commonality using a comparative genomics approach.

Results: E.coli REPs were identified using a re-defined motif based on a conserved 29bp sequence. We observed a biased distribution of REPs with respect to the ORFs: REPs are not found between divergent gene-pairs and predominantly located between convergent gene-pairs.
A set of 465 publicly available microarrays (Many Microbe Microarrays database) was used to explore the effects of REPS on transcription under various conditions. This analysis revealed an association between REP-related gene-pairs and higher expression levels. This association is also evident when Codon Adaptation Index (CAI) values were compared. We identified a set of microarrays with significant effects on gene-REP-gene pair (co-) expression. These arrays all represented the transcriptional response to a certain kind of stress, i.e. related to biofilm formation, aerobiosis and stress specific sigma factors. In addition, we found that REP-containing operons frequently posses an alternative promoter which is known to modulate the stress responses.

Conclusions: This study shows that REPs potentially have a global role in regulation of differential expression. Our results imply that REPs enable differential expression specifically in cases were transcription-driven elevated pressure on DNA supercoiling can arise, i.e. expression of convergent gene-pairs and transcription regulated by an alternative promoter. Our findings imply that the phenomenon of REP-enabled differential expression is linked to the bacterial stress response in E. coli.

44

Pep2Path: Enabling high-throughput antibiotic discovery by automatic identification of peptide biosynthesis pathways

Yared Paalvast[1], Marnix H. Medema [1,2], Eriko Takano1 & Rainer Breitling [2,3]

[1] *Department of Microbial Physiology, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Nijenborgh 7, 9747 AG Groningen, The Netherlands.*
[2] *Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Nijenborgh 7, 9747 AG Groningen, The Netherlands.*
[3] *Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, Joseph Black Building, University of Glasgow, Glasgow G12 8QQ, UK.*

Nonribosomal peptides are secondary metabolites produced by nonribosomal peptide synthetases (NRPSs). Since nonribosomal peptides often show antibiotic or antitumor activity, they are much investigated in the hunt for new therapeutic compounds[1]. To facilitate this research, the NORINE database stores information on known nonribosomal peptides and their NRPSs [2]. For a substantial proportion of these peptides, the producing NRPS is unknown. Identifying the NRPSs of these non-ribosomal peptides could reveal new targets for synthetic biology peptide engineering [3, 4].
The aim of this study was to find NRPSs for nonribosomal peptides without a known synthetase. We developed the Pep2Path software that identifies putative NRPS gene clusters in genome sequences, predicts the encoded nonribosomal peptide, and finds the best matching peptide in the NORINE database, linking "orphan" peptides to their putative biosynthetic clusters. The underlying code is built on antiSMASH [5] and NRPSPredictor2 [6], and uses both a support vector machine classifier and a specificity code of amino acid sequences lining the NRPS adenylation domain active site to predict the amino acids incorporated in a nonribosomal peptide [6]. We show that Pep2Path is computationally very efficient and rapidly identifies the correct NRPS-to-peptide links for the majority of known nonribosomal peptides.

In the future, the Pep2Path software can form the basis for accelerated high-throughput discovery of novel antibiotics by combining genome sequencing and chemical analysis of nonribosomal peptides from a multitude of bacterial strains, through the automated linking of partial peptide sequences inferred from peptide mass spectra to genomically encoded NRPS gene clusters.

References

1. Fischbach MA, Walsh CT. (2006) Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: Logic, machinery, and mechanisms. Chem Rev 106(8): 3468-3496.
2. Caboche S, Pupin M, Leclère V, Fontaine A, Jacques P, et al. (2008) NORINE: A database of nonribosomal peptides. Nucleic Acids Res 36(suppl 1): D326-D331.
3. Medema MH, Breitling R, Bovenberg R, Takano E. (2010) Exploiting plug-and-play synthetic biology for drug discovery and production in microorganisms. Nature Reviews Microbiology 9(2): 131-137.
4. Medema MH, van Raaphorst R, Takano E, Breitling R. (2012) Computational tools for the synthetic design of biochemical pathways. Nature Reviews Microbiology .
5. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, et al. (2011) antiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res 39(suppl 2): W339.
6. Röttig M, Medema MH, Blin K, Weber T, Rausch C, et al. (2011) NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. Nucleic Acids Res 39(suppl 2): W362.

45

Towards (almost) closed genomes with SSPACE and GapFiller

Marten Boetzer and Walter Pirovano

BaseClear

De novo assembly is a commonly used application of next generation sequencing (NGS) experiments. The ultimate goal is to puzzle millions of reads into one complete genome, although in practice this is still a challenging task. In particular the presence of repeated elements and low-coverage regions are difficult hurdles to take. Where draft assemblies usually result in a number of contiguous sequences (contigs), additional paired-end and matepair sequencing experiments can help to jump over difficult areas and place contigs into larger stretches called scaffolds. A final step in the assembly is to manually close the gaps between and within these scaffolds using standard Sanger sequencing. However this is an expensive procedure, especially for larger genomes.

Here we introduce an automated strategy, called GapFiller, to accurately close gaps within scaffolds using paired-reads. The method is a true novelty in the field and shows good results on both bacterial and eukaryotic genomes. Moreover on a human dataset we underscore that GapFiller can be helpful in reconstructing chromosomic regions that carry important functional information. We demonstrate the GapFiller method, in combination with our widely-used SSPACE scaffolding algorithm, provides an excellent strategy to automatically finish draft genomes. As a consequence the amount of additional wetlab work needed to close a genome is drastically reduced. Also we will discuss the importance of these methods in the light of third generation sequencing strategies such as the IonTorrent and the Pacific Biosciences systems. The software is freely available for non-commercial users at www.baseclear.com/bioinformatics-tools/.

46

NBIC as IDP - what does it mean for you?

Gera Pronk, SURFnet

In the past Months SURFnet and NBIC worked together in a SURFconext pilotproject to connect NBIC as Identity Provider (IDP). March 2012, NBIC was connected technically as IDP. Still legal and organisational steps are needed because NBIC is the first 'Virtual Organisation' (VO) that is connected as IDP in the Dutch community and this is 'off the beaten path'. The way we will treat IDP's may have implications for future agreements between SURFnet and other IDP's and Service Providers.

In this presentation we will tell about the new possibilities for NBIC as a Identity Provider but we also want to ask the audience to reflect at the challenges we encounter.

What does it mean for Dutch bioinformatics that NBIC is an IDP? In short term and in nearer future?
What does the NBIC pilotproject mean for other VO's and the Dutch research community?
Which scenarios are possible?

47

Stochastic transitions in oncogenic transformation

Anas Rana[1,2], Jon Armond[1], Mario Nicodemi[3], Sach Mukherjee[2,4]

[1] *Centre for Complexity Science, University of Warwick, Coventry, UK*
[2] *Netherlands Cancer Institute, Amsterdam, The Netherlands*
[3] *Dip.to di Scienze Fisiche, Univ. di Napoli "Federico II", INFN Napoli, Italy*
[4] *Department of Statistics, University of Warwick, Coventry, UK*

In many biological processes time-varying transcriptional changes are related to underlying changes in cell state. Examples of state transitions abound in development, differentiation and disease. For example, in tumorigenesis initially normal cells acquire a transformed phenotype following modulation of one or more oncogenes or tumour suppressor genes. At any given time during a transition process, the cell population may be heterogenous, since in general cells will not change state synchronously. However, mainstream genome-wide assays yield only population average data, making it difficult to elucidate state-specific biology. On the other hand, genome-wide single cell assays remain challenging. Here, we present a class of stochastic models that allow identification of state transition dynamics and state-specific expression signatures from population-average time course data. We use these approaches to investigate oncogenic transformation using data from an in vitro mammary epithelial cell system containing an inducible Src oncogene (Hirsch et. al. '10). We identify an intermediate, pre-oncogenic state as well as gene lists that are specific to states and that may shed light on events leading to acquisition of the transformed phenotype.

48

Bioinformatics and Modeling at DSM

Christian Rausch, Marco de Groot, Herman Pel, Hans Roubos

DSM Biotechnology Center, Delft/NL

Royal DSM is a global science-based company active in health, nutrition and materials. By connecting its unique competences in Life Sciences and Materials Sciences DSM is driving economic prosperity, environmental progress and social advances to create sustainable value for all stakeholders. DSM delivers innovative solutions that nourish, protect and improve performance in global markets such as food and dietary supplements, personal care, feed, pharmaceuticals, medical devices, automotive, paints, electrical and electronics, life protection, alternative energy and bio-based materials. DSM's 22,000 employees deliver annual net sales of around €9 billion. The company is listed on NYSE Euronext.

DSM's Bioinformatics research is mainly based at the DSM Biotechnology Center, Delft, The Netherlands. Technologies and expertise we distinguish are: DNA design, protein engineering, metabolic pathway engineering, DSM's global bioinformatics platform, omics data analysis, (re)sequencing and biostatistics. These are applied to classical and rational strain engineering to make metabolites and enzymes for applications in food, feed, pharma and bio-based products. On a regular basis we hire new employees and have MSc and PhD visiting scientists. The poster will present recent examples of bioinformatics and modeling applications at DSM.

49

Biologically relevant constraints for analysis of high-throughput data.

P. Reshetova[1], J.A. Westerhuis[1], A.H.C. van Kampen2, A.K. Smilde[1]

[1] *Biosystems Data Analysis Group, Swammerdam Institute for Life Sciences, University of Amsterdam;*
[2] *Bioinformatics Laboratory, Academic Medical Center, Amsterdam*

The latest high throughput technologies in systems biology such as microarrays in transcriptomics and mass spectrometry in metabolomics give a unique opportunity to catch simultaneously a huge range of biochemical processes happening in a tested system. As a result the system is presented by means of expression of thousands of genes or concentration levels of hundreds of metabolites. Statistical analysis of such high throughput data requires powerful data analysis tools that are able to uncover particular processes as well as the picture in a whole. Due to the small number of experiments many these tools easily overfit the data. The default solution to this problem is the use of mathematical constraints, however this often leads to models that are hardly interpretable or biologically irrelevant.

To improve the biological relevance of empirical models of high throughput data, new approaches have appeared in which biological information about the system is used. In this work we focus on the biological relationship between the variables in the data being genes, proteins or metabolites. The relationships can be defined by functional groups as e.g. defined by Gene Ontology, interaction networks or metabolic pathways. Various methodologies have appeared in which these relationships are used for biological validation or for improved interpretation (e. a. the gene set enrichment or metabolite set enrichment methods. Other methods use such biological prior information for improved classification of groups. In this work we give an overview of such methods and discuss their basic properties.

50

Immunolabeling artifacts and the need for live-cell imaging

Ulrike Schnell, Freark Dijk, Klaas A Sjollema & Ben N G Giepmans

Dept. of Cell Biology, University Medical Center Groningen, A. Deusinglaan 1, 9713 AV Groningen, The Netherlands

Protein identification in situ is highly important to understand the regulation of cellular processes. Although the use of genetically-encoded fluorescent proteins has revolutionized the examination of proteins in living cells, studies using these proteins still are met with criticism because proteins are modified and ectopically expressed. In contrast to fluorescent proteins, immunocytochemistry of endogenous proteins is believed to provide more reliable localization information. However, introducing immunoreagents inside cells can cause protein extraction or relocalization, not reflecting the in vivo situation. Such artifacts are often underappreciated and reports have received little attention.

Aiming to increase the awareness of artifacts induced by standard immunolabeling protocols, we compared the effects of different fixation and permeabilization conditions on the localization of GFP-fusion proteins in different subcellular compartments. Besides fluorescent (live-cell) imaging, we used electron microscopy to reveal the cellular ultrastructure after fixation and permeabilization. We find that optimizing the immunostaining procedure for each protein of interest and for each cell line is highly important to prevent misinterpretation of results, and that different immunofluorescent staining methods should be used in parallel. We conclude that immunostaining experiments in dead, permeabilized cells should be complemented with live-cell imaging, such that implementation of fluorescent protein–tagged chimeras becomes a standard routine when scrutinizing protein localization in situ.

Reference:
Schnell U, Dijk F, Sjollema KA & Giepmans BNG. Immunolabeling artifacts and the need for live-cell imaging. Nat Methods. 2012 Jan 30;9(2):152-8

URL: http://f1000.com/13988973

51

Integrating different data sources for orthology prediction

Edouard Severing

Wageningen University

Accurate orthology inference is essential for protein function prediction and phylogeny reconstruction. Several algorithms exist for detecting orthologous proteins in two or more species. These algorithms typically only use protein sequence information for inferring clusters of orthologous proteins.

However, there are other properties of proteins which could be used for predicting orthology. Such properties include gene-structure information, gene-order information and co-expression patterns. The aim of this project is to develop a method for orthology prediction that enables the integration of different types of data. We are currently investigating how we can derive a similarity score for any pair of proteins using multiple sources of information.

As we are mainly interested in identifying orthologs in plant species we are developing our method using protein sequence, gene-structure, gene-order and co-expression data from Arabidopsis, rice and tomato. However, the end product of this project will be a set of tools/scripts for orthology prediction between any set of species for which sufficient genomics data is available.

52

Developing knowledge based approach to Word Sense Disambiguation in biomedical text.

Bharat Singh, Erik M. van Mulligen, Jan Kors, Kang Ning

Department of medical informatics, Erasmus MC

The biomedical scientific literature is now so large that automated tools are necessary to access it effectively (Chapman and Cohen, 2009). Automated methods for cataloging, searching and navigating these documents would be of great benefit to researchers working in this area, as well as having potential benefits to medicine and other branches of science. However, this process is made difficult by the fact that some words have multiple senses. Word Sense Disambiguation (WSD) systems aim to solve this problem by identifying the meanings of ambiguous words in a specific context (Agirre and Edmonds, 2006; Navigli, 2009). A word is ambiguous when it has more than one sense, which is determined based on the context in which the word is used. For example, the word discharge could refer to patient discharge or a secretion like a body fluid discharge.

The ability to accurately identify the meanings of terms is an important step in automatic text processing. It is necessary for applications such as information extraction and text mining that is important in the biomedical domain for tasks such as automated knowledge discovery. Weeber et al. analyzed MEDLINE abstracts and found that 11.7% of phrases were ambiguous relative to the UMLS Metathesaurus. The NLM Indexing Initiative attempted to automatically index biomedical journals with concepts from the UMLS Metathesaurus and concluded that lexical ambiguity was the biggest challenge in the automation of the indexing process. Friedman et all reported that an information extraction system originally designed to process radiology reports had problems with ambiguity when it was applied to more general biomedical texts.

In this paper we describe knowledge based structural approach to WSD in the biomedical domain that analyzes and exploits the structure of an available concept network. The approach is unsupervised; it does not require any expensive labeled training data and relies on information derived from the UMLS Metathesaurus and semantic network instead. The UMLS Metathesaurus and semantic network is converted into a graph that can be traversed to calculate semantic similarity of word senses. The Structural Semantic Interconnection (Navigli and Velardi 2005) algorithm is applied to this graph to carry out WSD. An advantage of this algorithm is that it provides a justification (i.e. a set of semantic patterns) to support a sense choice.

We evaluated our results on the NLM-WSD dataset and our method outperforms other knowledge-based methods that rely on the UMLS Metathesaurus and semantic network alone.

53

Reproducibility of parameter learning in a Bayesian network for predicting WNT pathway activation

Shriprakash Sinha[1], Marcel J.T. Reinders[1], Wim Verhaegh[2]

[1] *Delft University of Technology*
[2] *Philips*

Background: Bayesian network (Bnets) being a collection of probabilistic classifiers or regressors constrained by conditional relationships [Heckerman:1998], serve as useful models for hypothesis testing when data is missing or certain prior causal relations need to be incorporated. Adopting a primitive structure of the Wnt signaling pathway from [Clevers:2006], the Wnt Pathway being active or inactive can be tested by inferencing whether the transcription complex (TrCmplx) is active or inactive. This is done by capturing the expression values of the Wnt target genes using a reference Bnet from [Verhaegh:2011]. The network has three main layers with nodes denoting (1) ß-catenin, Tcf4 and TrCmplx (ß-catenin and Tcf4 are parents of TrCmplx), (2) Wnt target genes as children of TrCmplx and (3) probesets as children of individual genes. The network is trained on colon cancer cell lines with measurements of expression levels for a particular set of Wnt target genes. Inferencing the state of the TrCmplx based on the evidence provided by expression levels of the same Wnt target genes for an unlabeled sample leads to prediction of the state of Wnt Pathway. Perfect prediction was obtained on normal colon and colon adenoma samples from GSE8671 [Sabates:2007].

Method: This work tests the reproducibility capacity of a Bayesian network for parameter learning and predictions based on the estimated parameters. This is done by comparing (i) the estimated parameters with initially assigned parameters of the reference Bnet and (ii) prediction results obtained using learned parameters with prediction results obtained using the assigned parameters of the reference Bnet. The parameter estimation and predictions generated using these parameters, is done using simulations of complete and missing data, sampled from the reference Bnet.

Results: Estimated parameters from simulations on complete observations sampled from the reference Bnet matched to the initial parameters that described the reference Bnet with limited variation. In case of missing observations, some flipping of parameter values for genes where found. This can be attributed to the Bnet's incapability to interprete the meaning of a node's state, leading to an equivalent swapped solution. Also, near chance parameter values of the probesets can cause the Bnet to yield different parameter values for genes, compared to the reference Bnet. Dispite theses anomalies, prediction on normal colon and colon adenoma samples (GSE8671), showed reproducibility of results while hiding (1) complexes, (2) genes and (3) both genes and complexes.

Conclusion: Learning parameter values apropos to observations sampled from a reference Bnet with missing data, indicate that Bnet is effective in reproducing prediction results for the Wnt pathway activation.

References:

[Clevers:2006] H. Clevers. Wnt/ß-catenin signaling in development and disease. Cell, 127(3):469–480, 2006.

[Heckerman:1998] D. Heckerman et al. A tutorial on learning with bayesian networks. Nato Asi Series D Behavioural And Social Sciences, 89:301–354, 1998.

[Sabates:2007] J. Sabates-Bellver, L.G. Van der Flier, M. de Palo, E. Cattaneo, C. Maake, H. Rehrauer, E. Laczko, M.A. Kurowski, J.M. Bujnicki, M. Menigatti, et al. Transcriptome profile of human colorectal adenomas. Molecular Cancer Re- search, 5(12):1263, 2007.

[Verhaegh:2011] W. Verhaegh, P. Hatzis, H. Clevers, and A. van de Stolpe. Personalized cancer treatment selection using computational signaling pathway models. Can- cer Research (San Antonio Breast Cancer Symposium), 71(S24):S524–S525, 2011.

54

A course with HOPE

M. Snelleman, G. Vriend

CMBI, UMC St Radboud, Nijmegen

Many bioinformatics tools have been developed over the last decennium that most life scientists do not know how to operate. Sometimes the interface is too non-intuitive, the scientist has to set parameters he or she has never heard of or the output is a meaningless number.

The web-server HOPE (Have (y)Our Protein Explained) was developed for life scientists with little experience in bioinformatics. HOPE predicts the effect of a mutation on a protein's structure and function and reports its conclusions as life-scientist-understandable text.

We developed three types of courses for different sets of audience to explain the basics of bioinformatics and teach him or her how HOPE gets its conclusions. The courses will not deal with the algorithms but rather explain the biological motivation for HOPE's conclusions using multimedia in hands-on exercises and real-life cases. The exercises and videos of the lectures will be available online, which enables every potential HOPE user to absorb the course.

At the end of each course the life scientist should be able to understand how HOPE gets its information, how it uses this information and he or she will also be able to look critically at the HOPE report.

55

Network Inference in Hidden Markov Models for Modelling Chromatin

Nicolas Stadler, Sach Mukherjee

Netherlands Cancer Institute

Chromatin is DNA and all associated proteins that make up the nucleus of a eukaryotic cell. Primary functions of chromatin are: to package DNA into a smaller volume to fit in the cell, to strengthen the DNA to allow mitosis and meiosis and prevent DNA damage, and to control gene expression and DNA replication.

Very roughly speaking the organization of the chromatin can be divided into hetero- and euchromatin. Heterochromatin is associated with structural proteins, which pack the DNA very compactly into a 30 nm fiber consisting of nucleosome arrays. Genes located in heterochromatin are transcriptionally inactive. On the other hand chromatin stands for loosely compacted DNA, wrapped around histones, which form the 10nm "beads-on-a-string" fiber.

There is strong evidence that DNA and its hundreds of chromatin proteins have a much finer organization than simply eu- and heterochromatin. We explore the organization of chromatin by segmenting genome-wide protein binding data using Hidden Markov Models. Our key point is that not only different binding levels of proteins characterize specific chromatin regions, but also the interplay of proteins within such regions defines different types of chromatin. For example, in order to transcribe a specific gene not only the presence/absence of a combination of proteins is important but also the interaction of such proteins can be decisive.

56

Detection of Genetic Disorders in Fetuses by Next Generation Sequencing of Maternal Blood Samples

Roy Straver

Delft University of Technology

Until now testing for fetal genetic disorders was dependent on invasive sampling. It has been shown that small bits of fetal DNA can be found in maternal blood, and together these bits make up for the complete fetal genome. Recent studies have shown it is possible to determine fetal trisomy 21 (Down Syndrome) using this DNA information.

This study attempts to determine smaller genetic disorders by sequencing a maternal blood sample. After obtaining and mapping the reads from the samples, the data is filtered to remove most of the peaks caused by sequencing and reference genome errors by deleting large read stacks.

The data used for this study is obtained by the Illumina/Solexa HiSeq2000 an consists of numerous samples of healthy fetuses, numerous trisomy cases and numerous cases with smaller disorders. If the results of this study show that it is possible to determine fetal genetic disorders, this non-invasive method can be used for any pregnancy. This allows any couple to know whether their baby will be healthy, without risking miscarriages and possibly removing the pre-test screening that allows only the highest risk fetuses to get tested.

57

FluxEs 2.0: An R Package for the Quantification of Metabolic Fluxes in Mammalian Cell Lines

Hilal Taymaz-Nikerel[1,2,3,4,*], Hannes Hettling[2,3,4,*] , Johannes H.G.M. van Beek[1,2,3,4]

[1] *Section Medical Genomics, VU University Medical Centre, Amsterdam, The Netherlands*
[2] *Centre for Integrative Bioinformatics, VU University, Amsterdam, The Netherlands*
[3] *Netherlands Consortium for Systems Biology*
[4] *Netherlands Bioinformatics Centre*
[*] *Equal contribution of authors*

Quantification of metabolic fluxes, an important tool for characterization of cellular metabolism, gives insight on broad range of topics, from developing new metabolic engineering strategies to understanding the mechanisms of diseases. 13C metabolic flux analysis (MFA), which makes use of labeled isotopes, has been shown to be an efficient method to quantify fluxes in microbial and mammalian systems. This method combines use of labeled stable isotopes with computational models and analytical methods such as mass spectrometry (MS) or nuclear magnetic resonance (NMR) spectroscopy. For this purpose, in our group an open source computational package (FluxEs, "written in the R language") was developed [1] in which metabolic systems can be assembled, simulated and analyzed in order to calculate fluxes from dynamic 13C-data. Despite the successful application to estimate TCA cycle flux in animal biopsies from a single time point in metabolic steady state and isotopic non-steady-state data, FluxEs has some limitations, such as incorporation of reversible reactions and bimolecular reactions. Presence of these features in the program is important to fully consider biological complexity. Simulation of the experiment with the computational model implemented in our package allows designing an experiment with an optimal tracer enrichment in substrates given to the cell culture. This makes it possible to optimize the measurement with the goal of quantifying the maximum number of fluxes in the network with great accuracy. The package assembles the mathematical model equations automatically based on text files containing the biochemical information. We present the newer version of FluxEs that overcomes the above mentioned limitations. As an example, a metabolic network of glycolysis + TCA cycle + anaplerotic reactions in mammalian cells, including all known reversible reactions, was implemented and simulated for a variety of different labeled glucose carbon tracers given to the cell culture in silico.

[1] Binsl et al, 2010, Bioinformatics 26(5), 653-660.

58

Nanopublications

M.Thompson and E. Schultes

Leiden University Medical Center

Recently, we have proposed a new mode of data publishing called nanopublications (nanopub.org). A nanopublication is the smallest unit of publishable information, and has the form of a semantic assertion (Subject-Predicate-Object) plus provenance metadata (e.g., author, time-stamp). When appropriately serialized in RDF, nanopublications can be machine readable, interoperable, and retrieved using semantic web query methods. Nanopublications are perfectly suited to expose high-throughput, heterogeneous genomics data ensuring their rapid dissemination, universal interoperability, and long-term persistence. Because nanopublications permit any data to be attributed to its authors, to institutions and to specific projects their scientific impact can be automatically tracked, creating powerful incentives for the exposure of both legacy and forthcoming data in nanopublication format. Here, we will describe our efforts in participation with the Open PHACTS consortium (http://openphacts.org) to create example nanopublications for life science data.

www.nanopub.org

59

How can a disordered protein bind specifically multiple partners?

Busra Topal, Attila Gursoy, Ozlem Keskin

Koc University

Proteins interact through their interfaces. We can define interfaces into 3 types according to structure of the interfaces and the global structures of the interacting proteins. Members of 'Type 1' have similar interfaces, global folds and functions. In 'Type 2', members often do not share similar functions and do not have globally similar structures, however they have similar interfaces. Members of a 'Type 3' cluster have similar binding sites on one side of the interface, but the partner proteins are different. In this work, we will try to define an additional type, 'Type 4' for interfaces of disordered proteins. Disordered regions are defined as entire proteins or regions of proteins that lack a defined tertiary structure. Proteins with disordered sequences cannot bury sufficient hydrophobic core to fold spontaneously into the highly organized 3D structures. When disordered proteins bind one of their pre-existing conformations becomes much more populated. This happens because upon binding, this conformation becomes stabilized. One of the aims of this project is finding characteristics of disordered proteins like hydrophobicity and compactness for define 'Type 4'. Other and most important aim of the project is finding how a disordered protein can specifically bind multiple partners (hub proteins). We are using PRISM template set as a dataset. We found fractions of disordered regions of proteins in our dataset. Majority of proteins in dataset have disordered regions.

60

Two-photon microscopy and Spectral phasor to monitor NADH (free/bound) and FAD in normal and cancer cells in 3D culture

Truong H., Knaus H., Bader A., Fereidouni F., Blab G.A., and Gerritsen H.C.

Utrecht University/Molecular Biophysics

Nicotinamide Adenine Dinucleotide (NADH) and Flavin Adenine Dinucleotide (FAD) are important coenzymes that play important role in regulating metabolism. It has been described by various groups that auto-fluorescence property of NADH and FAD can be exploited to study cancer progression and cell death (Skala MC et al, 2007;Wang HW et al, 2008) via two-Photon Fluorescence Lifetime imaging (FLIM) system. In the general direction, we developed a unique approach of imaging and measuring NADH (free in cytoplasm and bound to enzyme) and FAD in 3D culture by combining two-photon microscopy with anisotropy and spectral phasor (representing the spectra by polar representation). Our goal is to exploit this technique to characterize cancer and non-cancer cellular spheroids (cell aggregates) by un-mixing and quantifying the intrinsic fluorphores based on their emission spectrum. Furthermore, we can employ NADH as possible apoptosis biomarker for 3D culture drug screening and invivo imaging of tumor model. This is more feasible approach than standard histopathological assessment which require fixation which likely alter fluorescence lifetimes compared with unfixed tissue.

61

An enzyme redesign method for improved production rates in Aspergillus niger

B.A. van den Berg[1,2,3], M.J.T. Reinders[1,2,3], H.J. Pel[4], J.A. Roubos[4], D. de Ridder[1,2,3]

[1] The Delft Bioinformatics Lab, Faculty of Electrical Engineering, Mathematics & Computer Science, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands
[2] Netherlands Bioinformatics Centre, Nijmegen, The Netherlands
[3] Kluyver Centre for Genomics of Industrial Fermentation, Delft, The Netherlands
[4] DSM Biotechnology Center, Delft, The Netherlands

High yields are required for industrial production of enzymes. Previous work showed that in the microbial cell-factory Aspergillus niger a protein's amino acid composition is predictive for high-level production. To improve production rates of enzymes for which we did not observe high-level production, we propose a design method that increases resemblance to proteins for which high-level production was observed. With all functional and buried residues untouched, mutations to amino acids that are often observed on the same position in homologous proteins are applied to obtain sequences with an amino acid composition that better matches that of structurally similar, but high-level secreted proteins. Experimental work will be done to test if the rational design method results in improved production rates.

62

Dynamic Transcriptomics of Helper T Cells

Henk-Jan van den Ham [1,3], Leon de Waal[1], Fatiha Zaaraoui[1], Maarten Bijl[1], Wilfred F. van IJcken[2], Albert D.M.E. Osterhaus[1], Rob J. de Boer[3] & Arno C. Andeweg[1]

[1] *Dep. of Virology*
[2] *Erasmus Center for Biomics, Erasmus MC, Rotterdam, The Netherlands*
[3] *Theoretical Biology & Bioinformatics, Utrecht University, The Netherlands*

Helper T cells form a major pillar of the adaptive immune system. After stimulation with their cognate antigen upon infection, helper T cells produce specific cytokines that play a crucial role in determining the phenotype of the immune response that is raised. Classically, the Th1 phenotype is associated with cellular immunity, while Th2 is associated with humoural immunity and antibody responses. Profiling of the T helper cell response is complicated by the fact that expression of key cytokines is thought to be dependent on the time since activation. The Th1 cytokine IFNgamma is typically observed to be an early cytokine, while the Th2 cytokine IL4 tends to be measured at later time points and/or after a number of cell divisions. In this study, we perform time-course experiments in order to obtain a comprehensive view of helper T cell differentiation.

We performed microarray assisted mRNA profiling on antigen stimulated, TCR transgenic murine splenocytes that were cultured in the presence of polarising cytokines. Transcriptome snapshots of Th cells differentiating into Th1 and Th2 phenotypes were obtained.

Multiple immune regulatory pathways are differentially expressed, including pathways specific to Th differentiation. Principle component analysis shows that: 1) activation, 2) time since activation, and 3) Th skewing are the largest sources of variance in our profiling experiments. We show that divergence between the Th1 and Th2 phenotypes does not increase in terms of number of differential genes from day 1 to day 4 after stimulation, indicating that divergence between these two phenotypes takes place within 24h post-activation. Applying a recently developed polar score data analysis method, enabled us to identify novel clusters of skewed genes associated with Th1 and Th2 differentiation. Among these Th1 and Th2 specific genes are a number of key players previously associated with other T lymphocyte phenotypes or the maturation of other lymphoid cell types.

We show that T cell activation, time, and skewing are the largest sources of variance in our dataset. In terms of differential mRNA expression, Th1 and Th2 cells diverge within 24 hours post activation. Members of the Batf transcription factor family are expressed in different helper phenotypes, suggesting an important role for this family in helper T cell phenotype differentiation.

63

Determining phylogenetic relationships of fish owls using next-generation sequencing

Jaap W.F. van der Heijden[1], Seyed Yahya Anvar[1,2], Ken Kraaijeveld[1,2], Johan T. den Dunnen[1,2], Jeroen F.J. Laros[1,2]

[1] *Center for Human and Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands*
[2] *Leiden Genome Technology Center, Leiden University Medical Center, Leiden, The Netherlands*

Molecular phylogenetic studies of rare or complex organisms are far from trivial. Due to the absence of reference sequence and potential contamination and DNA damage, conventional methods cannot be used to gain insights on evolutionary relationships of such organisms. These approaches heavily rely on the detection of variants and protein encoding proteins as the basis for classification. Hence, new algorithms are needed to overcome these limitations by characterizing interspecies relative abundance of sequence-contexts that reflects shared and species-specific genomic information. In 2009, the discovery of a population of Ketupa Zeylonesis Semenowi (Brown fish owl) in Turkey, where it was believed to be extinct, provided a unique opportunity to identify the phylogenetic relationships among fish owls. In particular, distinct morphology and vocalization of Brown fish owl potentially constitute an un-described species that calls for special efforts on phylogenetic characterization of various populations of fish owls. Here we developed a new algorithm that uses dynamic k-mer profiles to improve on the identification and characterization of phylogenetic relationships among fish owls. We have obtained DNA samples from museum specimens of different fish owl populations from Syria, Nepal, India, and Bangladesh. Subsequently, we performed a whole-genome shotgun sequencing on Illumina HiSeq. After a rigorous quality-control and pre-processing regime, the algorithm measures the overall distance between generated k-mer profiles. Intriguingly, our model could classify these profiles on the basis of their intra- or inter-species relationships. Moreover, integration of the maximal information coefficient between different profiles helped to better characterize the phylogenetic tree of these fish owls. Our study explores several potential strategies in determining phylogenetic relationships and highlights some of the pitfalls associated with this endeavour, such as methods that heavily rely on small sets of target genes for molecular phylogenetic studies. In the dawn of next-generation sequencing, we believe that the work presented here and the future developments can provide powerful tools for the identification of shared and species-specific genomic information. These advancements would lead to a better understanding and characterization of phylogenetic relationships of rare or complex species.

64

GPCR structures: What moves where?

Rob van der Kant, Gert Vriend

CMBI, UMC St. Radboud Nijmegen

G protein-coupled receptors (GPCRs) are responsible for the majority of cellular responses to hormones and neurotransmitters as well as the senses of sight, olfaction and taste. Recently several crystal structures have been elucidated representing different GPCRs in different states of activation. Unfortunately these GPCRs were modified to great extent to allow them to form crystals. These modifications might induce conformational changes in the GPCRs structures, which makes them inaccurate descriptions of the true structures in living cells. We performed statistical analyses using Random Forests to separate facts from fiction and to elucidate the movements of the trans membrane helices upon activation.

65

Cross-Species Alignment of Co-expression Networks

M.E. van der Wees, M. El-Kebir, U.K. Nandal, P.D. Moerland, J. Heringa, G.W. Klau

Centrum Wiskunde & Informatica, Vrije Universiteit Amsterdam, AMC Amsterdam

Small animals, like rodents, are extensively used to study human diseases and to develop treatment methods. However, biological differences between species give rise to the question whether experimental results are actually transferable from modeling organisms to human. In this work we want to identify conserved sub-networks of genes among species by means of network alignment of co-expression networks. A network alignment is a mapping from nodes (genes) in one network to those in a second network and takes into account both sequence similarity and topological information of genes.

We derive networks by applying a threshold to co-expression matrices obtained from micro-array assays. Together with sequence similarity scores, obtained from all-against-all BLAST alignments, these node-to-node correspondences contribute to an objective function, which has to be optimized in order to find the best scoring network alignment. However, NP-hardness of this optimization problem complicates the search for the best scoring alignment. Therefore, lower and upper bounds to the solution are obtained with a Lagrangian relaxation approach, implemented in the network alignment tool Natalie.

A first step towards identifying highly conserved sub-networks, we composed two co-expression networks and one sequence similarity network from human and mouse liver sample data sets. Next, we defined an additive scoring function for sequence and topological similarity. First computational results indicate that the method scales well and produces meaningful conserved co-expressed clusters.

Future work includes identifying sub-networks that contain known disease genes and are not conserved between mouse and human samples. Furthermore, we will extend the network alignment algorithm with a branch and bound procedure, an exact approach to find provably optimal network alignments. Finally, we will apply our method to new mouse and human data sets from diabetes and metabolic syndrome samples.

66

Apoptotic cells are cleared by directional migration and elmo1-dependent macrophage engulfment

Tjakko van Ham[1], David Kokel[2], Randall Peterson[2]

[1] *Giepmans lab, Department of Cell Biology, UMCG, Groningen & Massachusetts General Hospital, Harvard Medical School, Boston, MA*
[2] *Department of Cardiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA*

Cell death by apoptosis shapes embryonic morphogenesis and is essential to tissue homeostasis in aging and disease. Failure to clear apoptotic cells can disrupt the functions of neighboring cells and lead to chronic inflammation and autoimmune disease. How disposal of apoptotic cells occurs in vertebrates is not well understood, and in vivo data are largely lacking.

Here, we use in vivo imaging to study disposal of apoptotic cells during zebrafish brain development. We find that apoptotic cells are remarkably motile, frequently migrating several cell diameters in minutes. Apoptotic cell motility is directional; cells move to the periphery of the brain via an actinomyosin-dependent process. Actin remodeling is observed within the apoptotic cell during migration, suggesting active rather than passive conveyance. During the first 2 days of development, engulfment is very rare, and most apoptotic cells lyse upon arrival at the brain periphery. At later stages, we find apoptotic cells are rapidly engulfed. Ablation of primitive macrophages by PU.1 knockdown blocks engulfment, indicating that macrophages are the main phagocytic cells. Genetic knockdown studies further demonstrate that engulfment is controlled by guanine exchange factor elmo1, which mediates Rac GTPase dependent cytoskeletal arrangements in engulfment. Elmo1 deficient macrophages are capable of finding apoptotic cells and binding to them, but engulfment is rare and occurs through apparently random macropinocytosis rather than directed phagocytosis.

In all, these findings suggest that clearance of apoptotic cells in brain development is accomplished by the combined actions of apoptotic cell migration and elmo1-dependent engulfment by macrophages. This provides novel insight into apoptotic cell clearance mechanisms in vivo, which may be particularly relevant to understanding chronic inflammatory and autoimmune conditions, including neurodegenerative diseases.

67

Regulatory Interaction plugin: a PathVisio plugin to visualize microRNA-target interactions

Stefan van Helden, Martina Kutmon, Lars M. Eijssen, Susan L. Coort, Chris T. Evelo

Bioinformatics-BiGCaT, Maastricht University, the Netherlands

Since the discovery of microRNAs (miRNAs), which are post-transcriptional regulators, many studies are performed to investigate their role in physiological and disease processes. Microarray technology makes it possible to measure miRNA expression on a large scale. The pathway visualization tool PathVisio (www.pathvisio.org) can provide valuable assistance to gain insight in which biological processes miRNAs play a role. In PathVisio pathways are schematic representations of the interactions between biological entities such as genes and proteins. These biological entities are annotated with identifiers and accession numbers from online databases with additional information about the selected item.

To include miRNAs in the pathway analysis, PathVisio should be able to visualize them. WikiPathways is an open, public platform dedicated to the curation of biological pathways by and for the scientific community. At the moment, most of the 1674 available pathways do not contain information about miRNAs. By expanding PathVisio's functionality with the Regulatory Interaction plugin, it is possible to automatically visualize the regulatory interaction partners of a selected pathway element in a side panel, as well as show the expression levels of these biological entities by loading a data set containing regulator and target expression data, e.g. miRNA and mRNA expression data.

To do this, we use a simple text file format containing two columns, regulator and target identifiers, in our case-study these are the interactions between miRNAs and genes. It is possible to load multiple interaction files simultaneously, e.g. MicroCosm Targets, miRTarBase and miRecords. If the user selects an element in a pathway, the plugin displays the interaction partners in a side panel. By loading an expression data set, this will also show the expression levels of the biological entities involved in the regulatory interaction.

The regulatory interaction plugin enables the integration of miRNA regulation in pathway analysis. In the near future, we will add the functionality to perform pathway statistics based on a set of miRNAs. The target genes of those miRNAs will be used to perform a gene set enrichment analysis. This could give some indication if a pathway is regulated by a set of miRNAs or not.

68

Project HOPE: Providing the last piece of the puzzle….

H. Venselaar

CMBI, UMC St Radboud Nijmegen

Introduction:
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 3D-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 a significant 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 prediction-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. Therefore we developed project HOPE, a web-server for automatic point-mutant analysis. Project HOPE predicts the structural effect of a mutation in a way that is easy-to-understand for anyone in the (bio)medical field. The web server is available on www.cmbi.ru.nl/hope

Methods:
The interface of Project HOPE is a website where the user can upload the sequence and mutation of interest. The core of Project HOPE is a Java-implemented module that collects information from a series of sources and stores this in a database. This information consists of calculations on the real structure or on a homology model by WHAT IF web-services (including accessibility scores, secondary structure, contact distances, variability etc), sequence based predictions by Distributed Annotation (DAS) Servers (including predictions for transmembrane domains, phosphorylation sites, etc), structural features annotated in the Uniprot database (including domain locations, known variations, motifs, etc.), conservation scores from HSSP, and GO-terms for known InterPro-domains. All information is stored per residue in a PostGreSQL database. A decision scheme is used to reach an informed conclusion about the effect of the mutation, thereby focusing on the structural changes. The conclusion is shown on the website and illustrated with figures and animations.

Results/Conclusion:
During the recent years we have collaborated in numerous projects with researchers from several (bio)medical departments. Our contribution to these projects consisted mainly of mutation studies (non-sense and missense mutations) that provided insight necessary to understand the effect of the mutation and to design new experiments. We used these projects to validate Project HOPE by comparing our manual analyses with the automatically generated reports. Additionally, we validated Project HOPE by analyzing over 100 mutations that were described in the 2010 volume of The American Journal of Human Genetics and other journals. We compared Project HOPE's reports with the conclusion drawn by the authors of these articles, and with the predictions generated by other widely-used methods such as SIFT, PolyPhen 2 and Grantham-scores. The results show that Project HOPE can correctly identify 95% of the mutations as damaging and that, in most cases, the generated reports are similar to the analysis by a well-trained bioinformatician. This result equals the score obtained by the popular PolyPhen 2 webserver. However, in contrast to this server, Project HOPE provides a detailed and illustrated report describing the effect of the mutation, and

thus can contribute to a better understanding of the protein and the mutation. The validation of Project HOPE also revealed several possible points for improvement which will be implemented in the (near) future.

In general we can conclude that Project HOPE is an easy accessible and understandable structural analysis server that provides information that can be beneficial for research in the (bio)medical field and can therefore be regarded as the last piece of the molecular puzzle.

69

Generating complex descriptions of sequence variants using HGVS nomenclature based on sequence comparison

J.F.J. Laros, M. Vermaat, J.T. den Dunnen, P.E.M. Taschner

Center for Human and Clinical Genetics, Leiden University Medical Center, The Netherlands

Descriptions of sequence variants can be checked and corrected with the Mutalyzer sequence variation nomenclature checker (https://mutalyzer.nl/) to prevent mistakes and uncertainties which might contribute to undesired errors in clinical diagnosis. Construction of variant descriptions accepted by Mutalyzer requires comparison of the reference sequence and the variant sequence and basic knowledge of the Human Genome Variation Society sequence variant nomenclature recommendations (http://www.hgvs.org/mutnomen/). With the advert of sophisticated variant callers (e.g., Pindel) and the rise of long read sequencers (e.g., PacBio), the chance of finding a complex variant increases and so does the need to describe these variants. An algorithm performing the sequence comparison would help users to describe complex variants.

The algorithm closely follows the human approach to describe a variant. It will first find the "area of change", and then finds the largest overlap between the original area and the area in the observed sequence. This process is repeated until the smallest description is found.

This algorithm ensures that the same description will be generated every time researchers observe this variant. Furthermore, no knowledge of the HGVS nomenclature is required to generate this description. This not only helps clinicians to generate the correct description, but its implementation also allows automation of the description process.

We have incorporated this algorithm in the Mutalyzer suite under the name Description Extractor (https://mutalyzer.nl/descriptionExtract).

Funded in part by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 200754 - the GEN2PHEN project.

70

Patterns recognition on microbiota

X. Wang[1] ,M.J.C. Eijkemans[2], G. Biesbroek[1], W.T. Hendriksen[1], W. J. Wallinga[3], E.A.M. Sanders[1], D. Bogaert[1]

[1] *Department of Pediatric Immunology and Infectious Diseases, UMC Utrecht, Utrecht, The Netherlands.*
[2] *Julius Center for Health Sciences and Primary Care, UMC Utrecht, Utrecht, The Netherlands*
[3] *Department of Infectious Diseases Epidemiology, National Institute of Public Health and the Environment, Bilthoven, The Netherlands.*

It is believed that microbiota of the upper respiratory tract of healthy humans generally represent a balanced ecosystem, where an imbalanced composition might predispose to development of respiratory infections. High-throughput sequencing methods enable us to investigate microbial composition of ecological niches. However, new applied analysis methods for identifying patterns in these large data sets are needed.

We developed a flexible strategy that determines the type of clustering method needed to analyse a given metagenomic dataset. With this strategy, we determined the optimal combination of the number of clusters, the distances, and linkages by calculating the silhouette index, the Davis Bouldin Index and applying the consensus clustering method. The developed strategy was applied to a published dataset containing 16S-rRNA-based microbiota composition data of 96 nasopharyngeal samples from children 18 months of age.

The absolute correlation distance and complete linkage performed best, using a structure with 13 clusters. With these settings, we confirmed the previously culture-proven negative correlation between Moraxella and Haemophilus influenzae. Newly identified correlations were correlations between Pasteurellaceae, Lachnospiraceae, and Pseudomonas, Streptococcus and Fusobacterium, and Dolosigranulum and Neisseria.

With the proposed strategy, we feel we have a reliable tool to identify bacterial correlations within a given metagenomic dataset. The usefulness of this approach was illustrated by the observation of several clinically relevant bacterial interactions in our nasopharyngeal dataset. This (flexible) approach can be applied to any microbiome data sets for identifying interaction patterns.

71

Robust genome-wide recombination patterns following heritable epigenetic perturbation in Arabidopsis

René Wardenaar[1], Maria Colomé-Tatché[1], Sandra Cortijo[2], Ritsert C. Jansen[1], Vincent Colot[2] & Frank Johannes[1]

[1] National Centre for Scientific Research UMR8186, Department of Biology, Ecole Normale Supérieure, 75230 Paris Cedex 05, France.
[2] Groningen Bioinformatics Centre, University of Groningen, Nijenborgh 7, 9747 AG, Groningen, The Netherlands.

Recombination is an essential aspect of genome evolution. It is an important source of variation for natural selection to work upon. In Arabidopsis thaliana, meiotic cross-over patterns are tightly controlled along chromosomes and correlate with DNA methylation levels and broad DNA sequence features.

Here we test the robustness of these patterns in response to an experimental perturbation of global cytosine methylation in an approximately isogenic population of epigenetic recombination inbred lines (epiRILs). This population of epiRILs was constructed from a cross between a Columbia wt plant and a Columbia ddm1 mutant plant (Johannes F. et al., 2009). DDM1 mutant plants have a ~ 70 % reduction in methylation levels mostly over repeat sequences (Vongs A et al., 1993) and these alterations are heritable for many meiotic generations even in the absence of the conditioning locus.

Using genome-wide methylation data from 123 epiRILs we derived a core recombination map on the basis of 867 meiotically stable Differentially Methylated Regions (DMRs) covering 91.5% of the total genome. The total estimated genome length of this map was 417.7 cM which is similar to previous reports in Arabidopsis.

We evaluated our map by comparing it to 17 recently published Arabidopsis F2 populations which were derived from pairs of 18 distinct natural accessions (Salomé et al., 2011). This allowed us to place the epiRIL map in the larger context of natural variation and to quantify the impact of ddm1-induced loss of DNA methylation on genome-wide recombination patterns. A comparison of the genetic lengths for each of the five chromosomes revealed very similar estimates across these diverse populations, suggesting that global recombination rates are remarkably robust. We show that these similarities also hold at a local scale.

72

PyPaSWAS: CUDA-based Smith-Waterman on graphics cards in Python

S. Warris, J.P. Nap

Expertise centre ALIFE
Institute for Life Science & Technology
Hanze University of Applied Sciences Groningen, The Netherlands

The CUDA-based Parallel Smith-Waterman Alignment Software (PaSWAS) performs local alignments on graphics hardware. The CUDA language is developed by NVIDIA Corporation to enable software developers to create applications other than graphics that use the computational power of the graphics processing unit. CUDA is an extension of the C++ language.

There are two issues when developing CUDA applications for bioinformatics research. First, not many bioinformatics researchers have sufficient knowledge of C++ to develop entire applications in this language. Python, Perl or Java are much more common languages used in bioinformatics. Second, many bioinformatics toolkits, libraries and other extensions are available for these common languages. We here demonstrate that Python is a suitable environment for developing CUDA applications.

The CUDA library for Python is called pyCUDA. Other important bioinformatics-related packages are numpy for numerical calculations and bioPython. The latter is used for reading and writing different file types and performing basic analyses on sequence data. These packages are combined with other publicly available packages to create pyPaSWAS. This software is used for different types of sequence alignments based on the Smith-Waterman algorithm, for example for primer detection in next generation sequence reads, or for large scale read mapping. It automatically detects the type of graphics card present and calculates the amount of GPU memory available. The use of Python for developing such CUDA-based applications should facilitate the advance of graphics hardware applications in bioinformatics.

73

Comparing score-optimal protein structure alignments

Inken Wohlers[1], Noël Malod-Dognin[2], Rumen Andonov[3], and Gunnar W. Klau[1]

[1] *Life Sciences, Centrum Wiskunde & Informatica, the Netherlands*
[2] *INRIA Sophia Antipolis - Méditerranée, France*
[3] *INRIA Rennes - Bretagne Atlantique and University of Rennes 1, France*

Background: Comparison and alignment of protein structures is important for detecting structural, functional and evolutionary relationships. Structure alignment is difficult, firstly because finding the alignment of maximum score is an NP-hard problem and secondly because of the wide spectrum of qualitatively different structural similarities. There is a large number of competitive programs for structure alignment, each optimizing its own scoring scheme. To date, there is no consensus which scoring best captures all biologically relevant similarities.

Methods: We developed a general framework and an exact algorithm that aims to find the top-scoring alignment according to four different scoring schemes, those of CMO, PAUL, DALI and MATRAS. If the provably top-scoring alignment is not determined within the time limit, the best alignment found so far is returned, together with an upper bound on the optimal score.

Results: Our algorithm computes CMO, PAUL, DALI and MATRAS alignments via our web server which is called CSA. CSA additionally provides a variety of functionalities for comparing these and other, user-uploaded alignments by the use of many alignment quality measures and various visualizations. Such a multivariate comparison helps identifying the most appropriate alignment for a given protein pair.

74

Analysis and Prediction of Interactions of PDZ Domain on PICK1 using PRISM

B. Tugce Yildizogly, Ozlem Keskin, Attila Gursoy

Koç University, Turkey

Protein interacting with C kinase (PICK1) is an evolutionarily highly conserved protein which interacts with over 60 proteins, including receptors, transporters, kinases and ionic channels. Being a critical player in synaptic plasticity, development and neural guidance, it regulates the trafficking and posttranslational modification of its interacting proteins taking part in neuronal function. Correspondingly, the role of PICK1 in disorders such as epilepsy, pain, brain trauma and stroke, drug abuse and dependence, and schizophrenia has been shown several times before. Strikingly, PICK1 carries out its repertoire of interactions mostly through the single PSD95/DlgA/Zo1 (PDZ) domain it bears. Differently from most PDZ domains, the PICK1 PDZ domain interacts with partners of type I, type II and atypical based on the properties of the residues of binding sequences, thereby being promiscuous to a certain degree while keeping the selectivity. Given all that, understanding the binding properties and interaction mechanisms of PICK1 emerges as an important research area. In this study, PRISM (Protein Interactions by Structural Matching) which is an algorithm to predict protein-protein interactions based on evolutionary and structural similarity to template structures given a target structure, is revisited. The algorithm is modified to tailor smaller structures for investigation of interactions of PICK1 with different types of PDZ binding peptides and structure of PDZ interactions more generally. It is aimed to compare the binding affinities of various PDZ binding peptides, study the effect of mutations of critical residues and identify potential partners to be targeted for therapeutical studies.

75

The FUNG-GROWTH database: linking growth to genome

M. Zhou[1], A. Wiebenga1, Vincent Robert[1], Pedro M. Coutinho[2], Bernard Henrissat[2], Ronald P. de Vries[1]

[1] *CBS-KNAW Fungal Biodiversity Centre, Utrecht, The Netherlands;*
[2] *AFMB, Marseille, France;*
e-mail: r.devries@cbs.knaw.nl

Aspergillus genome sequences demonstrate the potential to utilize a variety of different carbon sources. Natural carbon sources for many fungi are based on plant biomass and often consist of polymeric compounds, such as polysaccharides. They cannot be taken up by the fungal cell and are extracellularly degraded by a complex mixture of enzymes. Plant polysaccharide degrading enzymes have been studied for decades due to their applications in food and feed, paper and pulp, beverages, detergents, textile and biofuels. These enzymes have been classified based on amino acid sequence modules (www.cazy.org).

Based on the hypothesis that Aspergillus genomes have evolved to suit their ecological niche, we have performed a comparative study using 34 Aspergillus species/strains. In this study we have compared growth profiles on 35 different carbon sources (consisting of mono-, oligo- and polysaccharides, lignin, protein and crude plant biomass) to the CAZy annotation of the genomes to identify correlations between growth and genomic potential. Further analysis involves comparative transcriptomics using various system biology strategies, focusing on degradation related gene regulators.

Highlights of these integrative bioinformatics analysis will be presented as well as the public database in which the growth data is stored and the developments of the database anticipated for the next two years.

URL: www.fung-growth.org

76

Detection of genes essential for growth of respiratory pathogens

Aldert L. Zomer[1,3], Peter Burghout[1], Stefan de Vries[1], Hester J. Bootsma[1], Jeroen Langereis[1], Hendrik G. Stunnenberg[2], Peter W.M. Hermans[1] and Sacha van Hijum[3]

[1] *Laboratory of Pediatric Infectious Diseases, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands*
[2] *Department of Molecular Biology, Radboud University, Nijmegen, The Netherlands*
[3] *Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands*

Background:
Respiratory tract infections are a leading cause of global mortality and morbidity. The WHO estimates that annually 4-5 million people die of pneumonia. Streptococcus pneumoniae is among the most important respiratory tract pathogens. Infection by and growth of S. pneumoniae is a complex process dependent on a number of essential pathways, of which some could form ideal candidate targets for drug design and/or vaccine development.

Methods:
To identify microbial genes essential for growth of respiratory pathogens, we have used the Tn-Seq insertion knockout and sequencing strategy and developed a bioinformatics tool that allows rapid identification of disrupted genes. Genes underrepresented in the knockout library are likely essential for growth. To identify shared essential pathways in bacterial species we have used statistical analysis, pathway analysis and functional category enrichment methods.

Results:
In S. pneumoniae, Haemophilus influenzae and Moraxella catarrhalis we observed that ~15% of all genes is essential, similar to what has been found in knockout studies. These genes primarily encode functions involved in transcription, translation or replication, but also encode for hypothetical proteins orthologous between the tested species.

Conclusions:
High throughput screening of essential genes is feasible using Tn-Seq. To this end, a large knockout library is preferred to decrease the number of false positives. Importantly, genes encoding for orthologous proteins in all three species have been found to be essential, including hypothetical proteins, which are likely to play important roles in critical cellular processes and might form novel candidate targets for drug design and/or vaccine development

URL: http://bamics2.cmbi.ru.nl/websoftware/essentials/essentials_start.php

77

Computational analysis of the modular architecture of secondary metabolite biosynthesis gene clusters

Konrad Zych[1,2,3], Marnix H. Medema[1,2], Eriko Takano[1] & Rainer Breitling[2,4]

[1] *University of Groningen, Nijenborgh 7, 9747 AG Groningen, The Netherlands.*
[2] *Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Nijenborgh 7, 9747 AG Groningen, The Netherlands.*
[3] *Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University, Gronostajowa 7, 30-387 Krakow, Poland*
[4] *Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, Joseph Black Building, University of Glasgow, Glasgow G12 8QQ, UK.*

The biosynthetic pathways of the great majority of secondary metabolites with pharmaceutical activities are encoded by huge clusters of genes, termed secondary metabolite gene clusters (SMGCs). Inside SMGCs, genes are further grouped into conserved multigene modules, each of which is responsible for the biosynthesis of a part of the end product. To gain deeper insight into the evolution of SMGCs and to open up new ways to discover novel compounds, we analysed the modularity of SMGCs computationally in a high-throughput fashion.

As a starting point of our analysis, we created a list of SMGCs from all available actinomycete nucleotide data using our previously published software pipeline antiSMASH [1]. To identify all highly conserved modules, we then studied co-conservation of the genes within these gene clusters. Based on a classification of all SMGCs genes into clusters of orthologous groups (COGs), we reconstructed interaction networks linking COGs by gene synteny and by co-localization of genes within the same gene clusters. A simple algorithm that overlaid the two networks could then identify highly connected motifs of COGs. These motifs represent conserved modules that can be directly linked to the chemical moieties of the secondary metabolite end product. Using these data, we were able to identify a number of gene clusters with conserved architectures (module compositions) that had not been reported earlier, which may be responsible for the biosynthesis of compounds with novel chemical structures.

Intriguingly, our catalogue of conserved modules enables a deep analysis of the evolution of SMGCs, by comparing module compositions of homologous gene clusters from different species and using parsimony or likelihood methods to infer the most probable evolutionary route that has resulted in the observed variety of architectures. It also enables screening for novel types of gene clusters on an unprecedented scale. Finally, this new approach can be integrated into the antiSMASH pipeline to allow more detailed predictions of the secondary metabolite end products of unknown SMGCs from their module compositions.

[1] Medema et al. (2011), antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. 39: W339-W346.

78

Pheno2geno - Constructing genetic maps from molecular phenotypes.

Konrad Zych[1,2], Danny Arends[2], Ritsert C. Jansen[2]

[1] *Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University, Krakow, Poland*
[2] *Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, The Netherlands.*

Pheno2geno is a package for the computational generation of markers and construction of genetic maps from molecular phenotypes in segregating inbred line populations, what is possible because phenotypes with a clearly separated bimodal or trimodal expression distribution can be used as genetic markers. Pheno2geno offers: de novo map construction, saturation of existing maps and detection of sample mix-ups.

Pheno2geno starts by finding the phenotypes that are suitable as markers. These phenotypes should show differential expression between founders and are selected using a Student's t-test or RankProd. In the next step mixture modeling is used to select phenotypes that show bimodal (e.g. for RIL, BC) or trimodal (e.g. for F2) expression patterns with mixing proportions comparable to the expected segregation frequencies (e.g. 1 to 1 in RIL) across all the offspring individuals. Only these phenotypes are transformed from continuous measurements into discrete markers.
After marker detection the markers are used to create de novo map. Additional information, like known physical and/or genetic positions for all/some of the markers could be used and it will improve the quality of the resulting map.

When a genetic map is available pheno2geno can be used to saturate it. For each of the markers it is assessed whether it has a single significant QTL. If so, it is being placed on the map in the position of this peak, if not it is dropped from the analysis.

Pheno2geno can be used to point out sample mix-ups and errors. Using QTL information the package compares the observed phenotype value with the expected value. This results in calculating a mismatch score, and obtaining phenotype-based idealized genotypes.

Tests on an 180 individuals from A. thaliana inbred population show that pheno2geno is able to point out wrongly measured samples and saturate the genetic map. Saturation results in a decrease of the average distance between markers by ~50%.

Poster list

Researcher	Title abstract	Page	Poster number
Anvar, Y.	Deep-sequencing of TALENs targeted embryonic stem cells to estimate their efficacy in genome editing	13	1
Atici, P.	PRISM (Protein Interactions by Structural Matching)	14	2
Baakman, C.	MicroWeb: making BD Pathway Data more Accessible	15	3
Babaei, S.	Detecting recurrent gene mutations in pathway context using multi-scale graph diffusion	16	4
Backus, L.	Random Forest based data analysis of ~omics data: a novel method for detecting sub-classifications governed by interacting variables.	17	5
Bargsten, J.	Structural homology in Solanaceae: analysis of genomic regions in support of synteny studies in tomato and potato	18	6
Bawono, P.	How do you recognise your distant relatives? Application of profile HMM for aligning distant proteins	19	7
Bayjanov, J.R.	Grid and cloud computing for high throughput assembly and annotation of (meta)genome sequences	20	8
Bayjanov, R.	PhenoLink? A web-tool for linking phenotype to ~omics data for bacteria: application to gene-trait matching for Lactobacillus plantarum strains	21	9
Beek, D.	Detection of pathogenic copy number variations in shallow next generation sequencing data	22	10
Beyelas, G.	Using grid middleware as a computational resource for large-scale NGS and imputation in the eBioGrid project	23	11
Bijl, M.	VASP3 - Central Storage and Analysis Pipeline for Transcriptomics Data	24	12
Binsl, T.	Ready-made IT platform for diagnostic analysis of medical data	25	13
Bosdriesz, E.	Optimal and robust regulation of gene expression	26	14
Bot, J.	BiG Grid infrastructure: Facilitating the computational needs for current bioinformatics applications	27	15
Bouwman, J.	Structured data analysis and storage to facilitate systems biology analysis	28	16
Bruning, O.	Sampling the Sweet Spot	29	17
Chibon, P-Y	Marker2sequence: From QTLs to potential candidate genes.	30	18
de Hollander, M.	Ongoing efforts of a Galaxy solution for the BiGGrid/SARA HPC Cloud	31	19
de Leeuw, W.	Managing cloud computing for life sciences research via smart interfaces	32	20
de Ruiter	Determining gene regulation mechanisms through multi-scale integrative analysis of genomic signals	33	21
Deelen, P.	Effect of population specific imputation reference set using second genotype chip as gold standard	34	22
Dinkla, K.	Compressed Adjacency Matrices: Untangling Gene Regulatory Networks	35	23

Researcher	Title abstract	Page	Poster number
Dymecka, M.	Bioinformatical analysis of the GRAS family of proteins	36	24
Ederveen, T.	COMPANION: Comparative genome annotation in prokaryotes: a halt to error propagation?	37	25
El-Kebir, M.	NatalieWEB: A web server for topology-aware global protein-protein interaction network comparison	38	26
Gritsenko, A.	A sequence optimization method for increased gene expression	39	27
Hanssen, B.	Dockland's got talent! Determining relations between structure quality and docking quality.	40	28
Heideman, H.	Eighteen-fold Performance Increase for Short-Read Sequence Mapping on Genome of the Netherlands (GoNL) data using a Hybrid-Core Architecture	41	29
Hill, S.	Data-driven characterisation of protein signalling networks in cancer	42	30
Jansen, G.	Advanced SKOS-aided Query Builder (ASQ-builder): Building smarter queries for public databases (PubMed, OMIM) using SKOS Vocabularies	43	31
Joosten, R.	PDB_REDO: optimized protein structures for bioinformatics research	45	32
Lauber, C.	A genetics-based approach to partitioning the natural diversity of an RNA virus family	46	33
Liu, D.	Towards Watching the Bottlenecks of Membrane-Protein Biogenesis by Single-Molecule Methods	47	34
May, A.	Fast and accurate calculation of protein-protein interaction: contribution of surface and interface residues	48	35
McClellan, E.	MS PRUNA: Mass Spectrometry Peptide Re-mining Using Network Analysis	49	36
Mina, E.	e-Science in bioinformatics: enabling the reuse of methods while unravelling the epigenetic factors of Huntington's Disease	50	37
Morgado, L.	Transgenerational analysis of DNA methylation inheritance in Arabidopsis thaliana	51	38
Nandal, U.	Compendium database of rodent and human metabolic syndrome related datasets	52	39
Nijveen, H.	Small Multiples View Cytoscape plugin for comparing gene expression in a network context	53	40
Nooren, E.	e-BioGrid: Building a Dutch e-science infrastructure for life science research	54	41
Obbink, H.	TraIT: an integrated translational research IT platform	55	42
Overmars, L.	Are REPs genetic insulators that enable differential regulation of gene expression in bacteria?	56	43
Paalvast, Y.	Pep2Path: Enabling high-throughput antibiotic discovery by automatic identification of peptide biosynthesis pathways	57	44
Pirovano, W.	Towards (almost) closed genomes with SSPACE and GapFiller	58	45
Pronk, G.	NBIC as IDP - what does it mean for you?	59	46
Rana, A.	Stochastic transitions in oncogenic transformation	60	47

Researcher	Title abstract	Page	Poster number
Zhou, M.	The FUNG-GROWTH database: linking growth to genome	90	75
Zomer, A.	Detection of gene essential for growth of respiratory pathogens	91	76
Zych, K.	Computational analysis of the modular architecture of secondary metabolite biosynthesis gene clusters	92	77
Zych, K.	Pheno2geno - Constructing genetic maps from molecular phenotypes.	93	78