Poster Abstracts

Describing Complex Biological Phenomena by Switched Systems

Rodrigo Assar, Alice Garcia and David James Sherman

University Bordeaux 1, INRIA, Bâtiment A30 Domaine Universitaire, 351 cours de la Libération, 33405 Bordeaux, France

Analysis of Complex Biological Phenomena requires approaches combining continuous, discrete, stochastic, deterministic and non-deterministic behaviors in a non-ambiguous way. We studied how to model and simulate these systems by descriptions human-comprehensible and machine-readable.
To start, we defined Stochastic Switched Systems (SSS) whose continuous dynamics is modeled by differential equations and its discrete dynamics by transition systems, allowing stochastic and non-deterministic behaviors. Such description is human-comprehensible, while BioRica codes are machine-readable. We decomposed the continuous dynamics into diverse interacting modules, integrating different descriptions or implementations, allowing the reuse and reconciliation of models.
We developed some SSS applications and gave rules to reduce interacting systems into SSS that, applied to Gene Regulatory Networks, simplify the analysis of cell division and differentiation.
Switched systems, transition systems and hierarchical modular design allow to obtain human and machine-readable descriptions of a wide range of complex biological phenomena.

Mathematical modeling of protein aging and turnover in live yeast cells

Joseph Barry, Anton Khmelinskii, Michael Knop and Wolfgang Huber

European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany

An understanding of the subcellular dynamics of the eukaryotic proteome is essential to infer protein function. However, a comprehensive genome-wide study of the dynamics of protein age and turnover at subcellular spatial resolution is so far lacking. To this end, we endogenously tag proteins in S. cerevisiae with a tandem fluorescent protein construct and measure fluorescence intensities using live cell imaging. Robust image segmentation techniques and mathematical models of protein turnover are presented here that are necessary to relate protein fluorescence intensity to the number of protein molecules. The relationship between stochastic and deterministic models of protein turnover is explored, as is protein inheritance variation between mother and bud yeast cells. The results of this study will ultimately provide a foundation upon which other eukaryotic proteomes, including human, may be understood.

Random walks on graphs to predict drug treatment efficacy and potential side-effects

Jacques Colinge, Uwe Rix and Giulio Superti-Furga

CeMM.Research Center for Molecular Medicine GmbH, Lazarettgasse 14/AKH BT 25.3, 1090 Vienna, Austria

We present new developments of global network scoring methods we use to build disease and drug action models, which we apply to predict cancer drug treatment efficacy and potential side-effects. These models are based on random walks on graphs. They are very flexible and allow us to easily integrate different sources of data such as mutation frequencies, drug target affinities, and protein interactions to cite the most common, but also biological pathways, protein phosphorylations, etc.

The random walk scores are especially useful when drugs target hubs in the human interactome as they naturally avoid loss of specificity in models growing exponentially fast. We shall illustrate the application of the new scores to several kinase inhibitors used to fight different forms of leukemia and show good adequacy with experimental data and literature. We shall also show how new potential applications of existing drugs can be proposed and diseases compared.

An Empirical Codon Hidden Markov Model

Nicola De Maio and Carolin Kosiol

Institute of Population Genetics, VetMedUni Vienna, Veterinärplatz 1, 1220 Vienna, Austria

Codon models are widely used for estimation of positive selection, phylogenetic trees, and many other tasks.
Empirical codon models summarize mutational patterns from huge amounts of data and have hundreds of parameters, thus are computationally requiring but also promising tools.
We estimated empirical codon models from full-genome Drosophila data (12 genomes and 50 genomes projects data). Choosing different species subtrees we study the effect of species choices on the estimated parameters.
We propose several semi-empirical models in order to assess which level of model complexity describes the data most efficiently. Furthermore we introduce hidden Markov classes to model and detect variation in selective pressure and codon usage among alignment sites. Finally, we investigate branch-specificity of evolutionary rates.

Genome-Wide Associations for Carcass traits in Irish Holstein Friesian Cattle

Anthony Doran, Chris Creevey, J McInerney, and Donagh Berry

Teagasc, Animal and Bioscience Research Department, 0000 Dunsany, Ireland

In recent years there has been an ongoing revolution in genomic research. Genome-Wide methods have been identified as a major approach to fully investigate complex traits such as disease susceptibility, health and animal agricultural performance. For the analysis reported here a Genome-Wide association study for carcass traits (carcass weight, fat, muscle conformation and cull cow carcass weight) on Irish cattle was performed. The analysis was performed using the results from the 50k Bovine SNP chip from across 1061 animals, that were extensively phenotyped for the traits of interest. Bayesian regression models as described by Meuwissen et al. (2001) (BayesA and BayesB) were employed along with a derivative of each (BayesA_P and BayesB_P) to include polygenic effects within the models. The study provided insights into the quantitative trait loci (QTL) associated with carcass traits from Irish Holstein Friesian cattle. These QTLs will be used as the foundation for further analysis and identification of genes and gene pathways involved in muscle production traits.

Species disambiguation using random walks over a mixture of adjacency matrices

Nathan Harmston

Imperial College London, Biochemistry Building/Department of Biological Sciences, London SW7 2AZ, UK

The scientific literature contains a plethora of information about biological systems. Manual curation lacks the scalability to extract this information due to the ever-increasing numbers of papers being published. The development and application of text mining technologies has been proposed as a way to deal with this problem. However, the inter-species ambiguity of the genomic nomenclature makes mapping of gene mentions identified in text to their corresponding Entrez gene identifiers an extremely difficult task. We propose a novel method which transforms a MEDLINE record into a mixture of adjacency matrices; by performing a random walk over the resulting graph we can perform multi-class supervised classification. Such graph mixtures add flexibility and allow us to generate probabilistic classification schemes that naturally reflect uncertainties in data. Our method achieves state of the art performance on a publicly available corpus and improves over standard classification techniques in a number of ways: flexibility, interpretability and is resistant to the effects of class bias in the training data.

Analyzing In Vitro Selection Experiments using Next Generation Sequencing Technologies

Barbara Keil and Peter F Arndt

Max Planck Institute for Molecular Genetics, Department Vingron, Ihnestrasse 63-73, 14195 Berlin, Germany

We conducted in vitro selection experiments using the SELEX (Systematic Evolution of Ligands by Exponential Enrichment) protocol. Starting with a large pool of random DNA, sequences are selected iteratively with respect to their binding affinity to a target molecule. Recent advances in DNA sequencing technologies give us the unprecedented opportunity to sequence a fraction of the DNA pool after each iteration. Therefore, we can analyze the dynamics of the evolving sequence pool in great detail. As in vitro selection experiments mimic an evolutionary process, they offer the possibility to study the dynamics of population genetics models of selection analytically and experimentally. We model the number of copies of a sequence after each round as a Markov chain where the mean copy number is computed using a biophysical model. A statistical test that we developed allows us to judge whether a sequence in the initial library of DNA sequences was enriched.

Nonparametric Bayesian data fusion to integrate time-course and static data

Paul D. W. Kirk, Emma J. Cooke, Richard S. Savage and David L. Wild

University of Warwick, Warwick Systems Biology Centre, Coventry House (Top floor), Coventry CV4 7AL, UK

Integrating different sources of data remains a fruitful means by which to generate hypotheses and gain new insights in systems biology. Time-course measurements have become increasingly common, requiring the development of novel tools for modelling these data sets and exploiting the information they contain. We present a form of hierarchical Dirichlet process mixture model for combining different sources of data, which uses Gaussian process regression to model gene expression time-courses. Our approach clusters genes together, and determines (on a gene-by-gene basis) whether we should cluster on the strength of the data sets considered separately or in combination. We demonstrate our method by applying to S. cerevisiae data, and show the advantages of using realistic models of time-courses compared to those that do not respect the covariance structure of the data.

Learning false discovery rates by fitting sigmoidal threshold functions

Bernd Klaus and Korbinian Strimmer

University of Leipzig, Härtelstraße 16-18, 04107 Leipzig, Germany

False discovery rates (FDR) are widely used for multiple testing in genomic analyzes, for instance in microarray-based transcriptomic or proteomic studies. Typically, FDR is estimated from a mixture of a null and an alternative distribution. Here, we study a complementary approach proposed by Rice and Spiegelhalter (2008) that uses as primary quantities the null model and a parametric family for the local false discovery rate. Specifically, we consider the half-normal decay and the beta-uniform mixture models as FDR threshold functions. Using simulations and analysis of real data we compare the performance of the Rice-Spiegelhalter approach with that of competing FDR estimation procedures. If the alternative model is misspecified and an empirical null distribution is employed the accuracy of FDR estimation degrades substantially. Hence, while being a very elegant formalism, the FDR threshold approach requires special care in actual application.

Unbiased estimates of transposable element population frequencies in a natural population of D. melanogaster using next generation sequencing

Robert Kofler, Andrea Betancourt and Christian Schloetterer

VetMedUni Vienna, Insitute of Population Genetics, Veterinaerplatz 1, 1210 Wien, Austria

Transposable elements (TE) are stretches of genomic DNA that are able to amplify in the genome. They have been found in most organisms investigated so far and typically constitute 3 – 80% of the genomic DNA. Genome-wide studies of TE abundance and TE insertion frequencies in populations have been hampered by the high cost and the bioinformatics challenges involved. Here we introduce a novel and cost efficient approach to estimate TE population frequencies for TEs that are present in the reference sequence as well as for novel TE insertions. This approach utilizes paired-end Illumina sequencing of a pooled population. Using a natural population of Drosophila melanogaster from northern Portugal we demonstrate the utility of our novel approach. We found 6,824 previously not annotated TE insertions. We show that with respect to TE dynamics the low recombining regions of the centromere are significantly different from the low recombining regions of the telomere. Furthermore by scanning for fixed TE insertions in high recombining regions that are close to low Tajima's D values we present 31 novel candidates of positively selected TE insertions.

Sensitivity, robustness and identifiability in stochastic chemical kinetics models

Michal Komorowski and Michael Stumpf

Imperial College London, Centre for Bioinformatics, Division of Molecular Biosciences, London w14 9ss, UK

We present a novel and simple method to numerically calculate Fisher Information Matrices for stochastic chemical kinetics models. The linear noise approximation is used to derive model equations and a likelihood function which leads to an efficient computational algorithm. Our approach reduces the problem of calculating the Fisher Information Matrix to solving a set of ordinary differential equations. This is the first method to compute Fisher Information for stochastic chemical kinetics models without the need for Monte Carlo simulations. This methodology is then used to study sensitivity, robustness and parameter identifiability in stochastic chemical kinetics models. We show that significant differences exist between stochastic and deterministic models as well as between stochastic models with time-series and time-point measurements. We demonstrate that these discrepancies arise from the variability in molecule numbers, correlations between species, and temporal correlations and show how this approach can be used in the analysis and design of experiments probing stochastic processes at the cellular level.

Improvement of RNA-Seq precision

Pawel P. Labaj, German G. Leparc, Bryan Linggi, Lye Meng Markillie, H. Steven Wiley and David P. Kreil

Chair of Bioinformatics, Boku University Vienna, Muthgasse 18, 1190 Vienna, Austria

Measurement precision determines the power of any subsequent analysis, such as the detection of differential expression. The collection of RNA-Seq datasets with technical replicates now allows a systematic analysis of the precision of expression-level estimates from massively parallel sequencing.

A good target coverage of 84% of the estimated true transcript population could be achieved with 331 million 50bp reads, and diminishing returns were seen from increased sequencing-depths. In contrast, the majority of the measurement power (75% of all reads) was spent on only 7% of the transcripts, making less strongly expressed genes hard to measure: less then 30% could be quantified reliably with a relative error < 20%.

We introduce a new approach for analysing sequencing-reads, increasing the number of reliably measured transcripts by 50% to ~57,000. Still less than what can be achieved with standard microarrays, extrapolations to higher sequencing-depths highlight the need for more efficient experimental strategies combining technologies with complementary strengths.

Investigating primary and secondary regulatory mechanisms for hypertension using genome-wide gene expression in multiple tissues and radiotelemetric blood pressure in the rat

Sarah Langley, Matthias Heinig, Leonardo Bottolo, Josef Zicha, Jaroslav Kunes, Vaclav Zidek, Ted Kurtz, Michal Pravenec, Sylvia Richardson, Norbert Hubner, Stuart Cook, Timothy Aitman and Enrico Petretto

Imperial College Faculty of Medicine, Hammersmith Hospital, Du Cane Road, London W12 0NN, UK

Hypertension is a risk factor for several cardiovascular diseases and the underlying genetic factors can be explored using the heritability of blood pressure (BP) variation across species. In this study, we used the BXH/HXB rat recombinant inbred strains to disentangle the role of primary and secondary regulatory mechanisms underlying hypertension. We modeled variation in multiple BP phenotypes using spectral and wavelet methods, and integrated these data with genome-wide expression in seven rat tissues both at the individual gene- and gene network-levels. Cross-tissues comparative analysis of transcripts, expression quantitative trait loci and co-expression networks correlated with BP (FDR<5%) identified candidate genes (primary regulation) and functional pathways (secondary regulation) for hypertension. We translated these findings into humans, and found significant enrichment (FDR<10%) of GWAS signals for rat genes that specifically associated with pulse pressure, spectral and wavelet phenotypes, indicating increased power of these analyses to identify new candidates for hypertension across species.

Bulk segregant analysis with next generation sequencing

Gen Lin, Paul Theodor Pyl, Lars Steinmetz and Julien Gagneur

EMBL Heidelberg, Meyerhofstrasse 1, 69117 Heidelberg, Germany

Quantitative trait loci (QTLs) mapping interprets the link between phenotypic differences and genomic variations. This valuable provides insights in predicting disease and breeding plants and animals [1].

In Bulk segregant analysis, a large pool of genetically distinct cells is grown several generations under selective conditions and genotyped population-wise [2]. Enriched beneficial alleles in the final population are identified through changes in allele frequency profiles. Calls for causative variants are then posited as local maxima in the profile. Though this method holds promise in the advent of high-throughput sequencing [3], no computational method is established to assess statistical significance of the calls.

We have developed a method that (i) reports confidence interval on allele frequency peak positions and (ii) provides hypothesis testing for difference in allele frequency. Our method is non-parametric and partially controls for sequence bias at polymorphic positions.

Analysis of preliminary data from a cross between 2 strains of Saccharomyces cerevisiae shows cases that the method correctly identifies known QTLs

Identification of interactions between transcription factors in tissues using DNA binding affinity

Alena Mysickková and Martin Vingron

Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin, Germany

Tissue-specific gene expression is generally regulated by combinatorial interactions among transcription factors (TFs) which bind to the DNA. Despite this known fact, previous discoveries of the mechanism that controls gene expressions usually consider only a single TF. Here, we provide a prediction of interacting TFs in 30 human tissues based on their DNA binding affinity in promoter regions. We analyzed all possible pairs of 130 vertebrate TFs from JASPAR database. First, all human promoter regions were scanned for single TF-DNA binding affinities and for each TF, a rank list of all promoters ordered by the binding affinity were created. We then studied the similarity of the rank lists in promoters of tissue-specific genes and detected candidates for TF-TF interaction by applying the conditional independence test for multiway contingency tables. Our candidates were validated by both known protein-protein interactions (PPIs) and the co-occurrence of their binding sites in promoters. We found that the known PPIs are significantly enriched in the group of our predicted TF-TF interactions (4-10 times of random expectation). In addition, we could find analogous TF-TF interactions in other species for some of our predictions.

Adaptation in allopolyploid Dactylorhiza: a story from beyond genetics

Ovidiu Paun, Richard Bateman, Michael Fay and Mark Chase

University of Vienna, Rennweg 14, 1030 Vienna, Austria

Epigenetic information includes heritable signals that modulate gene expression but are not encoded in the primary nucleotide sequence. We have studied natural epigenetic variation in three allotetraploid sibling orchid species that differ radically in geography/ecology. The epigenetic variation released by genome doubling has been restructured in species-specific patterns that reflect their recent evolutionary history, and have an impact on their ecology and evolution, hundreds of generations after their formation. Using two contrasting approaches that yielded largely congruent results, epigenome scans pinpointed epiloci under divergent selection that correlate with eco-environmental variables, mainly related to water availability and temperature. The stable epigenetic divergence in this group is largely responsible for persistent ecological differences, which then set the stage for species-specific genetic patterns to accumulate in response to further selection and/or drift. Our results strongly suggest a need to expand our current evolutionary framework to encompass a complementary epigenetic dimension when seeking to understand population processes that drive phenotypic evolution and adaptation.

HTSeq -- Analysing high-throughput sequencing data with Python

Paul Pyl and Simon Anders

European Molecular Biology Laboratory, Huber Group - Genome Biology Unit, Meyerhofstrasse 1, 69117, Heidelberg, Germany

When analysing high-throughput sequencing data there are many readily available tools for the "big" tasks such as alignment, variant-calling, assembly, peak-finding, and browsing. However, once a project deviates from standard workflows, custom scripts are needed.

We present HTSeq, a Python library for rapid development of such scripts.

HTSeq offers parsers for many common data formats as well as classes to represent that data in a manner independent of input file formats, and performant data structures for convenient access to data and metadata.

HTSeq has proven helpful for many different tasks, such as quality control, demultiplexing, removing adapter contamination and calculating coverage vectors.

We also used HTSeq to genotype reads in crossing experiments, study ChIP-Seq profiles and determine read-gene overlap for RNA-Seq data (including handling complex overlap scenarios).

HTSeq allows for rapid development and testing of complex analyses, and ensures reproducibility and re-usability.

SpeedGene -- Handling the data management needs of high throughput-sequencing data and GWAS: A C++ library for the fast and efficient storage of genetic data

Dandi Qiao

Harvard University, Unit 134, 199 Park Drive, 02215 Boston, U.S.

The most commonly used input format for family-based genetic data is the pedigree file-type. For high-throughput sequencing data or genome-wide data, pedigree files can reach sizes of several Tetra-bytes or hundreds of Gigabytes of disk space. Large file size results in wasting of disk space and loading time during analysis. We introduce here a library that includes an optimized "storage-and-load" algorithm that, depending on the minor allele frequency, chooses among three different compression-algorithms to minimize the disk space for storage. The library also provides functions for loading the compressed files and retrieving any original data. We show that our new algorithm performs better than the currently available compressed formats of pedigree file by several magnitudes. The compression factor of the algorithm can range from 16 to several hundreds. Also, it takes only few seconds/minutes to load an entire file with the library, and parallel processing of the dataset is supported.

Graphical Integration Tools for Genomic Data

Michael G. Schimek and Karl Kugler

Medical University of Graz, IMI, Auenbruggerplatz 2/V, 8036 Graz, Austria

Current genomic research produces large amounts of data from studies of varying size and quality, often addressing similar questions. Most recently, new methods have been developed that combine or integrate such findings (e.g. from public data repositories). Owing to the dimensionality and complexity of these data of interest, such methods are exploratory in nature and involve technical parameters, hence integrated findings cannot be unique. Here we introduce new graphical techniques that assist the users of these methods in the selection of parameter values and in the interpretation of obtained aggregation results. Their usage can reduce the work load of biological verification, improve the reliability of outcomes, and reduce the costs of planned experiments. The new graphical tools – soon available in TopKGraphics, part of the R package TopKLists - are applied to microarray meta-analysis problems.

Repeated significance tests for high-dimensional data

Sonja Zehetmayer, Martin Posch and Peter Bauer

Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria

It is well known that increasing the sample size and testing a single hypothesis repeatedly after non-significant results at unadjusted critical levels inflates the overall Type I error rate severely. Surprisingly, if a large number of hypotheses are tested controlling the False Discovery Rate, such ‘hunting for significance’ has asymptotically no impact on the error rate.

More specifically, if the sample size is increased for all hypotheses simultaneously and only the test at the final interim analysis determines which hypotheses are rejected, a data dependent increase of sample size does not affect the False Discovery Rate. This holds asymptotically (for an increasing number of hypotheses) for all scenarios but the global null hypothesis where all hypotheses are true. To control the False Discovery Rate also under the global null hypothesis, we consider stopping rules where early stopping is possible only if sufficiently many null hypotheses can be rejected.