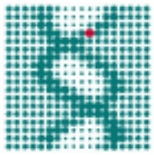

MASAMB 2012 – Posters' abstracts

Approximate Maximum Likelihood Inference for Coalescent Models

Johanna Bertl, Andreas Futschik, Greg Ewing

Statistical inference is often based on the likelihood. Under coalescent models, the likelihood cannot be obtained analytically, except for very simple situations. Therefore, simulation-based inference methods like ABC and MCMC are often used. Employing different strategies, they explore the parameter space to detect regions of high posterior probability.

We investigate a stochastic gradient algorithm, which is related to the Kiefer-Wolfowitz algorithm. It moves along the direction of a simulated gradient of the likelihood towards the maximum likelihood estimator. An advantage of this approach is that it keeps the number of simulations from low likelihood regions in the parameter space low.

We present simulation results that illustrate the performance of the algorithm in terms of speed and the number of dimensions it can cope with. Since it turns out that a good choice of tuning parameters is essential for the performance of the algorithm, we propose some good tuning strategies.

Modelling Temporally Varying Metabolic Fluxes

Justina Zurauskiene, William Bryant, John Pinney, Michael Stumpf

In order to investigate how metabolism functions in biochemical networks, flux balance analysis (FBA) has become one of the generic tools. By adding constrains to the stoichiometric analysis of a metabolic system FBA is able to determine the metabolic fluxes at steady state. The main problems, however, are that FBA does not uniquely specify fluxes (but instead has to invoke additional criteria and constraints, such as optimality of biomass production) and that it cannot be used for modelling the dynamical behaviour of fluxes. Here we use a non-parametric Bayesian approach that allows us to extend conventional FBA to temporally varying flux data. In contrast to traditional approaches to metabolic flux estimation, which employ uniform sampling of the space, the novel technique proposed here allows us to capture the temporal evolution of flux dynamics at successive time-points. The correlation between fluxes is here captured using a multiple-output Gaussian process, which allows us to describe how the metabolic fluxes change over physiological time-scales. I will illustrate the usefulness of this approach by application to metabolic models of Mycobacterium tuberculosis.

Improved MSMAD approach for highly noisy aCGH data of hepatocellular carcinomas

Michael G. Schimek, Tomas Reigl, Eva Budinska

Genome analysis has become one of the most important tools for understanding the complex process of cancerogenesis. With increasing resolution of CGH arrays, computationally efficient algorithms are in demand, which are effective in the detection of aberrations even in highly noisy data.

Budinska, Gelnarova and Schimek (2009; Bioinformatics 25/6, 703-713) have developed a non-parametric statistical technique called MSMAD (Median Smoothing Median Absolute Deviation) for aCGH analysis. They combine quantile smoothing for pre-processing with a median absolute deviation concept for double breakpoint detection (one based on the unsmoothed and the other on the smoothed data) followed by a merging step to reduce the number of false discoveries. In a systematic comparison, MSMAD outperformed a popular

segmentation method, a Hidden Markov Model-method, and a fused lasso-based method for real as well as simulated data in most instances.

In cancer research, the aCGH data sets are very large and usually characterized by a poor signal-to-noise ratio. Therefore analysis methods of high computational efficiency are required that enable us to cope with very noisy measurements. Here we present an improved MSMAD algorithm in R and apply it to selected hepatocellular carcinoma data.

Hybrid Monte Carlo for ODE models in Systems Biology

Andrei Kramer, Nicole Radde

Ordinary Differential Equations (ODEs) are powerful tools for the modeling of intracellular processes in systems biology. The free parameters of such ODE models (usually reaction rate coefficients) need to be estimated to fit observations. It is quite common that this forms an ill posed inverse problem, i.e. problems where straight-forward optimization techniques are not the right method. Bayesian methods deal with uncertainties by associating a probability distribution to the parameters; observations then exclude regions in parameter space by assigning low posterior probabilities to them. This approach requires effective Markov Chain Monte Carlo (MCMC) algorithms -- to draw representative samples from the posterior distribution. One such method is the Hybrid Monte Carlo algorithm. It is adaptive to a given problem (its posterior) and aims to provide samples with low auto-correlation. We introduce an adaptation of this algorithm to typical problems in systems biology. We show that equilibrium state measurements (e.g. western blotting) or low resolution time series measurements offer (as compensation) possibilities for performance improvements during HMC sampling.

References

[1] A. Kramer, and N. Radde. Optimal experimental design using a Bayesian framework for parameter identification in dynamic intra-cellular network models. Int. Conf. on Comput. Sci. (ICCS), Amsterdam, Netherlands, 2010.

[2] Mark Girolami, Ben Calderhead, and Siu A. Chin. Riemannian manifold hamiltonian monte carlo. July 2009.

Residual Component Analysis for Estimating Sparse Inverse plus Low Rank Structures

Alfredo Kalaitzis, Neil Lawrence

Probabilistic principal component analysis (PPCA) seeks a low dimensional representation of a data set in the presence of independent spherical Gaussian noise, $\mathbf{\Sigma} = \sigma^2\mathbf{I}$. The maximum likelihood solution for the model is an eigenvalue problem on the sample covariance matrix. In this paper we consider the situation where the data variance is already partially explained by other factors, e.g. conditional dependencies between the covariates, or temporal correlations leaving some residual variance. We decompose the residual variance into its components through a generalised eigenvalue problem, which we call residual component analysis (RCA, \citet{Kalaitzis:rca11}).

This new data analysis technique is combined with GLASSO in an EM framework to estimate the sparse inverse and low rank components of a covariance matrix model. Full covariance matrix models of data are often problematic as their parameterization scales with D^2.

Two separate approaches to a reduced parameterization of these matrices are to base them on low rank matrices (as in probabilistic PCA) or on a sparse inverse structure (as in GLASSO).

These two approaches have very different characteristics: one involves specifying sparse conditional independencies in the data, the other assumes that a reduced set of latent variables is governing the data. Clearly, in any given data set, both of these characteristics may be present. Our sparse plus low rank approach is the first approach to deal with both these cases in the same model.

We illustrate the ideas on the recovery of a protein-signaling network, and gene expression time-series.

Inferring Nucleosome Positioning from Sequencing Data

Alessandro Mammana

Determining where histones are located in the genome is one of the most fundamental problems in Epigenetics. While in the past several efforts have been done for the computation of the binding affinities from sequence data, it is now clear that the sequence cannot provide reliable prediction. For this purpose ChIP-seq is a technique that offers unprecedented and largely unexplored opportunities. The purpose of this work is to develop an algorithm for the identification of nucleosome positions from sequencing data. In contrast to previous approaches, our method aims at an accurate estimation of the binding probability of well-positioned as well as less stable nucleosomes, taking into account information from a control sample, for bias estimation, and from multiple Chip-seq experiments, for studying the relationship between different modifications. We plan to show that such a probabilistic description is suitable both for the detection of well-positioned nucleosomes and for genome-wide comparative analysis.

Portraying the expression landscapes of cancer subtypes

Lydia Hopp, Henry Wirth, Mario Fasold, Markus Loeffler, Hans Binder

Cancer is a complex disease that involves a sequence of gene-environment interactions in a progressive process that occurs with dysfunction in multiple systems. There are many biological pathways and multiple genetic, epigenetic and transcriptional influences working simultaneously in the expression of cancer phenotypes. The study of individual components in isolation does not allow an adequate understanding of phenotypic expression. Instead, an integrative approach is needed to investigate gene-environment interactions. We need to adapt a holistic view on the activation pattern rather than to consider single genes or single pathways.

With this motivation we apply self-organizing maps (SOM) which is a feature centred machine-learning clustering method to large scale patient expression data of different cancers (Glioblastoma Multiforme, Burkitt's Lymphoma, Prostate Cancer) in order to characterize the specifics of the genome wide expression landscapes in different molecular subtypes of each cancer entity. Our method simultaneously searches for features which are differentially expressed and correlated in their profiles in the set of samples studied. We aim to merge many of the functionally related genes into larger aggregates called functional modules defined as sets of genes related to one type of a cellular or functional process (e.g. inflammation, cell division, etc.) and to characterize disease-specific changes in the resulting interaction network. Characteristic differences between sample types and developmental stages can be clearly identified and further analyzed using so-called 'metagene-profiles' characterizing the intrinsic correlation modes. SOMs portray molecular phenotypes with individual resolution. We demonstrate the potency of the method in selected applications characterizing the diversity of gene

expression the cancer subtypes studied and to discover similarity relations between them. In addition to functional modules the individual portraits allow identifying misclassified samples and thus to improve quality control in large patient series.

Matapax: An online high-throughput genome-wide association study pipeline

Liam Childs, Jan Lisec, Dirk Walther

High throughput sequencing and genotyping methods are dramatically increasing the number of observable genetic intraspecies differences that can be exploited as genetic markers. In addition, automated phenotyping platforms and OMICS profiling technologies further enlarge the set of quantifiable macroscopic and molecular traits at an ever increasing pace. Combined, both lines of technological advances create unparalleled opportunities to identify candidate gene regions and, ideally, even single genes responsible for observed variations in a particular trait via association studies. However, as of yet this new potential is not sufficiently matched by enabling software solutions to easily exploit this wealth of genotype-phenotype information. We have developed Matapax, a web-based platform to address this need. Initially, we built the infrastructure to support association studies in Arabidopsis thaliana based on several genotyping efforts covering up to 1,375 Arabidopsis accessions. Based on the user-supplied trait-information, associated SNP-markers and SNP-harbouring or neighbouring genes are identified using both the GAPIT and EMMA libraries developed for R. Additional interrogation is facilitated by displaying candidate regions and genes in a genome browser and by providing relevant annotation information. In the future, we plan to broaden the scope of organisms to other plant species as more genotype/phenotype information becomes available.

Availability: Matapax is freely available at http://matapax.mpimp-golm.mpg.de and can be accessed using any internet browser.

Second order linear regression analysis of mouse microarray data provides increased specificity and reveals interactions between covariates like strain and cytokine stimulation

Olivia Prazeres Da Costa, Thorsten Buch, Achim Tresch

Background

The investigation of molecular mechanisms of transcriptional regulation requires the detailed analysis of genome-wide expression data. Current approaches typically examine gene expression data that has been obtained under the influence of multiple factors, such as genetic background, environmental conditions, time, or external perturbations. They often lack an appropriate analysis strategy that takes into account the effects of multiple factors and their potential interactions. Here, we apply second order multiple linear regression to microarray data of mouse bone marrow-derived macrophages of (van Erp, Dach et al. 2006). In this experiment, the following factors were actively varied: the genetic background (BALB/c or C57BL/6), the bacterial strain employed for infections in vitro (control strain WA(pTTS, pP60) or virulent strain WA(pYV)), and the presence or absence of IFN-γ. We calculate the main effects and the interaction effects of these three factors.

Results

We present a method to calculate the individual and combinatorial effects of multiple factors that have an effect on transcription. For each gene, we can qualify the interaction between two factors into neutral (no interaction),

alleviating (effects of co-occurrence of these factors results in weaker expression than expected from the single effects), or aggravating (combined effects stronger than expected from single effects). In a subsequent gene set enrichment analysis for Gene Ontology classes, we can show that genes pertaining to a certain interaction type are enriched in functional classes that are closely related with the respective factors.

Conclusions

We demonstrate our approach provides new insights in the transcriptional response of the host system, and may become a standard tool for the analysis of expression data obtained under the complex influence of multiple factors.

Experiments in transcript isoform expression and differential expression estimation from RNA-seq data

Peter Glaus, Antti Honkela, Magnus Rattray

We present comparisons of our recently developed Bayesian method for RNA-seq transcript isoform expression and differential expression estimation, BitSeq[1], to a number of alternatives in both problems.

In transcript expression analysis, we present comparisons against Cufflinks, RSEM and MMSEQ. All the methods perform approximately equally well in real data comparison agains TaqMan qRT-PCR. In within-gene isoform level analysis using synthetic data, fully Bayesian methods clearly outperform Cufflinks with BitSeq better than the rest especially for isoforms with few reads.

In differential expression analysis, we compare against Cuffdiff, baySeq, DESeq and edgeR using synthetic data, although the last three have not been designed for transcript-level analysis. For the last three of these methods, we use counts derived from BitSeq expression level mean estimates. Overall, BitSeq provides slightly higher accuracy than the other methods, followed most closely by baySeq. This order and differences are especially clear for more weakly expressed transcripts. For highly expressed transcripts, DESeq and edgeR are equally good or slightly better than the others.

References

[1] P. Glaus, A. Honkela, and M. Rattray. Identifying differentially expressed transcripts from RNA-seq data with biological variation. arXiv:1109.0863 [q-bio.GN]

Computational analysis of the transcriptional program of pluripotency

Kalaimathy Singaravelu, Andreas Beyer

Embryonic stem (ES) cells, which are derived from the inner cell mass of the developing blastocyst, possess the dual capacities of self-renewal and pluripotency. Specific patterns of histone modifications and DNA methylation are major characteristics of the pluripotent state of ES and induced pluripotent (iPS) cells. However, the molecular mechanisms driving these modifications are not fully understood. The recent successes in the reprogramming of somatic cells to an embryonic stem cell fate by expression of key transcription factors also clearly demonstrated that the epigenome of a differentiated cell can be rewired to support embryonic development. The regulatory mechanisms by which the molecular machinery (i.e histone methyl transferases) decide where to place the modifications remains an open question. In order to address these questions we are combining published and new DNA-binding information of relevant factors and genome-wide histone

modification data to search for patterns of binding factors that explain specific epigenetic marks. In order to obtain more detailed insights into the regulation of specific promoter classes, we are distinguishing binding events at CpG island and TATA boxes and we analyse the co-occurence of bindings and epigenetic marks separately for different promoter types. The co-occurrence of histone 3 lysine 4 tri-methylation (H3K4me3) with histone 3 lysine 27 tri-methylation (H3k27me3) has been particularly debated. We are therefore specifically discussing the correlation of these marks with protein binding events.

Comparative analysis of germline-cancer sample pairs

Paul Pyl, Anna Paruzynski, Anne Arens, Christof von Kalle, Manfred Schmidt, Wolfgang Huber

We analyse pairs of samples (control and leukemia) from 3 patients with T-ALL. We performed DNA and RNA Sequencing and will investigate common patterns in structural variants and SNVs in those patients.