Imperial College London

CISBIO

Abstracts: Oral Presentations

VENUE: Lecture Theatre 164
Skempton Building
Imperial College London
South Kensington Campus
London SW7 2AZ
http://www.theosysbio.bio.ic.ac.uk/masamb/

Contents

FRIDAY, 12th of April 2013

SESSION III: Next-Generation Sequencing **Pages 15 – 19**

9:00-9:20 *Discovery of protein binding patterns by joint modelling of ChIP-seq data*
Yanchun Bao, Veronica Vinciotti, Ernst Wit and Peter-Bram 't Hoen

9:20-9:40 *Approximate Inference for Transcript Quantification in RNA-Seq*
James Hensman, Panagiotis Papastamoulis, Peter Glaus, Antti Honkela, Neil Lawrence and Magnus Rattray

9:40-10:00 *Accelerating Integrative Modelling using GP-GPU Computing*
Sam Mason, Paul Kirk, Richard Savage, Faiz Sayyid and David Wild

10:00-10:20 *Accounting for technical noise in single-cell RNA-seq experiments*
Simon Anders, Philip Brennecke, Jong Kyoung Kim, John Marioni and Marcus Heisler

10:20-10:40 *NextGenMap*
Fritz J. Sedlazeck, Philipp Rescheneder and Arndt von Haeseler

SESSION IV: Systems Biology **Pages 20 – 24**

11:20-11:40 *Inference for single cell systems*
Sarah Filippi, Chris Barnes, Paul Kirk and Michael Stumpf

11:40-12:00 *Efficient inference of stochastic gene regulation models using fluorescence histograms and stochastic simulations*
Gabriele Lillacci and Mustafa Khammash

12:00-12:20 *Functional redundancy in the NF-κB signalling pathway*
Michal Komorowski, Michal Włodarczyk and Tomasz Lipniacki

12:20-12:40 *Gaussian process models of circadian rhythm*
Nicolas Durrande, James Hensman, Magnus Rattray and Neil Lawrence

12:40-13:00 *Parameter inference in complex biological systems using adaptive gradient matching with Gaussian processes and parallel tempering*
Benn Macdonald, Frank Dondelinger and Dirk Husmeier

SESSION V: Evolution **Pages 25 – 29**

14:20-14:40 *Direct estimation of the amino acid replacement matrix and phylogeny using rjMCMC*
Andrew Meade and Mark Pagel

14:40-15:00 *Stability-activity trade-off constrains the adaptive evolution of RubisCO*
Romain Studer, Pascal-Antoine Christin, Mark Williams and Christine Orengo

15:00-15:20 *POlymorphisms-aware phylogenetic MOdels*
Nicola De Maio, Christian Schlötterer and Carolin Kosiol

15:20-15:40 *A maximum likelihood approach for detecting coevolution in proteins*
David Talavera, Simon Lovell and Simon Whelan

15:40-16:00 *Gaussian process modelling of evolutionary time series*
Hande Topa, Agnes Jonas, Carolin Kosiol and Antti Honkela

THURSDAY, 11th of April 2013
SESSION I: Statistical Bioinformatics

13:00-13:20 *Clustering genes by phylogenetic similarity*
Kevin Gori, Christophe Dessimoz and Nick Goldman

Reconstructing the Tree of Life is a long-held aim of biology, but its structure is still uncertain, despite the abundance of data available from high-throughput sequencing. Due to genomic sequences having evolved through multiple processes of evolution - such as incomplete lineage sorting, lateral gene transfer or recombination - different segments of a genome can have distinct evolutionary histories. Given this, there is no reason to assume that all genes in a genomic dataset are congruent with a single phylogenetic tree. Therefore it is necessary to develop methods that incorporate incongruence.

We propose a system in which data can be described by multiple trees. Genes are partitioned into classes, with each class comprising genes that share similar phylogenetic histories, and a tree is calculated for each class. The division into classes is accomplished by first calculating a phylogenetic tree for each gene in the data set, and then clustering on a matrix of pairwise distances between the trees.

We assess the performance of a variety of tree-distance metrics and clustering methods at recovering the correct classes from data simulated from a known set of evolutionary histories. Several tree-distance metrics are investigated, combined with various clustering approaches. We find that the best combination is of geodesic distances with spectral clustering.

One particular challenge in our approach lies in choosing the number of classes. Using simulated and empirical data, we evaluate three approaches: parametric resampling, non-parametric resampling, and heuristic stopping criteria.

I will discuss two challenges for our future work. The first is in scaling our approach to large data sets. The second is to implement an optimisation procedure to improvement the assignment of genes into clusters, for example through an iterative, Expectation-Maximisation process. Finally, I will describe the biological questions our method can be applied to.

THURSDAY, 11th of April 2013
SESSION I: Statistical Bioinformatics

13:20-13:40 *Metabolite identification from liquid chromotography/MS data using a Bayesian modelling approach*
Rónán Daly, Joe Wandy, Simon Rogers and Rainer Breitling

The identification of molecules from mass spectra represents a significant bottleneck in the use of high-throughput mass spectrometry (MS) to routinely analyse the metabolic makeup of an organism. Typically, MS is preceded by a physical separation (e.g. by liquid chromatography) and data are collected as a series of spectra as molecules elute. It is well understood that molecules eluting at the same time (and with the same temporal peak shape) are likely to be related: e.g. they are different isotopomers of the same molecule or different adducts/fragments.

The presence of these related peaks can help in identifying molecules by, e.g., comparing the mass differences and intensity ratios with those theoretically computed for particular molecules and their isotopes and adducts. We have recently developed a Bayesian method that simultaneously groups related peaks (based on their elution time and/or shape) and assigns the groups to particular chemical formulas. However, formula identification is only one step towards metabolite identification due to the possibility of many isomers for some formulas.

In this work, we extend our model to incorporate freely available network data using the Bayesian approach introduced in [1] as well as predicted elution times using the models in [2]. By incorporating these additional forms of information, an extra level of identification is added to the model, that can distinguish metabolites with the same formula. We investigate the performance of this extended model for both synthetic and real LC/MS-derived metabolomics data.

[1] Rogers et al. (2009). Probabilistic assignment of formulas to mass peaks in metabolomics experiments. Bioinformatics, 25(4):512–518.

[2] Creek et al. (2011). Toward global metabolomics analysis with hydrophilic interaction liquid chromatography-mass spectrometry: Improved metabolite identification by retention time prediction. Analytical Chemistry, 83(22):8703–8710.

THURSDAY, 11th of April 2013
SESSION I: Statistical Bioinformatics

13:40-14:00 *Predicting protein beta-sheet contacts using a maximum entropy-based correlated mutation measure*
Nikolas Burkoff, Csilla Várnai and David Wild

Motivation: The problem of ab initio protein folding is one of the most difficult in modern computational biology. The prediction of residue contacts within a protein provides a more tractable immediate step. Recently introduced maximum entropy-based correlated mutation measures (CMMs), such as direct information, have been successful in predicting residue contacts. However, most correlated mutation studies focus on proteins that have large good-quality multiple sequence alignments (MSA) because the power of correlated mutation analysis falls as the size of the MSA decreases. However, even with small autogenerated MSAs, maximum entropy-based CMMs contain information. To make use of this information, in this article, we focus not on general residue contacts but contacts between residues in β-sheets. The strong constraints and prior knowledge associated with β-contacts are ideally suited for prediction using a method that incorporates an often noisy CMM.
Results: Using contrastive divergence, a statistical machine learning technique, we have calculated a maximum entropy-based CMM. We have integrated this measure with a new probabilistic model for β-contact prediction, which is used to predict both residue- and strand-level contacts. Using our model on a standard non-redundant dataset, we significantly outperform a 2D recurrent neural network architecture, achieving a 5% improvement in true positives at the 5% false-positive rate at the residue level. At the strand level, our approach is competitive with the state-of-the-art single methods achieving precision of 61.0% and recall of 55.4%, while not requiring residue solvent accessibility as an input.

Burkoff, N.S., Várnai, C and Wild, D.L., Predicting protein β-sheet contacts using a maximum entropy-based correlated mutation measure. Bioinformatics (2013) 29 (5) 580-587

THURSDAY, 11th of April 2013
SESSION I: Statistical Bioinformatics

14:00-14:20 *Going beyond static networks*
Thomas Thorne and Michael P.H. Stumpf

In the field of Systems Biology we are often faced with the task of performing inference that involves network structures, one particular example being the inference of gene regulatory network structures. We might be interested in changes to the network structure in a single time series, between differing conditions or even on an evolutionary time scale. To this end we employ Bayesian nonparametrics to provide the flexibility to allow for such changes in the network structure. By doing so we can infer these changes from the data, whilst also combining information from multiple sources to derive more accurate predictions.

THURSDAY, 11th of April 2013
SESSION I: Statistical Bioinformatics

14:20-14:40 *Decision forests on distance matrices for neuroimaging genetics studies*
Aaron Sim, Dimosthenis Tsagkrasoulis and Giovanni Montana

In genetic association studies there is a growing appreciation for the significant boost in power inherent in adopting quantitative phenotypic traits in place of categorical case-control indicators. In recent years, both the complexity and variety of such data have grown considerably. The classic example is in Imaging Genetics, where the availability of high-resolution neuroimaging data from multiple imaging modalities (e.g. MRI, PET) has made available multiple phenotypes ranging from multivariate volumetric measurements to complex brain connectivity networks for the search for the genetic bases of many neurological diseases.

In this investigation we propose and evaluate a methodology based on a generalisation of Random Forests – Random Forests on Distance Matrices (RFDM) – to utilise and integrate phenotypic data across multiple types for genetic association studies. In addition, we develop an information-theoretic measure for epistasis detection.

RFDM is built to accept responses of any representation. This is achieved via a representation-independent distance matrix. We obtain the distances between subjects either analytically using known metrics, or via supervised manifold-learning procedure. Genetic associations are determined from a proposed generalized information gain measure, and pairwise epistatic interactions from the conditional information measure.

We validate the methodology in two ways. First, applying RFDM to a cohort from the Alzheimer's Disease Neuroimaging Initiative (ADNI) public datasets, we recover several known genetic factors associated to Alzheimer's disease (AD). Second we perform simulations of several AD models. The population genetics data is simulated from the datasets of the International HapMap Project while MRI and brain connectivity data from ADNI. We show that RFDM detects both single-SNP and interacting causal factors with greater power, compared to both case-control studies and multivariate RF regression. With multiple endophenotypes, tests which integrate data via averaged distance matrices outperforms those that employ single datasets

THURSDAY, 11th of April 2013
SESSION II: Computational Cell Biology

15:20-15:40 *The Fidelity of Dynamic Signaling by Noisy Biomolecular Networks*
Clive Bowsher, Margaritis Voliotis and Peter Swain

Cells live in changing, dynamic environments. To understand cellular decision-making, we must therefore understand how fluctuating inputs are processed by noisy biomolecular networks. Here we present a general methodology for analyzing the fidelity with which different statistics of a fluctuating input are represented, or encoded, in the output of a signaling system over time. We identify two orthogonal sources of error that corrupt perfect representation of the signal: dynamical error, which occurs when the network responds on average to other features of the input trajectory as well as to the signal of interest, and mechanistic error, which occurs because biochemical reactions comprising the signaling mechanism are stochastic. Trade-offs between these two errors can determine the system's fidelity. By developing mathematical approaches to derive dynamics conditional on input trajectories we can show, for example, that increased biochemical noise (mechanistic error) can improve fidelity and that both negative and positive feedback degrade fidelity, for standard models of genetic autoregulation. For a group of cells, the fidelity of the collective output exceeds that of an individual cell and negative feedback then typically becomes beneficial. We can also predict the dynamic signal for which a given system has highest fidelity and, conversely, how to modify the network design to maximize fidelity for a given dynamic signal. Our approach is general, has applications to both systems and synthetic biology, and will help underpin studies of cellular behavior in natural, dynamic environments.

THURSDAY, 11th of April 2013
SESSION II: Computational Cell Biology

15:40-16:00 *Analysis of mitochondrial genetic variation due to bottlenecking and segregation*
Iain Johnston

Mitochondrial DNA (mtDNA) is subject to a high mutation rate, and many diseases result from buildup of mutational damage in mtDNA. The stochastic process of mtDNA inheritance, involving bottlenecking (varying the number of mtDNAs per cell) and segregation (increasing the cell-to-cell variance of mutant load) is currently poorly understood, severely limiting clinical ability to predict the inheritance of mitochondrial disease and the behaviour of embryos in which mtDNA content has been modified (for example, in nuclear transplantation as a means of preventing mitochondrial disease). We perform the first quantitative survey of proposed bottlenecking mechanisms, using parametric inference to identify the current experimental levels of support for different proposed mechanisms. We introduce and explore a physically motivated, analytically solvable stochastic mathematical model for bottlenecking, compatible with experimental data. Our model facilitates predictions about mtDNA content during and after development, identifying important controlling variables and suggesting optimal experimental schemes to address medical questions, placing the understanding of bottlenecking on sound mathematical foundations.

THURSDAY, 11th of April 2013
SESSION II: Computational Cell Biology

16:00-16:20 *Mathematical Modelling of the Unfolded Protein Response*
Kamil Erguler, Myrtani Pieri, Charalambos Stefanou and Constantinos Deltas

The Unfolded Protein Response (UPR) is a major signalling cascade for quality control of protein folding. The cascade consists a complex circuitry of three types of membrane receptors and their downstream pathways operating in concert in order to deliver the right response at the right time. The right response refers to a range of behaviour from adaptive to destructive depending on the degree and the duration of unfolded protein accumulation in the ER.

Although a good deal about the workings of specific sub-processes of the cascade has been learnt from experimental studies, the decision mechanism at a system level has not been studied to this day. By making use of the extensive literature, we developed the most comprehensive mechanistic model available so far for studying collectively a large portion of the cascade.

The analysis revealed three distinct activity states at which the system operates to manage adaptation, tolerance, and the initiation of apoptosis. The decision to adapt or destruct can, therefore, be understood as a dynamic process where the balance between the stress and the folding capacity plays a pivotal role. The model demonstrated for the first time that the UPR is capable of generating oscillations in translation attenuation and the apoptotic signals, and a Bayesian sensitivity analysis identified a set of parameters controlling this behaviour.

Putting together what is known about a system in a structured mathematical model proved to be an excellent way to question the existing knowledge while proposing novel and interesting hypotheses to be tested. Adopting a stochastic framework facilitated validation by augmenting the information content of the predictions.

In summary, we present the step by step construction of an informed hypothesis in the form of a mathematical model, and a valuable collection of initial attempts to experimentally validate some of the most challenging predictions.

THURSDAY, 11th of April 2013
SESSION II: Computational Cell Biology

16:20-16:40 *Investigating the molecular mechanism behind the broad classification of clear cell renal carcinoma*
Tammy Cheng, Sakshi Gulati, Rudi Agius, Marco Gerlinger, Charles Swanton and Paul Bates

Clear cell renal cell carcinoma (ccRCC) is the predominant histological subtype of kidney cancer (comprising 70-75% cases) and is currently treated as a single entity. However, large-scale gene expression analysis of clear cell kidney cancer has suggested that there are likely to be at least two molecular subgroups, named type A and B, which denotes the heterogeneity of ccRCC. With the intention of obtaining more insights into the molecular heterogeneity between type A and B subgroups, we analyse microarray data to identify differences in molecular mechanisms for the two subgroups. Identifying these genes will help us to define the prognostic implications of these two subtypes and improve our ability to predict sensitivity and resistance to ccRCC tumour drugs.

We have conducted high-dimensional multivariate statistical analyses on two independent microarray data sets to identify small clusters of genes that serve as effective signals for differentiating type A and B ccRCC patients. This consequently generated a list of susceptible genes that are likely to contribute to the differentiation between the two subtypes. Based on the susceptible genes, we have identified signaling pathways and biological functions that are likely to define the separation between the two groups. Currently we are building Boolean models to investigate further the impact of the susceptible genes on regulating cell proliferation. The ultimate goal of our study is to build biological models based on the identified pathways and biological functions to explain the level of heterogeneity in ccRCC.

THURSDAY, 11th of April 2013
SESSION II: Computational Cell Biology

16:40-17:00 *Bayesian Networks Reveal Interaction between NF-kB and Cellular Context*
Heba Sailem, Chris Bakal and Julia Sero

Automated microscopy is a powerful tool that enable extracting multi-dimensional phenotypic data from cellular images. However, there is still a desperate need for statistical and computational tools to analyse the tremendously rich data sets that can be generated through the imaging of individual cells. In this study we have used automated microscopy and advanced image analysis to model a causal relationship between cell shape and nuclear localisation of the transcription factor Nf-κB in 19 breast cancer cell lines.

NFkB is activated by inflammatory cytokines, such as TNFα, and regulates proliferation, survival, apoptosis and the cellular response to stress. Dysregulation of NF-κB contributes to a number of pathologies, for example cancer and cardiovascular disease. However, a classical question in biology is how can the activation of a single protein, such as Nf-κB, lead to strikingly different outcomes?

We hypothesised that the shape of cells could be an important determinant of the outcome of Nf-κB activation. Therefore, we generated a novel data set that describes single-cell morphology, local cell density and subcellular localisation of NF-κB in 19 cell lines before and after stimulation with TNFα We used Bayesian dependency modelling and determined a causal relationship between NFkB and cellular measurements. Our data is continuous and so we used regression trees to determine the best split when modelling the data. The resulting models revealed that NF-κB localisation is highly dependent on aspects of cell shape as well as population context. We experimentally validated these models and demonstrated that changing the cellular density affected NFkB localisation. We also perturbed cell shape using cytoskeletal drugs that resulted in changes in NFkB level. Using multivariate regression models we predicted the change in NFkB localisation in 11 cell lines based only on shape changes following 15 different drug treatments. We demonstrated the same pattern of dependency for the YAP transcription factor. We propose that using Bayesian methods is a useful tool that exploits the variability in cellular populations to gain a systems level understanding of the interaction between cellular phenotypes and signalling.

FRIDAY, 12th of April 2013
SESSION III: Next-Generation Sequencing

9:00-9:20 *Discovery of protein binding patterns by joint modelling of ChIP-seq data*
Yanchun Bao, Veronica Vinciotti, Ernst Wit and Peter-Bram 't Hoen

An important biological question is the one of detecting the regions in the genome bound by chromatin modifiers or transcription factors, as these give insight into the mechanism of gene regulation. ChIP sequencing (ChIP-seq) is a biological method to detect these, by generating sequence reads at the positions bound by a transcription factor. When data is collected on more than one protein, the interest is also in the discovery of regions that are uniquely bound by a protein. In this talk, we present a Markov random field model for the modelling of ChIP-seq data from multiple experiments, including technical and biological experiments as well as experiments on different proteins. The statistical model accounts for the fact the different antibodies used for experiments on different proteins are associated with different levels of efficiency. We compare our model with existing ones on a simulation study and on real ChIP-seq data on two histone acetyltransferases, p300 and CBP

FRIDAY, 12th of April 2013
SESSION III: Next-Generation Sequencing

9:20-9:40 *Approximate Inference for Transcript Quantification in RNA-Seq*
James Hensman, Panagiotis Papastamoulis, Peter Glaus, Antti Honkela, Neil Lawrence and Magnus Rattray

High throughput sequencing of RNA (RNA-seq) allows for the investigation of expression at the isoform level. Accurate and robust quantification of transcript abundance requires methods that account for uncertainties including biases and noise in the sequencing process, and the potential alignment of any single read to several isoform variants.

Probabilistic models are a natural way to deal with such problems, and several approaches have been proposed (RSEM [1], BitSeq [2], MISO[3]). The solution of the probabilistic model involves a large number of latent (hidden) variables, which detail the connections between isoforms and reads. A Bayesian approach via Gibbs sampling [2,3] shows significant benefits, but the associated computational burden limits its usefulness is practise.

Here we demonstrate an approximate inference scheme for the BitSeq model, based on recent advances in variational Bayes [4]. The method is an order of magnitude faster than a straightforward application of VB, which is itself an order of magnitude faster than a Gibbs sampler. Our proposed method is also faster than EM, whilst retaining some benefits of the Bayesian approach. We show that our variational scheme is effective in estimating the mean of the true posterior.

Since our approximate posterior tends to underestimate the variance, we develop a further variational scheme by integrating out the latent variables and considering a richer family of distributions. In this second scheme, we prove that we can always find a solution that is closer to the true posterior than the first, and we demonstrate empirically that we are able to approximate the true posterior closely.

Finally, we demonstrate that our approximate inference scheme gives excellent performance in detecting differentially expressed transcripts. We discuss how the additional computational performance given by the approximate inference scheme might make way for more complex statistical models, where aspects such a population structure are accounted for at the transcript quantification stage.

[1] B Li and C.N Dewey 2011, BMC bioinformatics 12(1)
[2] P.Glaus, A. Honkela and M. Rattray 2012, Bioinformatics 28(13)
[3] Y. Katz, E. T. Wang, E. M. Airoldi, C. B. Burge. Nature Methods 2010 (7, 1009-1015)
[4] J. Hensman, M. Rattrray and N. D. Lawrence, NIPS 2012

FRIDAY, 12th of April 2013
SESSION III: Next-Generation Sequencing

9:40-10:00 *Accelerating Integrative Modelling using GP-GPU Computing*
Sam Mason, Paul Kirk, Richard Savage, Faiz Sayyid and David Wild

In the forthcoming era of personal genomic medicine, genome sequences and other forms of high-throughput data such as gene expression, alternative splicing, DNA methylation, histone acetylation, and protein abundances will be routinely measured for large numbers of people. Modern high-throughput technologies generate a broad array of different data types, providing distinct—yet often complementary—information. To realise the promise of these technological developments, it is essential to develop statistical and computational methodology to allow the integrated modelling of information from these different experimental platforms. A limiting factor of data integration on a genome-wide scale is the trade off between model complexity and runtime, with methods often having to balance these conflicting requirements.

We describe a Bayesian method for the unsupervised integrative modelling of multiple datasets, which we refer to as MDI (Multiple Dataset Integration). We present ongoing work on a Markov Chain Monte Carlo (MCMC) sampler exploiting General Purpose computing on Graphical Processing Units (GP-GPUs) to provide to over three orders of magnitude speedup relative to our previously published implementation [1]. Furthermore, issues related to the implementation of high-performance MCMC code on massively parallel GP-GPUs will be considered.

[1] Kirk, P., Griffin, J.E., Savage, R.S., Ghahramani, Z. and Wild, D.L. (2012)
Bayesian Correlated Clustering to Integrate Multiple Datasets.
Bioinformatics. doi:10.1093/bioinformatics/bts595

FRIDAY, 12th of April 2013
SESSION III: Next-Generation Sequencing

10:00-10:20 *Accounting for technical noise in single-cell RNA-seq experiments*
Simon Anders, Philip Brennecke, Jong Kyoung Kim, John Marioni and Marcus Heisler

Recent studies have demonstrated the feasibility of using RNA-Seq to investigate the transcriptomes of single cells. These techniques promise valuable insights into transcriptional regulation, not only from comparing different cell types but also from studying the variability within a population of seemingly homogeneous cells. However, despite the high levels of technical noise present in such experiments, a quantitative statistical method to distinguish true biological variability from technical noise has not been established.

To address this issue, we have developed a general method that enables to assess the statistical significance of observed cell-to-cell variability in expression strength on a gene-by-gene basis. The method relies on judicious use of technical spike-ins gain information on the strength of technical noise across the dynamic range, which is the used to fit a noise model to be used in inference.

Using two distinct cell types from roots of Arabidopsis, we demonstrate our method by identifying genes that vary in their expression substantially above noise levels between otherwise similar cells. We find that many highly variable genes co-vary in their expression between cells and are enriched in functional categories. This provides a biological snapshot of each individual cell's regulatory state and suggests examples of co-regulation, thus providing a novel approach to gaining biological insights from single-cell transcriptomics.

FRIDAY, 12th of April 2013
SESSION III: Next-Generation Sequencing

10:20-10:40 *NextGenMap*
Fritz J. Sedlazeck, Philipp Rescheneder and Arndt von Haeseler

Keeping pace with the rapid developments of high throughput sequencing technologies constantly challenges the performance of bioinformatics tools. Here we report NextGenMap, a fast and accurate read aligner. NextGenMap aligns reliably reads to a reference genome even when the sequence difference between target and reference is large. NextGenMap efficiently utilizes the available hardware by exploiting multi-core CPUs as well as graphic cards (GPUs). NextGenMap estimates the most important mapping parameters from the input, and thus handles efficiently any read data independent of read length and sequencing technology. We show that NextGenMap outperforms current mapping methods with respect to running time and to the number of correctly mapped reads.

Availability:
NextGenMap source code and documentation are available at: http://cibiv.github.com/NextGenMap/

FRIDAY, 12th of April 2013
SESSION IV: Systems Biology

11:20-11:40 *Inference for single cell systems: origins of extrinsic noise*
Sarah Filippi, Chris Barnes, Paul Kirk and Michael Stumpf

At the molecular level every cell is unique. Differences between cells of the same type can be pronounced, and these differences can have profound biological and biomedical implications. The mechanisms driving this variability can be divided into two classes. Within-cell variability may arise from stochasticity in the biochemical reactions such as gene expression, protein localization, post-translational modification and formation of protein complexes. This source of variability is often called intrinsic noise. In addition, cell-to-cell fluctuations in biochemical reaction rates and other biophysical parameters can lead to so-called extrinsic noise. Given the growing abundance of 'omics data resolved at the single-cell level, it is becoming increasingly important to include these sources of noise in our models. Here we present a general modelling and inference methodology for single-cell data that explicitly takes into account extrinsic noise. We define hyperparameters to guide the variation of biophysical parameters among the cell population and use the unscented transform to derive a tractable likelihood function. Using quantitative image cytometry (QIC) single-cell proteomics data, we apply our methodology in order to study MEK/ERK phosphorylation dynamics.

FRIDAY, 12th of April 2013
SESSION IV: Systems Biology

11:40-12:00 *Efficient inference of stochastic gene regulation models using fluorescence histograms and stochastic simulations*
Gabriele Lillacci and Mustafa Khammash

In the noisy cellular environment, stochastic fluctuations at the molecular level manifest as cell-cell variability at the population level. Ongoing progress in single-cell measurements is producing high-throughput data sets for an increasingly accurate characterization of such variability. Consequently, stochastic gene regulation models have gained popularity as tools in computational biology, since they can capture the dynamics of the variability and extract information from it.
Their use, however, is still limited by issues of computational complexity. Even when network inference is able to reveal the key players in the process under study and their interactions, building a complete model requires overcoming a critical challenge: the determination of the many unknown parameters that will inevitably appear in it. These are numbers such as production and degradation rates, binding affinities, and so forth, which are very difficult to measure directly.
We describe a novel method based on approximate Bayesian computation (ABC) for parameter inference in stochastic gene regulatory networks using time-dependent flow cytometry measurements, referred to as INSIGHT. By introducing a suitable way of comparing experimental and model-generated fluorescence histograms, our proposed approach alleviates many of the limitations of the existing techniques, particularly in terms of computational cost.
We test our method on flow cytometry measurements of a synthetic gene network in E. coli, and we find that a detailed mechanistic model of this system can be estimated with high accuracy and high efficiency. Furthermore, we introduce the notion of Mismatch Index, which measures the fidelity of a given model in reproducing experimental observations, and we show how our proposed technique offers a basis to evaluate the quality of different models of the same process.

FRIDAY, 12th of April 2013
SESSION IV: Systems Biology

12:00-12:20 *Functional redundancy in the NF-κB signalling pathway*
Michal Komorowski, Michal Włodarczyk and Tomasz Lipniacki

The ability to represent intracellular biochemical dynamics via deterministic and stochastic modelling is one of the crucial components to move biological sciences in the observe-predict-control-design knowledge ladder. Compared to the engineering or physics problems, dynamical models in quantitative biology typically dependent on a relatively large number of parameters. Therefore, the relationship between model parameters and dynamics is often prohibitively difficult to determine. We developed a method to depict the input-output relationship for multi-parametric stochastic and deterministic models via information-theoretic quantification of similarity between model parameters and modules. Identification of most information-theoretically orthogonal biological components, provided mathematical language to precisely communicate and visualise compensation like phenomena such as biological robustness, sloppiness and statistical non-identifiability. A comprehensive analysis of the multi-parameter NF-κB signalling pathway demonstrates that the information-theoretic similarity reflects a topological structure of the network. Examination of the currently available experimental data on this system reveals the number of identifiable parameters and suggests informative experimental protocols.

FRIDAY, 12th of April 2013
SESSION IV: Systems Biology

12:20-12:40 *Gaussian process models of circadian rhythm*
Nicolas Durrande, James Hensman, Magnus Rattray and Neil Lawrence

Discovery and detection of genes with a periodic expression is of particular interest for inferring the gene network surrounding the circadian clock. We present a Gaussian process model of gene expression from microarray and next generation sequencing data, which elegantly deals with the small number of available time points, estimates the present periodicity and balances model complexity and fit under the probabilistic paradigm.

The novelty in our approach lies in the decomposition of the Gaussian process covariance function into a sum of periodic and aperiodic sub-covariances. This is obtained by extracting the subspace generated by the Fourier basis - and its orthogonal complement - from the RKHS associated with the Gaussian process covariance. This results in additional covariance function parameters which describe a sliding scale between priors over periodic and aperiodic functions.

Considering each gene in turn, we extract from the posterior Gaussian process distribution a periodicity statistic based on the variability of the sub-models, thus describing the periodicity. The power of the model lies in its ability to distinguish between high frequency content and noise in a probabilistic fashion, in contrast to existing models which limit high frequency components in order to prevent overfitting.

We illustrate the method on microarray data from several contemporary studies. The model is capable of detecting periodic signals which are more complex in nature (such as periodic spikes, triangles waves and other non sinusoidal patterns) than existing methodologies.

In further work, we incorporate our methodology into a Gaussian process mixture model. In this model, genes are clustered using a GP prior for the mean of each cluster. We specify periodic GP priors for the mean function, and allow each gene within a cluster to vary from the mean by a further Gaussian process function. Thus we are capable of clustering genes based on the periodic component of the signals alone.

FRIDAY, 12th of April 2013
SESSION IV: Systems Biology

12:40-13:00 *Parameter inference in complex biological systems using adaptive gradient matching with Gaussian processes and parallel tempering*
Benn Macdonald, Frank Dondelinger and Dirk Husmeier

Parameter inference in mathematical models of complex biological systems, expressed as coupled ordinary differential equations (ODEs), is a challenging problem. The systems depend on chemical kinetic parameters, which usually cannot all be measured and must be inferred from the data. Conventional methods using Markov Chain Monte Carlo (MCMC) sampling tend to involve integrating the system of ODEs at each iterative step, to see how well the sampled parameters correspond with the data. However, the computational costs associated with repeatedly solving the ODEs are often staggering, making many techniques impractical. Therefore, aimed at reducing this cost, new concepts using gradient matching have been proposed. These approaches generally work by first smoothing the signal (to avoid modelling the observational noise), then comparing the gradients from the resulting interpolant with those predicted from the ODEs. Our work, using a Bayesian approach, combines current adaptive gradient matching (AGM) techniques, using Gaussian process interpolation, with a parallel tempering scheme. The AGM allows the sampled parameters of our ODEs to reshape our interpolant, proceeding to less mismatch between our gradients. As well as tempering our posterior across parallel MCMC chains, we also temper the mismatch hyperparameter that governs the difference between the gradients. In this way, we are able to have the gradients from our interpolant closely match those from the ODEs, for chains closer to our target posterior. We use 2 ODE systems to assess our new technique: a simple model depicting neuron firing (Fitz-Hugh Nagumo) and a model of the behaviour of autocatalytic reactions (Lotka-Volterra). We present a comparative evaluation with other methods that are representative of the current state of the art.

FRIDAY, 12th of April 2013
SESSION V: Evolution

14:20-14:40 *Direct estimation of the amino acid replacement matrix and phylogeny using reversible-jump Markov Chain Monte Carlo methods*
Andrew Meade and Mark Pagel

Replacement matrices, such as BLOSUM, PAM, WAG and LG, are at the heart of some of the most popular bioinformatics and computational biology tools in use today (Blast, Fasta, Clustal, Muscle), as well as being used in phylogenetic inference. Due to their size and complexity, replacement matrices are typically estimated from large molecular databases and the resulting fixed matrices are applied to a wide range of proteins.

Here we present a novel reversible-jump MCMC (RJ MCMC) method to estimate amino acid replacement matrices directly from the data. RJ-MCMC is used to group replacement rates which are statistically indistinguishable into a smaller number of distinct parameters, thereby often greatly reducing the dimensionality of the problem and allowing the remaining parameters to be estimated, along with the phylogeny, while simultaneously accounting for uncertainty in.

Applying the RJ MCMC method to simulated data sets with 1, 5 and 10 unique replacement rates shows that the method can accurately identify the correct model of evolution and estimate the replacement rates. Simulating data using the WAG model shows that the RJ MCMC method performs as well as a complex model of evolution, both in terms of likelihood and estimating parameters. The RJ-MCMC method achieved significant likelihood improvements over traditional replacement matrixes, when applied to four published datasets, and led to changes in branch lengths and topologies. The estimated replacement matrices from published data sets showed significant deviations from the traditional replacement matrices, suggesting that evolutionary processes vary significantly between proteins.

FRIDAY, 12th of April 2013
SESSION V: Evolution

14:40-15:00 *Stability-activity trade-off constrains the adaptive evolution of RubisCO*
Romain Studer, Pascal-Antoine Christin, Mark Williams and Christine Orengo

Evolution of proteins is influenced by natural selection, controlling the fixation rate of amino acid changes. This rate is accelerated by positive selection during adaptation to environmental changes. A well-known case of such adaptation is the ribulose-1,5-bisphosphate carboxylase (RubisCO), the enzyme responsible for fixation of CO2 during photosynthesis. In flowering plants (angiosperms), two forms exist, the C3 and the C4, the latter being faster in terms of catalytic activity. The C4 forms result from convergent evolution in multiple clades, with substitutions at a small subset of sites under positive selection. The physicochemical forces behind these evolutionary changes remain unknown.

Using a phylogenetic framework and homology modelling, we reconstructed in-silico ancestral sequences and associated 3D structures. We were able to track precisely the past evolutionary trajectories, by identifying mutations on each branch of the phylogeny and evaluating the stability effect of these mutations. A higher number of slightly destabilising mutations were observed at the base of C4 clades, with subsequent stabilising mutations to restore the global stability. Some of these mutations are buried in the structure but close to the loop that gives access to the enzymatic cavity. Others are in interface regions and may be having an impact on activity, via an allosteric effect. These results show that RubisCO evolution is constrained by a “stability-activity trade-off”.

FRIDAY, 12th of April 2013
SESSION V: Evolution

15:00-15:20 *POlymorphisms-aware phylogenetic MOdels*
Nicola De Maio, Christian Schlötterer and Carolin Kosiol

Comparative analysis of genomes of related species, and of different individuals of the same species, can reveal adaptive trends in the history of the considered taxa, as well as show intensity and genomic variation of evolutionary patterns of species that undergo speciation. However, these intra and interspecific data also bring new challenges, such as the presence of incomplete lineage sorting and ancestral shared polymorphisms. Previous methods for genome-scale data of within and between-species diversity are mostly based on the coalescent process, therefore are restricted to very few populations and cannot handle selection.

Here, we present a new method called POlymorphisms-aware phylogenetic MOdel (PoMo). It is a phylogenetic Markov model with states representing fixed alleles as well as polymorphisms at different allele frequencies. A substitution is hereby modeled through a mutational event followed by a gradual fixation. Polymorphisms can either be observed in the present (tips of the phylogeny) or be ancestral (present at inner nodes). With this approach, we naturally account for incomplete lineage sorting and shared ancestral polymorphisms. Our method can accurately and time-efficiently estimate the parameters describing evolutionary patterns for phylogenetic trees of any shape (species trees, population trees, or any combination of those). Furthermore, we are able to disentangle the contribution of mutation rates and fixation biases in shaping substitution patterns.

We apply PoMo to genome-wide synonymous sites alignments of human, chimpanzee, and two orangutan species. From each taxon, we included data from several individuals. We present accurate estimates of mutation rates and GC-biased gene conversion (gBGC) in great apes. We also find that both mutation rates and gBGC vary with GC content, resulting in the well-known differences in substitution rates between regions with different GC content. Finally, our results are consistent with directional selection acting on coding sequences in relation to exonic splicing enhancers.

FRIDAY, 12th of April 2013
SESSION V: Evolution

15:20-15:40 *A maximum likelihood approach for detecting coevolution in proteins*
David Talavera, Simon Lovell and Simon Whelan

Detecting coevolving residues within or between proteins is a major computational challenge, with applications in molecular evolution, structural biology, and functional bioinformatics. It is assumed that coevolution mostly affects neighboring residues. The mutual information approach has been widely used, but has some significant drawbacks, namely: i) it requires a large number of sequences; ii) it lacks an explicit model for describing coevolution and cannot account for the tree structure inherent in biological data; and iii) it is not easy to integrate with statistical methodology. There have been several attempts to build models of (molecular) coevolution based on biologically and statistically sound reasoning. Although more conceptually sound than MI, currently these approaches can only identify correlations between changes in residues or identify coevolution between pairs of binary characters. Here we present a maximum likelihood approach for testing specific hypotheses about molecular coevolution. Our model is based on a propensity matrix that provides an explicit description on the tendency for pairs of amino acids to co-occur. The model parameterization uses this matrix to adjust the equilibrium frequency of residue pairs over evolution. This approach allows us to test the contribution of different physicochemical features on coevolution of pairs of residues; e.g., identifying pairs that are coevolving due to their charge, size or polarity. A series of likelihood ratio tests can then be used to identify pairwise interactions within a protein, with multiple testing corrected using a false discovery rate adjustment.
We applied this model to the study of coevolution of trypsin homologues, using 7 different coevolution propensity matrices. In this way, we identified sets of coevolving residues and the diverse reasons they are likely coevolving. Some sites seem to have an unrealistic tendency to coevolve with many other sites. However, filtering them out results in enrichment for coevolving pairs in close proximity.

FRIDAY, 12th of April 2013
SESSION V: Evolution

15:40-16:00 *Gaussian process modelling of evolutionary time series*
Hande Topa, Agnes Jonas, Carolin Kosiol and Antti Honkela

In population genetics, changes in the allele frequencies play an important role in the evolutionary process of the species. However, distinguishing the alleles that are changing under selection from those just displaying genetic drift is challenging due to the large number of false positives.

Here we present a Gaussian Process (GP) approach to model the evolutionary time series data. First we infer the Single Nucleotide Polymorphism (SNP) frequencies and their observation noise variances from sequencing data under a Poisson-Gamma model. Then, we fit time-dependent and time-independent GP models to logistic transformed frequencies, while incorporating the inferred noise variances in the models. Finally, we compute the Bayes Factors between time-dependent and time-independent GP models and rank the SNPs according to their Bayes Factors.

We compare the performances of our method and Cochran-Mantel-Haenszel statistical test on a simulated dataset which mimics the sequencing data of Drosophila, with four replicates along eight generations. Results indicate that our method outperforms with a higher precision at the same recall rate and making use of the inferred noise variance in the GP models helps to decrease the number of false positives.