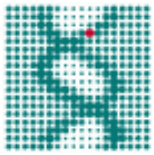


MASAMB 2012 – Talks' abstracts

Modelling epistasis in protein evolution: The evolutionary Stokes shift

Richard Goldstein, David Pollock

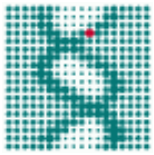
Standard models of protein evolution generally assume each location evolves independently, although it is well appreciated that substitution rates in a protein are influenced by changes in the amino acids at other locations. Generating more accurate but computationally tractable models of the evolutionary process will require a better understanding of the properties of such interactions. We simulate the evolution of a purple acid phosphatase with a fitness criterion based on thermodynamic stability. We find, contrary to standard models, that the amino acid propensities at a given location vary widely due to these epistatic interactions, with the changes in propensities occurring at a wide range of timescales. We also observe that, following a substitution at a given location, the rest of the protein undergoes co-evolution that increases the propensity of this new amino acid. As a result, subsequent substitutions at this location will be, in general, less favourable. This has similarities to the phenomenon known as the "Stokes shift" in spectroscopy, where vibrational relaxation results in stabilisation of the current electronic state relative to other electronic states, increasing the energy change during electronic transitions. Because of this similarity, we call this new evolutionary process the "evolutionary Stokes shift". The results of the simulations give excellent agreement with the distribution of changes in stabilities calculated for real proteins that also show back substitutions becoming less favorable over time. The observation of an evolutionary Stokes shift has important consequences for our understanding and modelling of protein evolution.

Effect of genomic population dynamics on the genealogy of insertion sequences

Jaime Iranzo, Susanna Manrubia

The ubiquity, diversity and persistence of mobile genomic elements constitutes an intriguing fact ever since the first genomes were sequenced. Formerly considered as selfish DNA able to persist at the cost of their host genomes, mobile elements are at present known to provide benefits by increasing genomic plasticity and facilitating adaptation. The puzzling role that these elements play in the genomes has fuelled long debates on which are the processes determining their abundance and distribution. In this contribution we focus on insertion sequences (IS's), that can be considered as the simplest transposable elements in prokaryotic genomes: an IS consists of a single gene coding for a transposase, usually flanked by inverted repeats. IS elements enter genomes through horizontal gene transfer (HGT), then they experience duplications and deletions that determine their intra-genomic abundance. At a higher level, the persistence of IS's in a population may be regulated by natural selection, although the contribution of an IS to organismal fitness (and even its sign) remains an open question.

Theoretical models are useful tools to understand the diversity and distribution of IS's since they allow to disentangle the effects of natural selection, HGT, duplications and deletions. Moreover, the increasing amount of sequencing data is allowing to test such models and to extract new knowledge from them. One particular aspect that has become tractable is the genealogy of IS elements in a population. The study of genealogies on IS elements is relevant as it is more discriminative than the classical study of abundance distributions when it comes to distinguish among HGT, duplications and deletions. Specifically we have developed a theoretical framework that considers the intra-genomic genealogy of IS (driven by duplications and deletions) nested into a population genealogy of genomes (driven by natural selection and neutral drift). Within this framework, HGT becomes apparent by merging the intra-genomic genealogical trees. The resultant model predicts the abundance distribution, diversity and genealogical structure of IS elements in different scenarios. Finally, comparison with genomic data allows for a quantitative estimation of the effect that different processes exert over the IS distribution and asks for a reanalysis of the previously assumed role of natural selection.



MASAMB 2012 – Talks' abstracts

The evolution of GC content in avian genomes

Carina Mugal, Peter Arndt, Hans Ellegren

The genomes of many vertebrates, including mammals and birds, show a characteristic heterogeneous distribution of the local GC content, the so-called isochore structure of the genome. By now, the origin of isochores has been explained via the mechanism of GC-biased gene conversion (gBGC), i.e. short-scale, unidirectional exchanges between homologous chromosomes in the neighborhood of recombination-initiating double-strand breaks, where AT/GC heterozygotes produce more GC- than AT-gametes. Whereas the isochore structure is declining in many mammalian genomes, the heterogeneity in GC content is being reinforced in the avian genome. Despite this discrepancy, examinations of individual mammalian and avian substitution frequencies, are both consistent with the gBGC model of isochore evolution. However, a negative correlation between the local substitution rate and the local recombination rate present in the avian genome appears to be inconsistent with the gBGC model of evolution. Hence, it seems important to consider along with gBGC other consequences of recombination on the origin of mutations and their probability of fixation, as well as to take relationships of recombination rate to other genomic features into account.

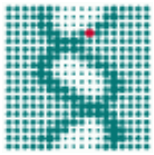
In order to investigate the negative correlation between the local substitution rate and the local recombination rate, we developed a minimal analytical model to describe the substitution pattern found in the avian genome and compared simulated data to observed data. This analysis sheds light into which other genomic features and aspects of recombination impact on the local substitution pattern and the evolution of GC content in the avian genome. The results indicate that the local GC content itself, either directly or indirectly via interrelations to other genomic features, has an impact on the local substitution pattern by affecting the rate of mutation. Further, we suggest that this phenomenon is specific to avian genomes due to their unusually slow rate of chromosomal evolution, where many chromosomes have remained more or less intact during avian evolution. Because of this, interrelations between the local GC content and other genomic features are more pronounced in the avian genome.

Statistical evaluation of tandem repeat detection algorithms and tandem repeat classification

Elke Schaper, Andrey Kajava, Alain Hauser, Maria Anisimova

Tandem repeats represent one of the most prevalent features of genomic sequences. Due to their abundance and functional significance, a plethora of tandem repeat detection tools has been devised over the last two decades. Despite the long-standing interest, tandem repeat detection is far from being resolved. Our large-scale tests reveal that current detectors produce highly incoherent inferences of tandem repeats, reflecting characteristics of the underlying algorithms rather than the true distribution of tandem repeats in genomic data. Our simulations show that the power of detecting tandem repeats depends on the degree of their divergence, as well as on basic repeat characteristics - the length of the minimal repeat unit and the number of units in tandem. Both accuracy and predictive power vary significantly between the different detectors.

To reconcile the often-conflicting predictions of current algorithms, here we evaluate several statistical criteria for measuring the quality of predicted repeat units. In particular, we propose a model based phylogenetic classifier. Applied in conjunction with the state-of the art detectors, our statistical classification scheme for inferred repeats allows to filter out false positive predictions. Since different algorithms appear to "specialize" at predicting repeats with certain properties, we advise applying multiple detectors with subsequent classification to obtain the most complete set of true repeats. Finally, we show how our findings will support the development of tandem repeat detectors in future.



MASAMB 2012 – Talks' abstracts

Population dynamics of normal and leukaemia stem cells in the haematopoietic stem cell niche

Adam McLean, Cristina Lo Celso, Michael Stumpf

Haematopoietic stem cells are responsible for maintaining immune cells, red blood cells and platelets throughout life, and their location within the niche is crucial for correct functioning. Increasingly, data that provide spatio-temporal information about this niche and its constituent cells in vivo allow us to investigate hypotheses about the importance of the makeup of this niche for stem cells and haematopoiesis. Here we propose that in the case of leukaemia, healthy haematopoietic species must compete with their cancerous counterparts for the niche's resources, and thus we can investigate mechanisms of competition, persistence and survival using concepts from population biology. Here we develop a suite of models that capture niche dynamics between cancer and healthy leukaemic stem cells. Given the current absence of suitable data -- and we will briefly outline the scope for generating such data -- our analysis proceeds by exploring the high-dimensional parameter space in a comprehensive way, which allows us to map out regions in parameter space that lead to distinctly different biological outcomes. In particular, we infer therapeutically desirable regimes of these models using approximate Bayesian computation. Our results show that in order to survive an invasion of the niche from leukaemia, maintaining a healthy pool of haematopoietic stem cells is crucial and much more important --- in the sense of "necessary" and "almost sufficient" --- than interventions targeted at suppressing leukaemic stem cells and their progeny. We conclude with a brief outline of how these insights can be applied in practice.

Bayesian experimental design to probe cellular decision making processes

Juliane Liepe, Sarah Filippi, Michal Komorowski, Michael Stumpf

The construction of parameterised mathematical models allows us to predict the behaviour of biological systems, if the system parameters are known. However, most likely we have only limited knowledge about the parameters, and need to infer them from available experimental data. In a Bayesian framework these parameter estimates are described by posterior distributions. But depending on the nature of the experiments these distributions can differ considerably, as different experimental setups may result in discrepant behaviour. The problem then is to define a set of experiments, which jointly allow us to capture maximal information about the system's parameters. This, we show, can be naturally incorporated into a generalization of the conventional Bayesian inference framework.

Here we develop and implement an algorithm that estimates the information content of a given experimental setup. Our method is based on estimating the mutual information between parameter space and system output. We demonstrate its performance and usefulness using the classical repressilator model and Hes1 gene regulation data. Finally, we use our method to design the optimal experimental setup to predict the time course of AKT phosphorylation in response to several stimuli.

Ridge regression for risk prediction with applications to genetic data

Erika Cule, Maria de Iorio

In recent years, technological developments have increased the availability of genetic data. We address the challenge of using this data to predict disease risk. Standard regression techniques traditionally used to fit

prediction models cannot be applied to contemporary genetic data sets, due to the high dimensionality of the data and the correlation structure among genetic variants. Penalised regression methods are a family of regression techniques that can be used with such data. Among penalised regression methods, ridge regression has been demonstrated to offer the best predictive performance. Ridge regression requires the specification of a penalty parameter, which controls the amount of shrinkage of the regression fit. Several methods have been proposed in the literature to choose the ridge parameter from the data. However, previously proposed methods fail when the data comprise more predictors than observations, as is typically the case in current genetic data sets.

Here, we propose a semi-automatic method to guide the choice of ridge parameter from the data when the data comprise many more predictors than observations. We propose choosing the penalty parameter such that the degrees of freedom for variance is the same as that of a principal components regression with a specified number of principal components. We discuss ways to choose the number of components to use, with the aim of good predictive performance. Using simulation studies, we demonstrate that when the number of causal variables is large and effect sizes are small, a plausible situation in the case of complex diseases, our method offers improved predictive performance over other penalized regression methods. We apply our method to out-of-sample prediction using two Bipolar Disorder genome-wide association studies, with data from the Wellcome Trust Case-Control Consortium and the Genetic Association Information Network.

Transfer Learning for Cancer Theranostics

Christian Widmer, Alexander Zien, Gunnar Raetsch

Genome-wide gene expression profiling by microarrays has exerted a massive impact on the clinical management of several diseases, in particular on breast cancer [1]. Based on the clustering of many gene expression profiles, several molecularly and clinically distinct subtypes of breast cancer have been identified that call for different treatment. Corresponding diagnostic tests have been developed that rely on sets of indicative genes, so-called signatures.

However, so far predictive signatures, ie tests that predict the success of potential treatment options, based on microarray measurements are not considered to be sufficiently reliable for clinical application [2]. One reason for this failure seems to be the large variability of measurement results that is introduced by circumstancial factors, like sample treatment and measurement protocol details. Several efforts have been undertaken in order to understand [3] and counteract [4] these effects. In this work, we explore the potential of multitask learning [5] for this purpose.

In this work, we aim to predict the response to preoperative chemotherapy of breast cancer, categorized as a pathological complete response (pCR = no residual invasive cancer in the breast or lymph nodes) or residual invasive cancer (RD) based on microarray gene expression data. We mined EMBL-EBI for suitable data sets and compiled a list of 4 studies.

In the setting we have in mind, each study is treated as a different task in a multitask setting. This effectively means learning a classifier for each study, while sharing information between studies by coupling these classifiers using multitask learning.

Out methods [6] are largely based on regularization-based Multitask Learning [7], of which we use an instance that extends the well established Support Vector Machine. Furthermore, we consider Multitask Feature Selection [8] and compare its results to established gene sets, such as MammaPrint.

In preliminary experiments, comparisons to baseline methods clearly show that it is in fact beneficial to combine information from different studies.

References

[1] Shiang, C., & Pusztai, L. (2010). Molecular profiling contributes more than routine histology and immonohistochemistry to breast cancer diagnostics. Breast cancer research : BCR, 12 Suppl 4, S6. doi:10.1186/bcr2735

[1a] Colombo, P.-E., Milanezi, F., Weigelt, B., & Reis-Filho, J. S. (2011). Microarrays in the 2010s: the contribution of microarray-based gene expression profiling to breast cancer classification, prognostication and prediction. Breast cancer research : BCR, 13(3), 212. doi:10.1186/bcr2890

[2] Bonnefoi, H., Underhill, C., Iggo, R., & Cameron, D. (2009). Predictive signatures for chemotherapy sensitivity in breast cancer: are they ready for use in the clinic? European journal of cancer (Oxford, England : 1990), 45(10), 1733\u20131743. doi:10.1016/j.ejca.2009.04.036

[3] MAQC Consortium, Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., et al. (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nature biotechnology, 24(9), 1151\u20131161. doi:10.1038/nbt1239

[4] Yasrebi, H., Sperisen, P., Praz, V., & Bucher, P. (2009). Can survival prediction be improved by merging gene expression data sets? PloS one, 4(10), e7431. doi:10.1371/journal.pone.0007431

[5] Caruana R. Multitask Learning. Machine Learning. 1997;28(1):41

[6] Widmer C, Rätsch G. Transfer Learning in Computational Biology, ICML2011. Proceedings on the Workshop on Unsupervised and Transfer Learning. 2011.

[7] Evgeniou T, Micchelli CA, Pontil M. Learning multiple tasks with kernel methods. Journal of Machine Learning Research. 2005;6(1):615-637.

[8] Argyriou A, Evgeniou T, Pontil M. Convex multi-task feature learning. Machine Learning. 2008;73(3):243-272

Learning Gene Networks Underlying Somatic Mutations in Cancer

Navodit Misra, Ewa Szczureck, Martin Vingron

Cancer progression is characterized by an accumulation of somatic mutations across the genome. Recent advances in sequencing technology have allowed us to explore the spectrum of somatic mutations in unprecedented detail. However, the processes behind cancer progression remain poorly understood. Standard models of sequence evolution assume different loci evolve independently and are ill-suited for modeling the complex set of gene interactions that are likely present in cancer cells. On the other hand, most existing models for cancer progression either constrain the set of possible pathways or model parameters, in order to overcome the noise due to small sample sizes. In this paper, we introduce a model for learning gene networks underlying somatic mutations by utilizing evolutionary information. Gene interactions are encoded by a Bayesian network and inferring the evolutionary history is interpreted as a structure learning problem in the presence of missing data. We establish a theoretical result to prune the set of feasible solutions to the inference problem. We combine the pruning criterion with a local search strategy to develop an efficient heuristic for structure learning and demonstrate its accuracy on simulated data sets. Finally, we apply our algorithm to learn the network

structure and infer the ancestral mutation patterns for glioblastoma and lung adenocarcinoma. Our results suggest the presence of subtle combinatorial effects may influence probable pathways during the early stages of cancer progression.

Application of Random Survival Forests on Gene Set Scores for Gastric Cancer Prognosis and Biological Interpretation

Tomas Martin-Bertelsen, Lennart Friis-Hansen, Ole Winther

Predicting survival for stomach cancer patients from genome-wide gene expression measurements on primary tumours have already shown promising preliminary results. Previous methods have used gene set scores as representations for biochemical pathway activations in individual patients but did only use a limited set of pathways of well-known relevance to gastric cancer.

We hypothesise that 1) using prior biological knowledge in the form of gene sets as predictors for gastric cancer survival enables interpretation of data in the same robust way as gene set enrichment analysis does over single-gene differential expression analysis, and 2) unbiased variable selection in a large collection of diverse gene sets could lead to improved predictive power and new insight into clinically relevant gastric cancer biology and treatment options.

We have applied an existing recent nonparametric method for survival regression for high-dimensional data, random survival forests [Ishwaran et al., Random survival forests for high-dimensional data. Statistical Analysis and Data Mining, 2011], to select strong variables predictive of gastric cancer survival. Using gene set t-statistics [Luo et al., GAGE... BMC Bioinformatics 2009] as predictor variables for individual patients, we evaluated the robustness of random survival forests for variable selection and survival prediction in comparison with using individual genes as predictors. Random survival forests provides survival curves for individual patients which, together with the biological interpretations of predictive gene sets, may enable personalised treatment of gastric cancer patients.

Our collaborators at Copenhagen University Hospital are engaged in experimental validation of our preliminary results.

Assessing technology-specific error signatures in next-generation sequencing, with an application to the 1000 Genomes Project data

Michael Nothnagel, Alexander Herrmann, Andreas Wolf, Stefan Schreiber, Matthias Platzer, Reiner Siebert, Michael Krawczak, Jochen Hampe

Next-generation sequencing (NGS) is a key technology in understanding the causes and consequences of human genetic variability. In this context, the validity of NGS-inferred single-nucleotide variants (SNVs) is of paramount importance. We therefore developed a statistical framework to assess the fidelity of three common NGS platforms and to estimate the proportion of false-positives heterozygotes based on read distributions. Application of this framework to aligned DNA sequence data from two completely sequenced HapMap samples as included in the 1000 Genomes Project revealed remarkably different error profiles for the three platforms. Newly identified SNVs showed consistently higher proportions of false positives (3-17%) when compared to confirmed HapMap variants. We show that this increase was not due to differences in flanking sequence features, read coverage or quality, nor was this observation limited to a particular data set or variant calling

algorithm. Consensus calling by more than one platform yielded significantly lower error rates (1-4%). This implies that the use of multiple NGS platforms may be more cost-efficient than relying upon a single technology alone, particularly in physically localized sequencing experiments that rely upon small error rates. Our study thus highlights that different NGS platforms suit different practical applications differently well.

Fiona: A tool for automatic correction of sequencing errors in genome sequencing experiments

Hugues Richard, David Weese, Manuel Holtgrewe, Marcel Schulz

Next generation sequencing technologies can produce a large amount of artifacts. These artifacts are in general limited to one or a few positions, like base substitution errors or short insertions and deletions (indels). In this context, automatic read error correction is an important step, as it allows to improve the performance of the downstream analysis tasks, such as genome assembly or SNP calling. All of the approaches do not rely on a previous alignment of the sequences to the genome, but rather use the relatively low proportion of errors (<5%) to detect the substrings likely to be erroneous. However most of the proposed methods until now suffer from at least one of the following drawbacks: (1) multiple parameters to set up for the user, (2) large memory consumption, or (3) no detection of indels.

We propose a new standalone read error correction method, Fiona, based on suffix arrays and which supports substitution and indel corrections for any next generation sequencing platform. The only parameters to set by the user is the average error rate, which is generally reported with the sequencing experiment. Fiona is provided with an efficient implementation in the Seqan library with very small memory consumption that can be run on inexpensive hardware, but also supports multi-core parallelization if available. When assessing the accuracy of Fiona over an extensive set of conditions -varying read length, error rate and coverage- it always performed better (produced more correct reads) than all other suffix tree based methods.

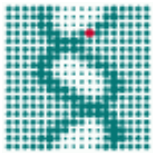
A probabilistic method for RNA-Seq read error correction

Marcel Schulz, Hai-Son Le, Ziv Bar-Joseph

Sequencing of RNAs with next generation sequencing technologies (RNA-Seq) has revolutionized the field of transcriptomics for genetics and medical research. RNA-Seq experiments are routinely applied to study mRNAs, miRNAs, and other short RNAs in a diverse range of organisms.

Error correction of RNA-Seq data is an important research direction to improve data analysis. In this talk we are going to devise the first general method to remove sequencing errors from RNA-Seq reads. Removal of sequencing errors in RNA-seq data is challenging because of the overlapping effects of non-uniform abundance, polymorphisms, and alternative splicing (mRNAs). We present the SEECER algorithm based on a formulation of profile hidden Markov models that addresses all the above mentioned challenges. We show that SEECER reduces the amount of sequencing errors, significantly increases the performance of downstream analyses with or without available reference sequence, and vastly outperforms ad-hoc approaches that researchers currently use.

We believe that SEECER will be useful for the RNA-seq community and help to get most out of the data.



MASAMB 2012 – Talks' abstracts

The allele distribution in next-generation sequencing data sets is accurately described as the result of a stochastic branching process

Thorsten Dickhaus, Verena Heinrich, Jens Stange, Peter Imkeller, Ulrike Krueger, Sebastian Bauer, Stefan Mundlos, Peter N. Robinson, Jochen Hecht, Peter M. Krawitz

Nowadays, whole-genome sequencing of the human deoxyribonucleic acid (DNA) has become possible with next-generation sequencing (NGS) technologies. However, the availability of such fine-grained genetic information poses new substantial challenges for statistical analysis of genetic data because of the enormous informational size of the human genome. Consequently, ultra high-dimensional techniques for simultaneous statistical inference have become a major branch of research in mathematical and applied statistics over the last two decades.

Here, we model the crucial (amplification) steps in an NGS protocol as a stochastic branching process and derive a mathematical framework for the distribution of alleles at heterozygous loci before sequencing. Specifically, we show that two-type Bienaymé-Galton-Watson branching processes with discrete time steps are suitable to model the amplification mechanism mathematically. Moreover, we apply a central limit theorem of the form derived in [1] for the fraction Q(k) of reference alleles at heterozygous genetic loci after k cycles of polymerase chain reaction (PCR) for the case that the initial number of amplifiable fragments tends to infinity. To this end, an analysis of the probability generating function of the process is crucial in order to calculate the limiting moments of Q(k).

Furthermore, we confirm our theoretical results by analyzing technical replicates of human exome data and demonstrate that the asymptotic variance of Q(k) is higher than expected by assuming a binomial distribution which does not capture the dynamics of the PCR process, but which has nevertheless often been utilized as a stochastic model in this situation. Hence, our findings contribute to reducing the number of false negatives in identification of heterozygous loci. The presentation is based on [2].

References

[1] Yakovlev, A. Y. and Yanev, N. M. (2009). Relative frequencies in multitype branching processes. Ann. Appl. Probab., 19, 1-14.

[2] Heinrich, V., Stange, J., Dickhaus, T., Imkeller, P., Krüger, U., Bauer, S., Mundlos, S., Robinson, P. N., Hecht, J. and Krawitz, P. M. (2012). The allele distribution in next-generation sequencing data sets is accurately described as the result of a stochastic branching process. Nucleic Acids Research, to appear (with technical supplement).

Normalization of DNase-seq data for classification of cell types

Michael Love

Different cell types can produce different programs of transcriptional regulation in part through variations in the chromatin accessibility of regulatory elements. This variation can be seen as a tuner, modulating the binding potential of transcription factors leading to more or less transcriptional activity of target genes. The ENCODE and Roadmap Consortia have made available genome-wide measurements of chromatin accessibility across hundreds of cell types, creating a unique opportunity to study regulatory elements. Chromatin accessibility is measured by DNaseI digestion followed by high-throughput sequencing (DNase-seq). The resulting pattern of reads along the genome informs as to the location, genomic width and degree of accessibility. The distributions of DNase-seq read counts can vary widely across experiments and laboratories, therefore the proper statistical

treatment of the data is critical to obtain a clear picture of cell-type differences. Here we show preliminary results in applying different methods of normalization and comparison of DNase-seq samples, including quantile normalization, mixture modeling, and discriminant analysis. We find that the variance arising from different protocols often dominates the variance from different cell-types. Within datasets using a common protocol, we are able to successfully identify cell-types and isolate regions which are most useful in identifying cell-type. The localization of these regions makes possible further study of regulatory elements, including searching for motifs of transcription factors and linking enhancers with potential target genes.

Detection of statistical significant differences in ChIP-seq data

Gabriele Schweikert, Guido Sanguinetti

ChIP-Seq has rapidly become the dominant experimental technique in functional genomic and epigenomic studies. Statistical analysis of ChIP-Seq data however remains challenging, due to the highly structured nature of the data and the paucity of replicates. Current approaches, largely borrowed from RNA-seq data analysis, focus on total counts of fragments mapped to a peak, mainly ignoring any information encoded in the shape of the peaks.

We demonstrate empirically on real data that higher order features of ChIP-Seq peaks carry important and often complementary information to total counts, and hence are potentially important in assessing differential binding. We then propose to incorporate higher order information in testing for differential binding by adapting recently proposed kernel-based statistical tests to ChIP-Seq data. Our results demonstrate that this approach can lead to very different and complementary predictions compared with standard tests based on total count.

A Bayesian bidirectional Hidden Markov model for ChIP sequencing data

Yanchun Bao, Veronica Vinciotti

An important biological question is the one of detecting the regions in the genome bound by histones 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. In the absence of binding, reads would be expected to be scattered across the genome, following some non-specific binding pattern which generates a background signal across the genome. Statistically, the observed count data, summarised into bins across the genome, are assumed to come either from a background or a signal distribution. Given these data, the interest is in inferring the state of the latent binary variable, which can be either "Enriched" or "Not Enriched". It is realistic in this application to assume a Markov property, whereby the probability of a region being enriched depends on the two neighbouring ones. In this talk, we present a bidirectional Hidden Markov model that can capture all these assumptions. The parameters of the model are estimated in a Bayesian framework, using either a Poisson or a Negative Binomial distribution for the counts.

An extensive simulation study shows that the model is competitive with existing methods, which are either based on a simple mixture model (without latent states) or on a uni-directional Hidden Markov model. Finally, the approach is used to detect the binding sites of two Transcription Factors (TFs), p300 and CBP, for which six ChIP-seq datasets have been recently generated, as well as four further datasets from a previous study. The experiments are run by two different labs, and different antibodies are used for each of the two transcription factors and from the two different labs, respectively. We discuss how, when comparing the binding profiles of the two TFs from ChIP-seq data, it is essential for any statistical model to account also for the different

antibodies' efficiencies. Our model does so, as antibody efficiency is reflected in the mixture portion of signal and background. In return, the model will also provide an estimate of relative antibody efficiency for different experiments, which aids the biological interpretation of the enriched regions detected by the model.

Sequencial Monte Carlo - Particle Gibbs inference of Transcriptional Landscape from RNA-Seq Data

Mirauta Bogdan, Pierre Nicolas, Hugues Richard

Sequencing technologies applied to transcriptome interrogation (RNA-Seq) permit a high precision in the inference of transcript localization and relative expression level. Namely, RNA-Seq produces millions of reads that, once aligned to the reference genome, provide counts that reflect the transcriptional landscape. This landscape is, even in the presence of post transcription processes, directly correlated to the expression activity. Read counts can serve to estimate, with an existing annotation, gene expression up to the isoform level [1] assuming homogeneous distribution of the reads inside predefined genome segments. Transcript boundaries can also be identified from abrupt variations in the local abundance of reads using sliding windows. However, realistic probabilistic models that simultaneously account for transcript boundaries and expression levels are still not available to describe RNA-Seq data. This precludes the use of a parametric inference framework to obtain estimate expression at the basepair level.

In this work, we design a strand specific model that includes the changes in expression level along the genome and the randomness induced by read sampling. Through Hidden Markov Model formalism this problem reduces to estimating the hidden path for the unobserved variable ut - the expression level at basepair t.

Ut is defined as the product between the relative abundance of the position t and the total number of reads. We then use a Sequential Monte Carlo (SMC) approach to infer hidden expression level (ut)1:T along a genome sequence of length T from observed read counts (yt)1:T . Model parameters are estimated in Bayesian framework. The SMC approach allows us to specify a reasonable model that fully exploits the complexity of RNA-Seq data.

Modeling RNA Pol-II Dynamics

Ciira Maina, Jaakko Peltonen, Antti Honkela, Magnus Rattray, Neil Lawrence

Gene transcription by RNA polymerase II (pol-II) is a key step in gene expression. Transcription consists of a number of dynamic events such as recruitment of pol-II to the promoter, elongation and termination. Furthermore, each of these steps may be rate limiting and therefore affect the level of gene expression. This motivates the study of transcription dynamics and in particular the effect of these dynamic events on gene expression.

High throughput technologies such as ChIP-seq allow us to determine protein-DNA interaction and have been used to investigate DNA-pol-II interaction. This provides a direct measure of gene activity since presence of pol-II on the gene body is correlated to gene expression (Welboren \textit{et al.} \textit{The EMBO Journal} 28, 1418-1428, 2009). Recently, global run-on and sequencing (GRO-seq) has been used to directly measure the presence of transcriptionally engaged polymerase on the gene body (Hah \textit{et al.} \textit{Cell} 145, 622-634, 2011). Examination of these data shows `transcription waves' of polymerase moving down the gene in response to stimuli.

In this work we present a mathematical model that directly models the movement of pol-II down the gene body. This model allows us to compute the transcription speed for each gene and also determine the genes responding differentially to stimuli using high thoughput sequencing data.

Model Description

In order to capture the movement of the transcription wave down the gene, we divide the gene into I segments and compute time series of pol-II occupancy for each of the segments. Due to the low temporal resolution characteristic of high-throughput data sets, the time series between measurements must be inferred. To this end, we model the pol-II occupancy for each segment $i\in\{0,\ldots,I\}$ as the convolution of a latent process $f(t)$ which is shared by all segments and a (possibly delayed) smoothing kernel $k_i(\tau-D_i)$ corrupted by an independent white Gaussian noise process $\epsilon_i(t)$.

The latent process $f(t)$ is modeled as a random function drawn from a Gaussian process (GP). GPs have been successfully applied to inference problems involving gene expression time series with limited data (Honkela \textit{et al.} \textit{PNAS} 107, 7793-7798, 2010). The smoothing kernel is assumed to be Gaussian. The estimated delay D_i of each smoothing kernel models the amount of time it takes the `transcription wave' to reach the corresponding gene segment. This is used to estimate the transcription speed.

Results

We apply our methodology to a data set obtained using GRO-seq to measure pol-II occupancy genome-wide when MCF-7 cells are treated with estradiol (E2)(Hah \textit{et al.} \textit{Cell} 145, 622-634, 2011). Our method is able to determine transcription speeds which as expected vary on a gene by gene basis. However, the values of transcription speeds we obtain are consistent with the literature. An advantage of our method is that it allows the investigation of transcription dynamics genome-wide as opposed to gene by gene as is prevalent in the current literature.

Joint models of gene expression and global phenotypes in the context of genetic and environmental variations

Julien Gagneur, Oliver Stegle

Phenotypic traits such as disease susceptibility are controlled by the joint influence of genotype and environment. Dissecting the molecular mechanisms that underlie these dependencies promises to deliver the necessary insights to develop drugs tailored to the genetic background and life circumstances of the patient. To this end, we have combined cellular and molecular phenotyping with statistical modelling to explain the molecular state as a function of genotype and environment and to provide for specific predictions of molecular intervention points.

Growth rates as model of global health traits have been determined for 160 yeast strains across more than 20 environments, revealing multiple loci with a condition-specific effect on fitness. In addition, a total of 190 expression profiles, sampling these strains across five representative environments have been performed.

First, we set out to predict the transcriptional state of the cell from genotype and environment. We employed a multi-task Gaussian process model with a structured covariance, sharing predictive strength across environments and groups of transcripts. The trained method provides for interpretable genetic models of each transcript, including genetic factors that act in cis or trans, regulatory hotspots across transcripts and the environment-specific contributions of each of these factors to expression variability. We found that sharing

information across environments and transcripts greatly improved the robustness to estimate gene-environment interactions, yielding improved predictive power.

Second, we developed novel inference techniques to pinpoint molecular intervention points with an environment-specific role on growth. Our approach builds on Bayesian model comparisons, assessing the statistical evidence that a particular transcript carries a mediating role between genetic signal and its environment-specific effect on growth. We applied the approach to genome-wide identified transcripts specific for each environment-specific growth QTLs, some of which are literature-validated. We have further promising candidates for which a wet lab validation is underway.

Together, the computational tools we have devised and the data we collected demonstrate how genomics and physiological data can be integrated to aid the design of personalized therapies.

Finding topics in diseases through the analysis of RNA-seq data

Nicolo Fusi, Neil Lawrence

Obtaining meaningful diagnostic or prognostic biomarkers from gene expression data is a critical factor in achieving a better understanding of disease dynamics. For this reason, several biomarker discovery approaches have been proposed over the years, proving quite successful in classifying disease subtypes or predicting survival in patients. In order to gain further insights, practitioners often have to conduct follow up analyses on the list of selected biomarkers, for instance by clustering them into functional categories or pathways.

Some methods [3] try to bridge this gap by clustering the genes as a preprocessing step, reducing the dimensionality of the data before classification. This has been shown to improve the results, but it discards the label information during the clustering phase. The method we propose is able to cluster functionally related genes while modelling the class labels at the same time. It builds on relatively recent approaches emerging from the natural language processing community, and in particular from the latent Dirichlet allocation model (LDA) [2]. In LDA, documents (such as web pages) are modelled as a mixture of topics, and each topic has several words associated to it. Similarly, we model biological samples as a mixture of pathways (topics) that are active at any given time, with each pathway containing one or more active genes (words). We also assume that one or more of these pathways have an active impact on the disease, and thus are associated to the class labels. This is closely related to the model proposed in [1], but with a different distribution for the labels. The generative process for the model we are proposing, for each sample N, is:

1. draw the mixing coefficients \theta according to Dir(\alpha)
2. for each gene/probe w_D
    (a) draw a pathway z_D according to Multinomial(\theta)
    (b) draw an expression level w_D Multinomial(\beta_{z_D})
3. draw a class label y | z_{1…Q} according to softmax(zb,\eta), where zb = 1/D sum_{i=1..D} z_i

A graphical representation of the model is provided in Figure 1. We show on simulated and real data from human and mouse that the proposed approach achieves a striking classification accuracy, while being able to extract meaningful disease pathways at the same time.

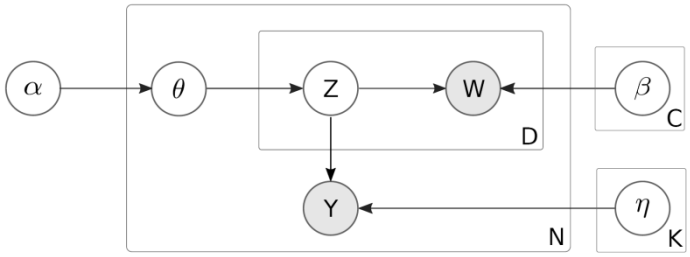


Figure 1: Graphical representation of the model. N is the number of samples, D is the number of probes, K is the number of pathways, C is the number of classes, W is the RNA-seq data, Y are the class labels, Z is the latent variable for the class allocations, \theta is a Dirichlet distributed vector of mixing coefficients.

References

[1] D.M. Blei and J.D. McAuliffe. Supervised topic models. Arxiv preprint arXiv:1003.0783, 2010.

[2] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. The Journal of Machine Learning Research, 3:993{1022, 2003.

[3] M. Yousef, M. Ketany, L. Manevitz, L. Showe, and M. Showe. Classification and biomarker identification using gene network modules and support vector machines. BMC bioinformatics, 10(1):337, 2009.

Evidential model ranking without likelihoods

Vladislav Vyshemirsky

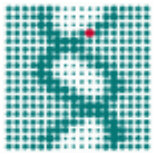
We present a probabilistic formulation of the Approximate Bayesian Computation scheme that allows evidential ranking of alternative models without direct use of a likelihood function. This approach is particularly important when ranking of several sophisticated stochastic models is desired, and the likelihood is either too complex or impossible to define. We suggest a modification of a Sequential Monte-Carlo sampler that uses ideas of Path Sampling to estimate an approximation to marginal likelihoods as a measure of evidence support. We demonstrate applications of this method on a problem of ranking alternative models of cancerous tumour growth using unique data from three cancerous spheroid lines.

The Effect of Subsample Size in Stability Selection

Andre Beinrucker, Gilles Blanchard, Ueruen Dogan

Selecting relevant variables is an important task in many field, where high dimensional datasets are involved. Unfortunately many variable selection methods are unstable, which makes the conclusions drawn from its selections unreliable. 'Stability Selection' is a meta method introduced by Meinshausen and Bühlmann that allows to stabilise any variable selection method. It basically consists of applying the original variable selection procedure to subsamples of the data and averaging the results and was recently found useful in many fields of applications including Genetics.

In our work we show how the performance of Stability Selection depends on the sizes of the chosen subsamples. We present experimental results on simulated and real datasets which underline our theoretical findings.



MASAMB 2012 – Talks' abstracts

Using Log-Concave Functions to Describe Peaks in Spectrometry Data

Sven Rahmann

Spectrometric techniques are pervasive in modern life sciences, e.g., mass spectrometry in proteomics and metabolomics, ion mobility spectrometry (IMS) in metabolomics of volatile organic compounds (VOCs), nuclear magnetic resonance (NMR) spectroscopy in structural (bio)chemistry, etc.

For the purposes of this work, a spectrum is a signal s(t) over time t, measured at discrete time points. We interpret the signal as a sample from a mixture of several (overlapping) peak components and a flat background component modeling noise.

A frequently arising problem, especially for IMS and NMR, is the decomposition of the measured signal into its (overlapping) peaks, which is routinely solved with methods such as Expectation Maximization (EM). There are two challenges, however. First and foremost, most of the existing methods use parametric peak models (e.g., Gaussians, Cauchy distributions, skewed Gamma distributions). In some cases, such as for IMS spectra, there is yet no physical theory describing the true peak shape, so each parametric model might be questionable. Second, the EM algorithm depends on reasonable starting parameters to converge to a meaningful local optimum.

We address both challenges simultaneously by proposing a nonparametric peak decomposition method. We only assume that each peak is log-concave, which appears to be a reasonable assumption as all parametric peak models currently used in practice satisfy this property. Log-concave functions have many desirable statistical properties and have been used recently in a variety of applications [review by Walther, Inference and modeling with log-concave distributions, Statistical Science 24(3):319--327, 2010].

Our approach is to fit a log-concave function of largest area under the signal, anchored at the location of the signal maximum, in order to describe the most prominent peak. We iterate this process with the remainder of the signal until we decide that only noise remains. The resulting decomposition can be used to extract peak descriptors (location, area, etc.) directly, or used as a starting point for EM methods for likelihood optimization, both parametric and nonparametric.

We discuss results on several IMS spectra from the project ``Resource-Constrained Analysis of Spectrometry Data'', where we collaborate with B&S-Analytik, a Dortmund-based manufacturer of IMS devices. This work has been supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 ``Providing Information by Resource-Constrained Analysis'', project B1.