10TH INTERNATIONAL WORKSHOP ON MACHINE LEARNING IN SYSTEMS BIOLOGY

SUMMARY

Biology is rapidly turning into an information science, thanks to enormous advances in the ability to observe the molecular properties of cells, organs and individuals. This wealth of data allows us to model molecular systems at an unprecedented level of detail and to start to understand the underlying biological mechanisms. This field of systems biology creates a huge need for methods from machine learning, which find statistical dependencies and patterns in these large-scale datasets and that use them to establish models of complex molecular systems. MLSB is a scientific forum for the exchange between researchers from Systems Biology and Machine Learning, to promote the exchange of ideas, interactions and collaborations between these communities.

WORKSHOP DETAILS

Date: Saturday September 3 – Sunday September 4, 2016
Time: 09:00 – 17:00
Venue: World Forum, room: Kilimanjaro 1/2

ORGANISERS

Juho Rousu, Associate Professor of Information and Computer Science at Aalto University, Finland
Dick de Ridder, Professor of Bioinformatics at Wageningen University, The Netherlands
Harri Lähdesmäki, Assistant Professor in the Department of Computer Science at Aalto University, Finland
Aalt-Jan van Dijk, Assistant Professor of Plant Systems Biology at Wageningen University, The Netherlands

Contact organisers through: mlsb2016@easychair.org

SPONSORED BY:

PHILIPS
ENZA ZADEN
RIJK ZWAAN
HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY
BAYER
WAGENINGEN UR For quality of life
BASECLEAR

Oral presentations

1 How network topological models influence drug-target prediction

Simone Daminelli, Josephine Thomas, V. Joachim Haupt, Claudi Duran, Michael Schroeder and Carlo Vittorio Cannistraci

The identification of drug-target interactions (DTIs) is important for understanding drug mode of action, infer new indications and identify possible side effects. Nevertheless, it is still a challenging task especially if we consider its formal definition as link-prediction problem in complex networks. Moreover, since novel drug-target validation is a costly and time consuming endeavour, a reliable evaluation of predictors performance is an open problem.

In this work we compare state-of-the-art supervised methods and topology-based models for drug-target interaction prediction. Besides, we consider prediction algorithms both based on bipartite network projections as well as recently proposed topological models, based on the Local Community Paradigm (LCP).

We analyze 5 gold standard DTIs networks and provide an exhaustive performance evaluation based on two validation frameworks. Additionally, we include a novel independent benchmark set of both positive and negative drug-target interactions defined by the value of their experimental chemical affinity. Finally, we investigate differences and similarities of the novel predictions derived from methods inspired by different principles.

Our results show that drug-target networks have enough topological information to identify highly reliable predictions, with comparable performance to state-of-the-art supervised methods which exploit additional knowledge. Surprisingly, a detailed analysis of novel predictions revealed that each method prioritize distinct true interactions, suggesting that a combination of a variety of methodologies motivated by diverse principles might improve the current drug-target discovery strategies.

4 On the inconsistency of l1-penalised sparse precision matrix estimation

Otte Heinävaara, Janne Leppä-Aho, Jukka Corander and Antti Honkela

Various l1-penalised estimation methods such as graphical lasso and CLIME are widely used for sparse precision matrix estimation. Many of these methods have been shown to be consistent under various quantitative assumptions about the underlying true covariance matrix. Intuitively, these conditions are related to situations where the penalty term will dominate the optimisation.

In this paper, we explore the consistency of l1-based methods for a class of bipartite graphs arising from sparse latent variable models, which are strongly motivated by several types of applications. We show that all l1-based methods fail dramatically for models with nearly linear dependencies between the variables. We also study the consistency on models derived from real gene expression data and note that the assumptions needed for consistency never hold even for modest sized gene networks and l1-based methods also become unreliable in practice for larger networks.

5 Statistical modeling of isoform splicing dynamics from RNA-seq time series data

Yuanhua Huang and Guido Sanguinetti

Isoform quantification is an important goal of RNA-seq experiments, yet it remains problematic for genes with low expression or several isoforms. These difficulties may in principle be ameliorated by exploiting correlated experimental designs, such as time series or dosage response experiments. Time series RNA-seq experiments, in particular, are becoming increasingly popular(1,2,3), yet there are no methods that explicitly leverage the experimental design to improve isoform quantification.

Here, we report on a new methodology, DICEseq (Dynamic Isoform spliCing Estimator via sequencing data) to jointly estimate the dynamics of isoform proportions from RNA-seq experiments with structured experimental designs. DICEseq is a Bayesian method based on a mixture model whose mixing proportions represent isoform ratios, as in (4); however, DICEseq incorporates the correlations induced by the structured design by coupling the isoform proportions in different samples through a latent Gaussian process (GP). By doing so, DICEseq effectively transfers information between samples, borrowing strength which can aid to identify the isoform proportions. Full mathematical and algorithmic details are given in the full paper (5); here we focus on some of the result highlights, and briefly discuss how the work is currently being expanded.

Numerical experiments on simulated data sets show that DICEseq yields more accurate results than state-of-the-art methods, including MISO(4), Cufflinks(6), and IsoEM(7). By simulating RNA-seq reads for 8-time-point experiments with different coverages, Figure 1A shows that DICEseq is able to exploit effectively the temporal information, providing a significantly lower mean absolute error than the other methods, an advantage which is particularly marked at lower coverage (RPK<=200). In addition, DICEseq clearly is able to provide more confident predictions at all coverage levels (Figure 1B), particularly at lower coverage; this is important, as often the confidence of an estimate is used to select genes which are further analysed (2).

To further assess accuracy of our method with real data sets, we compare the correlation of isoform quantification between two replicates for 309 yeast intron-containing genes at 1.5, 2.5 and 5.0 minute. Figure 1C shows DICEseq improves the Pearson's correlation coefficient to 0.896, outperforming by between 4 and 6 percentage points existing methods. The improvement is particularly marked if looking at the lower third genes of the expression range: DICEseq still obtains a Pearson correlation of 0.860, while the other methods achieve much lower correlations, ranging from 0.657 (Cufflinks) to 0.775 (IsoEM). This is remarkable since, as there are only three time points, the improvement obtained by taking temporal information into account could be expected to be limited. In addition, we computed the number of transcripts that pass a frequently used threshold (95%CI<0.3) for further analysis. Figure 1D illustrates that at all time points around 20% more genes are retained using a joint analysis, compared to methods that analyse data points in isolation.

Furthermore, DICEseq permits to quantify the trade-off between temporal sampling of RNA and depth of sequencing, frequently an important choice when planning experiments (data shown in the full paper(5)). Our results have strong implications for the design of RNA-seq experiments, and offer a novel tool for improved analysis of such data sets. Python code is freely available at http://diceseq.sf.net.

Current work is aiming at extending and improving the isoform quantification problem by including additional side information. A particularly promising avenue is to consider chromatin state (e.g. as assayed by ChIP-seq towards particular histone modifications) as an additional covariate which can be incorporated in the prior over isoform proportions. Initial results show that isoform expression levels can be predicted from histone modification data using a simple linear predictor, yielding a high test-set correlation of 0.745 (Figure 2). The incorporation of histone modification data in an integrative model can therefore both improve isoform quantification, and highlight non-trivial associations between chromatin state and alternative splicing.

6 Computational reconstruction of NFkB pathway interaction mechanisms during prostate cancer

Daniela Börnigen, Svitlana Tyekucheva, Xiaodong Wang, Jennifer Rider, Gwo- Shu Lee, Lorelei Mucci, Christopher Sweeney and Curtis Huttenhower

One major goal in cancer research is still the identification of novel drug targets. Although prostate cancer has been linked to NFκB and associated pathways, neither the full repertoire of molecular players nor their mechanisms of interaction have been fully specified. Instead, a more detailed mechanistic perspective of this pathway may lead to targeted transcript assays or to sets of informative gene products, which might identify patients with high risk of micrometastatic disease at time of surgery or radiotherapy, who would then be in need of systemic adjuvant therapy to prevent relapse and death.
Still, the identification of such pathways remains a challenging problem—one which we addressed here by computationally reconstructing interaction mechanisms of the NFκB pathway in prostate cancer. So far, the construction of novel pathways from a set of genes or the inclusion of novel genes within existing pathways is based on literature curation, predictive computational models, or lab experiments, and only few studies have predicted novel pathways for specific diseases or tissues. Here, we address this challenge and associate NFκB with a novel predicted pathway activated during prostate cancer, from which we already experimentally confirmed predicted associations between a few pathway members. Our suggested computational method is generalizable to other tissue types, cancers, and organisms, while the predicted novel information about the NFkB pathway will allow us to further understand prostate cancer and to develop more effective prevention and treatment strategies.

7 Integrating gene set analysis and nonlinear predictive modeling of disease phenotypes using a Bayesian multitask formulation

Mehmet Gönen

Motivation: Identifying molecular signatures of disease phenotypes is studied using two mainstream approaches: (i)~Predictive modeling methods such as linear classification and regression algorithms are used to find signatures predictive of phenotypes from genomic data, which may not be robust due to limited sample size or highly correlated nature of genomic data. (ii)~Gene set analysis methods are used to find gene sets on which phenotypes are linearly dependent by bringing prior biological knowledge into the analysis, which may not capture more complex nonlinear dependencies. Thus, formulating an integrated model of gene set analysis and nonlinear predictive modeling is of great practical importance.

Results: In this study, we propose a Bayesian binary classification framework to integrate gene set analysis and nonlinear predictive modeling. We then generalize this formulation to multitask learning setting to model multiple related datasets conjointly. Our main novelty is the probabilistic nonlinear formulation that enables us to robustly capture nonlinear dependencies between genomic data and phenotype even with small sample sizes. We demonstrate the performance of our algorithms using repeated random subsampling validation experiments on two cancer and two tuberculosis datasets by predicting important disease phenotypes from genome-wide gene expression data. We are able to obtain comparable or even better predictive performance than a baseline Bayesian nonlinear algorithm and to identify sparse sets of relevant genes and gene sets on all datasets. We also show that our multitask learning formulation enables us to further improve the generalization performance and to better understand biological processes behind disease phenotypes.

Availability: Matlab and R implementations of our two methods are available at \href{https://goo.gl/1lRUpp}{https://goo.gl/1lRUpp}.

12 Spectral Consensus Strategy for Accurate Reconstruction of Large Biological Networks

Severine Affeldt, Nataliya Sokolovska, Edi Prifti and Jean-Daniel Zucker

Background: The last decades witnessed an explosion of large-scale biological datasets whose analyses require the continuous development of innovative algorithms. A striking example lies in the recent gut bacterial studies that provided researchers with a plethora of information sources. However, despite this deeper knowledge of microbiome composition, inferring bacterial interactions remains a critical step that encounters significant issues. First, available metagenomics datasets are under high-dimensional settings, with thousands to millions of features per observation. Second, for most of the gut bacterial taxa, the complete genome reference is still unknown making it difficult to properly identify species that participate in microbiome networks. Lastly, the multi-step procedure required for bacterial taxa identification, ranging from DNA extraction to genome catalogue mapping, introduces a significant amount of noise in already sparse datasets. Such high-dimensional data challenge state of the art reconstruction methods and make any a priori choice of a learning approach particularly difficult.

Results: We propose a consensus method, named Spectral Consensus Strategy, to reconstruct large biological networks based on spectral decomposition. The results obtained on benchmark datasets demonstrate the interest of our approach, in particular for high-dimensional cases. When applied to the human gut microbiome co-presence network, our method successfully retrieves biologically relevant interactions and gives new insights into the topology of this complex ecosystem.

Conclusions: The Spectral Consensus Strategy improves prediction accuracy and allows scalability of any reconstruction method to large networks. The integration of multiple reconstruction algorithms turns our approach into a robust learning method. Altogether, this strategy increases the confidence of predicted interactions from high-dimensional datasets without demanding computations.

13 DegreeCox: a network-based regularization method for survival analysis

André Veríssimo, Arlindo L. Oliveira, Marie-France Sagot and Susana Vinga

Background: Modeling survival oncological data has become a major challenge as the increase in the amount of molecular information nowadays available means that the number of features greatly exceeds the number of observations. One possible solution to cope with this dimensionality problem is the use of additional constraints in the cost function optimization. Lasso and other sparsity methods have thus already been successfully applied with such idea. Although this leads to more interpretable models, these methods still do not fully profit from the relations between the features, specially when these can be represented through graphs. We propose DegreeCox, a method that applies network-based regularizers to infer Cox proportional hazard models, when the features are genes and the outcome is patient survival. In particular, we propose to use network centrality measures to constrain the model in terms of significant genes.

Results: We applied DegreeCox to three datasets of ovarian cancer carcinoma and tested several centrality measures such as weighted degree, betweenness and closeness centrality. The a priori network information was retrieved from Gene Co-Expression Networks and Gene Functional Maps. When compared with Ridge and Lasso, DegreeCox shows an improvement in the classification of high and low risk patients in a par with Net-Cox. The use of network information is especially relevant with datasets that are not easily separated. In terms of RMSE and C-index, DegreeCox gives results that are similar to those of the best performing methods, in a few cases slightly better.

Conclusions: Network-based regularization seems a promising framework to deal with the dimensionality problem. The centrality metrics proposed can be easily expanded to accommodate other topological properties of different biological networks.

14 Predicting Internal Ribosome Entry Site (IRES) activity from sequence

Alexey Gritsenko, Shira Weingarten-Gabbay, Shani Elias-Kirma, Ronit Nir, Dick de Ridder and Eran Segal

Since its discovery in viruses, translation of mRNAs through initiation at Internal Ribosome Entry Sites (IRESs) has emerged as a prominent mechanism of cellular and viral initiation. It supports cap-independent translation of select cellular genes under normal conditions, and remains functional in conditions when cap-dependent translation is inhibited. IRES structure and sequence are believed to be involved, however, systematic investigation of sequence and structure determinants of IRES activity lagged behind. With the recent discovery of thousands of novel IRESs in human and viruses we provided a 50-fold increase over previously available data (Weingarten-Gabbay et al., Science, 2016). The next big challenge is to use these data to devise a quantitative model to predict IRES activity from RNA sequences. Here, we performed the first in-depth computational analysis of this large body of IRESs, in which we explored RNA sequence features predictive of their activity and provide novel insights on the effect of sequence features, their number and position on IRES activity.

18 DGW: an exploratory data analysis tool for clustering and visualisation of epigenomic

Saulius Lukauskas, Roberto Visintainer, Gabriele Schweikert and Guido Sanguinetti

Functional genomic and epigenomic research relies fundamentally on sequencing based methods like ChIP-seq for the detection of DNA-protein interactions. These techniques return large, high dimensional data sets with visually complex structures, such as multimodal peaks extended over large genomic regions. Current tools for visualisation and data exploration represent and leverage these complex features only to a limited extend.

We present DGW, an open source software package for simultaneous clustering and alignment of multiple epigenomic marks. DGW uses Dynamic Time Warping to capture the structure of epigenomic marks, by adaptively rescaling genomic distances to group regions of interest with similar shapes. We demonstrate the effectiveness of the approach in a simulation study and on a real epigenomic data set from the ENCODE project. Our results show that DGW automatically recognises and aligns important genomic features such as transcription start sites and splicing sites from histone marks.

DGW is available as an open source Python package at https://lukauskas.github.com/dgw/.

21 ChARM: Discovery of combinatorial chromatin modification patterns in hepatitis B virus X-transformed mouse liver cancer using association rule mining

Sung Hee Park, Sung-Min Lee, Young-Joon Kim and Sangsoo Kim

Background:

Various chromatin modifications, identified in large-scale epigenomic analyses, are associated with distinct phenotypes of different cells and disease phases. To improve our understanding of these variations, many computational methods have been developed to discover novel sites and cell-specific chromatin modifications. Despite existing methods, these remain some rooms for improvement when they applied to resolve histone code hypothesis. Hence, our approach provides new insights into de novo combinatorial pattern discovery of chromatin modifications to characterize epigenetic variations in distinct phenotypes of different cells.

Results:

We report a new computational approach, ChARM (Combinatorial Chromatin Modification Patterns using Association Rule Mining), that can be employed for the discovery of de novo combinatorial patterns of differential chromatin modifications. We used ChARM to analyse chromatin modification data from the livers of normal (non-cancerous) mice and hepatitis B virus X (HBx)-transgenic mice with hepatocellular carcinoma (HCC), and discovered 2,409 association rules representing combinatorial chromatin modification patterns. Among these, the combination of three histone modifications, a loss of H3K4Me3 and gains of H3K27Me3 and H3K36Me3, was the most striking pattern associated with the cancer. This pattern was enriched in functional elements of the mouse genome such as promoters, coding exons and 5’UTR with high CpG content, and CpG islands. It also showed strong correlations with polymerase activity at promoters and DNA methylation levels at gene bodies. We found that 30% of the genes associated with the pattern were differentially expressed in the HBx compared to the normal, and 78.9% of these genes were down-regulated. The significant canonical pathways (Wnt/ß-catenin, cAMP, Ras, and Notch signalling) that were enriched in the pattern could account for the pathogenesis of HBx.

Conclusion:

ChARM, an unsupervised method for discovering combinatorial chromatin modification patterns, can identify histone modifications that occur globally. ChARM provides a scalable framework that can easily be applied to find various levels of combination patterns, which should reflect a range of globally common to locally rare chromatin modifications.

Availability:

Source code will be available upon request from the authors.

Posters

2 The translation of lipid profiles to lipid biomarkers in the study of infancy nutrition

Animesh Acharjee, Philippa Prentice, Carlo Acerini, James Smith, Ken Ong, Julian L. Griffin, David Dunger and Albert Koulman

Links between early life exposures and later health outcomes may in part be due to nutritional programming in infancy. This hypothesis is supported by observed long-term benefits associated with breast-feeding, such as better cognitive development in childhood, and lower risks of obesity and high blood pressure in later life. Effects of early nutritional interventions in infancy, using nutrient-enriched milk formulas, include increased later risk of metabolic disease [2]. However, the underlying mechanisms are unknown and are difficult to study. We recently applied shotgun lipidomics to the study of infant metabolism. Distinct differences in 3-month lipidomic profiles were observed between exclusively breast-fed and formula-fed infants; mixed-fed infants showed intermediate profiles and suggested that it would be possible to predict the diet of an infant based on their lipid profile. Our aim was to validate if there were significant differences in the lipid profile of infants due to their early nutrition.

3 Identifying synthetic microbial communities by learning in-silico communities using flow cytometry

Peter Rubbens, Willem Waegeman, Ruben Props and Nico Boon

Single-cells can be characterized in terms of their phenotypic properties using flow cytometry. However, up to our knowledge there has not yet been a thorough survey which tries to predict the label of microbial single-cells based on flow cytometry data. This paper aims to not only assess the quality of flow cytometry data when measuring microbial species, but also to suggest a method for creating and monitoring future synthetic microbial communities. We will do this by creating so-called in-silico communities, communities which allow us to explore properties of microbial flow cytometry data using supervised learning techniques; moreover we will show that it is possible to extrapolate these properties when identifying their in-vitro counterpart.

8 Automated Functional Annotation in UniProtKB with a Novel Approach: UniProt-DAAC (Domain Architecture Alignment and Classification)

Tunca Dogan, Alistair MacDougall, Rabie Saidi, Diego Poggioli, Alex Bateman, Claire O'Donovan and Maria Martin

Similarity based methods have been widely used in order to infer the properties of genes and gene products containing little or no experimental annotation. We propose a novel approach for the automatic annotation of protein sequences in the UniProt Knowledgebase (UniProtKB) by comparing their domain architectures, classifying proteins based on the similarities and propagating functional annotation. The performance of this method was measured through a cross-validation analysis using the Gene Ontology (GO) and Enzyme Commission (EC) annotation of UniProtKB/Swiss-Prot proteins. The results indicate the effectiveness of this approach in detecting functions of proteins with an average F-score: 0.85. We applied the method on nearly 55.3 million uncharacterized proteins in UniProtKB/TrEMBL resulting in 44,818,178 GO term and 3,374,889 EC number predictions for

12,172,114 and 3,056,760 proteins, respectively. About a quarter (24%) of these predictions were for 4,462,290 previously non-annotated protein entries, indicating the significance of the value added by this approach.

10 Combining network proximity and drug similarity for side effect detection

Emre Guney

One of the biggest challenges in the pharma industry nowadays is the failure of candidate drugs due to unexpected side effects. Computational methods offer a cost effective alternative to experimental methods to characterize drug side effects, but they typically rely on training predictors based on drug and side effect similarity. Here, we present, ProXide, a novel network proximity based drug side effect prediction method that does not require neither drug nor side effect similarity data. We systematically analyze 819 FDA approved drugs and 537 side effect modules in the interactome and show that proximity can detect known drug side effects with prediction accuracy comparable to similarity-based approaches. We also demonstrate that ProXim, integrating ProXide with drug chemical and target similarity into a simple logistic regression classifier, performs better than using any single method individually and proves to be more resilient to data incompleteness. Lastly, we highlight two use cases in which ProXide and ProXim can give insights on drug side effects observed in the clinic.

11 Topic modelling of biomedical text: from words and topics to disease and gene links

Sarah Elshal, Jaak Simm, Mithila Mathad, Jesse Davis and Yves Moreau

Background

The massive growth of biomedical text makes it very challenging for researchers to review all relevant work and generate all possible hypotheses in a reasonable amount of time. Many text mining methods have been developed to simplify this process and quickly present the researcher with a learned set of hypotheses. Previously, we have focused on the task of identifying genes that may be related to a disease. We applied a word-based concept profile similarity to learn patterns between disease and gene entities and hence identify links between them. In this work, we study an alternative approach based on topic modelling to learn similar patterns between the disease and the gene entities and measure its effect on the identified links.

Results

We evaluated two setups: (1) learning the topics from an input set of abstracts, and (2) learning the topics from an input set of genes, each represented by a set of abstracts. In both setups, we used the learned topics to create topic profiles for the disease and gene entities, and then measured the similarity between profiles. We calculated the recall in the top 10, 25, 50, and 100 ranked genes. Using the first setup, we achieve a recall of 33% and 41.7% in the top 10 and the top 100 ranked genes. Using the second setup, we achieve a recall of 44% and 74% in the same rankings, which improves the performance of 37% and 66% recall obtained using our original methods.

Conclusions

Topic modelling is very powerful in finding links between genes and diseases inside biomedical text. Compared to other text mining techniques, it can easily improve the performance as long as we find the correct settings.

15 Bayesian integrative clustering of heterogeneous types of high-throughput sequencing data

Chantriolnt-Andreas Kapourani and Guido Sanguinetti

Modern high-throughput sequencing technologies generate large amounts of biological data from different sources, and these data are used to measure diverse, but often related and complementary information. Our aim is to perform integrative clustering on heterogeneous types of HTS data using a Bayesian integrative data modelling approach. The standard approaches either perform separate clustering followed by post hoc integration [Wang et al., 2011] or incorporate all data sources simultaneously and generate a single joint clustering [Kormaksson et al., 2012]. However, flexible clustering methods need to be developed that can simultaneously model information from different platforms and can also capture the underlying structural similarities across the data sources.

16 Tree based feature induction for biomedical data

Konstantinos Pliakos and Celine Vens

During the recent years, a great advance in both biomedical data acquisition technologies and feature extraction methods has been witnessed. Harnessing these new tools and technologies has led to an indisputable increase in the number of available biomedical datasets. Despite the efforts made so far, the representational power of features used to describe a sample in such datasets, such as a gene in gene function prediction datasets or a protein in protein interaction datasets has yet to be improved. Here, the performed study focuses on the feature representation power from a machine learning perspective.

17 Revealing discriminative network functional modules in omic sciences: an easy and fast unsupervised multivariate method

Sara Ciucci, Yan Ge, Alessandra Palladini, Víctor Jiménez Jiménez, Luisa María Martínez Sánchez, Susanne Sales, Andrej Shevchenko, Steve W. Poser, Maik Herbig, Oliver Otto, Andreas Androutsellis-Theotokis, Jochen Guck, Mathias J. Gerl and Carlo Vittorio Cannistraci

Background

Recent advances in high-throughput techniques made available a large number of omic datasets and consequently required the development of network-inference methods, to describe the biomedical systems under analysis. Precisely, reverse-engineering or inferring networks are the process of identifying associations between omic entities behind the complexity of a biosystem. However, the usually employed correlation-based network methods only highlight linear associations between omic features, but do not pinpoint the main actors that are responsible for the perturbation of the system under analysis. On the other hand, building correlation networks between significant molecules pre-selected by means of a univariate statistical test is an ‘over-the-counter’ solution that neglects the multivariate and collective mechanisms at the basis of the omic system complexity. Ultimately, a plethora of methods are available for gene network reverse engineering but few methods are developed and extensively tested for inferring discriminative associations in omic systems in general.

Results

In this study, we developed a new unsupervised multivariate algorithm named PC-corr, to reveal linear discriminative correlation network modules, where combinations of features contribute to distinguish two or more conditions under analysis. PC-corr is based on a preliminary sample exploration by Principal Component Analysis (PCA) of the omic dataset; it extracts the significant PC loadings (eigenvectors) and combines them with the omic-feature linear correlations to reveal multivariate discriminative network associations. The method is unsupervised, hence it is particularly suited for the analysis of small-size datasets - which frequently occur in biomedicine - or pilot studies in which knowledge uncertainty represents an important issue.

Conclusions

Our results demonstrate that PC-corr method can be used as a valid and fast tool for multivariate inference of discriminative associations between the variables of a general omic dataset and, as a consequence, for identifying the most relevant omic network modules. We therefore propose a paradigm shift that results particularly useful for small datasets - where supervised approaches are unfeasible - from the univariate selection of single discriminative features and their correlation, to the multivariate identification of discriminative and collective associations between omic features. PC-corr can thus represent a new tool in precision medicine for the definition of combinatorial and multiscale biomarkers in complex omic data.

22 Predictive modeling of binding affinities between chemical compounds and protein targets for drug discovery and repurposing applications

Anna Cichonska, Balaguru Ravikumar, Elina Parri, Tapio Pahikkala, Antti Airola, Krister Wennerberg, Juho Rousu and Tero Aittokallio

We aim to develop a kernel-based machine learning modeling framework for drug-target interaction prediction, which integrates multiple biological and molecular data sources, along with learning their importance for the prediction task.

A basic assumption in the machine learning algorithms is that similar compounds bind to similar targets. Molecular similarities can be encoded using different types of kernels, which have proven powerful in capturing nonlinear molecular properties. The model uses information on two- and three-dimensional structures of both compounds and protein targets. Moreover, we integrate additional network-centric knowledge into the model, for example, in the form of protein-protein interactions, which can further extend the search space of new drug candidates or target proteins. We focus on a regression problem, where the task is to predict quantitative binding affinities, instead of the common classification setting, which treats the molecular interactions as simple on-off relationships.

We applied our framework to predicting missing bioactivities in the large-scale profiling study by Metz et al. that generated kinome-wide compound-target interaction map. To assess the model’s accuracy, we experimentally validated 100 predicted compound-kinase interactions, and obtained a high Pearson correlation of 0.8 between the predicted and validated binding affinities. This result demonstrates the potential of our modeling framework for inferring novel interactions of drug-like compounds. The approach can be used also with other target families, and may lead to novel lead compound discoveries as well as drug repurposing opportunities.

23 Seeing the Trees through the Forest: Sequence-based Homo- and Heteromeric Protein-protein Interaction sites prediction using Random Forest

Qingzhen Hou, Paul De Geest, Wim Vranken, Jaap Heringa and K. Anton Feenstra

Motivation: To fulfil biological functions, proteins bind to their partners via specific amino acids.Investigation of the properties and sequential information of these residues is important to reveal the mechanisms of protein-protein interactions and protein functions. These properties, derived from the interacting amino acids at sequence level, are usually exploited as features for machine learning methods to predict protein interacting positions. In this paper, we include two novel features (backbone flexibility and Sequence Specificity) predicted from sequences for protein interface prediction and evaluate the importance of different features using Random Forest. Results: We observe that there is no single sequence feature which enables to pinpoint interacting sites. However, combination of different properties does help the interface prediction. After selecting and integrating multiple features, we developed a Random Forest predictor which is able to distinguish interface and other residues with AUC of ROC plot at 0.72 in our homomeric test-set, which is better than other sequence-based methods. Moreover, when applied to identify interfaces of an independent heteromeric dataset, our method performs slightly better than the best sequence-only predictor. Thus, our predictor trained on homodimeric proteins can not only predict homodimeric interfaces, but is also able to locate interface residues in the heterodimers which suggested that our predictor captures the common properties of both homodimer and heterodimer interfaces.

24 Simultaneous prediction of protein-protein contacts and interaction partners

Miguel Correa Marrero, Richard G.H. Immink, Dick de Ridder and Aalt D.J. van Dijk

Protein-protein interactions underlie virtually any biological process. How proteins interact with each other is therefore a fundamental question in biology. However, techniques that give fine-grained information about protein-protein interactions are low-throughput and labour-intensive, which makes the development of in silico approaches attractive.

One way to approach the problem is to exploit the phenomenon of coevolution. Protein-protein interaction leads to the coevolution of the interfaces between the interaction partners, meaning that there are correlations between their sequences. From these correlations, one can deduce which residues are involved in the interaction interfaces. This can be done by applying statistical models to multiple sequence alignments of homologs of the proteins of interest.

However, one can easily introduce pairs of sequences that have lost the interaction, or paralogs. This introduces noise in the analysis and has limited the application of these coevolutionary approaches. To surpass this obstacle, we are developing a novel approach. Our approach combines traditional correlated mutation analysis with the expectation-maximization algorithm. For each sequence pair in the input alignments, the algorithm will first predict whether they are interacting or not. Using proteins predicted to interact, the algorithm will then predict contacts between columns in the alignment. These two steps are repeated until convergence is reached. This approach is still being tested.

25 Predicting Cleavage Sites Using a Generic Machine Learning Framework for Local Protein Properties

Nadav Brandes, Dan Ofer and Michal Linial

Determining residue-level protein properties, such as sites of post-translational modifications (PTMs), is vital to understanding protein function. Experimental methods are costly and time-consuming, while traditional rule-based computational methods fail to annotate sites lacking substantial similarity. Machine Learning (ML) methods are becoming fundamental in annotating unknown proteins and their heterogeneous properties. We present ASAP (Amino-acid Sequence Annotation Prediction), a universal ML framework for predicting residue-level properties. ASAP extracts numerous features from raw sequences, and supports easy integration of external features such as secondary structure, solvent accessibility, intrinsically disorder, or PSSM profiles. Features are then used to train ML classifiers. ASAP can create new classifiers within minutes for a variety of tasks, including PTM prediction. We present a detailed case study for ASAP: CleavePred, an ASAP-based model to predict protein precursor cleavage sites, with state-of-the-art results. Protein cleavage is a PTM shared by a wide variety of proteins sharing minimal sequence similarity. Current rule-based methods suffer from high false positive rates, making them suboptimal. The high performance of CleavePred makes it suitable for analyzing new proteomes at a genomic scale. The tool is attractive to protein design, mass spectrometry search engines and the discovery of new bioactive peptides from precursors. ASAP functions as a baseline approach for residue-level protein sequence prediction. CleavePred is freely accessible as a web-based application. Both ASAP and CleavePred are open-source with a flexible Python API. ASAP's and CleavePred source code, webtool and tutorials are available at: https://github.com/ddofer/asap; http://protonet.cs.huji.ac.il/cleavepred.

26 Lipoinformatics – machine learning approach to study lipid profiles

Neetika Nath, Christian Klose, Mathias Gerl, Michal A. Surma, Kai Simons and Lars Kaderali

Lipids are the highly diverse class of molecules that are structural components of biological membranes and function as energy reserves and signalling molecules. Within the metabolomics field, shotgun lipidomics, providing absolute quantification and high reproducibility is perfectly suited for bioinformatics approaches to guide the biotechnologies to improve human health.

The objective of this study is to develop a robust bioinformatics approach to identify lipid diagnostic biomarkers in human plasma that support the classification of subjects with high or low body mass index (BMI). The second objective of this study is to compare different normalization strategies for lipoidomic data of 326 human subjects with high (BMI > 30) or low (BMI < 25) BMI. We applied a random forest method implemented in varSelRF (R package) executing 1000 bootstrap samples. This yielded the most important features distinguishing high and low BMI. The resulting set of discriminating lipids is selected by the backwards stepwise elimination of features with smallest cross-validation error.

In our analysis we found no significant differences between normalizations by total lipid content or lipid class. The models were equally good with accuracies close to 0.75 and sensitivities and specificities at 0.72 and 0.75, respectively. Our results suggest that if using random forest for the analysis, the focus of the analysis to determine the important features.

28 The Systems Toxicology Computational Challenge: Markers of Exposure Response Identification – Insights gained

Vincenzo Belcastro, Carine Poussin, Stephanie Boue, Yang Xiang, Florian Martin, Julia Hoeng and Manuel C Peitsch

Humans are constantly exposed to chemicals (e.g. pollutants and pesticides) that may trigger harmful molecular changes. Risk assessment in the context of 21st century toxicology relies on the elucidation of mechanisms of toxicity and the identification of markers of exposure response from high-throughput data. The development of relevant computational approaches for the analysis and integration of these large-scale data remains challenging.

The purpose of sbv IMPROVER (www.sbvimprover.com/) is the crowd-sourced verification of methods in systems biology via computational challenges. The latest challenge (2016) aimed to address questions on the identification of exposure response markers in human blood enabling to discriminate between (A) exposed and non-exposed subjects, and (B) subsequently between formerly exposed and never exposed subjects (sub-challenge1) as well as the translatability of those markers between species (sub-challenge2). Participants were provided with human and mouse blood gene expression datasets to develop human-specific and species-independent gene signature-based models for class label prediction of independent test samples. Anonymized participant's predictions were scored according to a predefined methodology approved by an independent scoring review panel of experts.

Twenty-three teams worldwide participated in at least one sub-challenge. Most of the teams provided highly accurate predictions (pval < 0.05) for the first task (A), while prediction performances were much lower for task B. Different classes of machine learning methodologies were applied including Linear Discriminant Analysis, and Random Forest. A small set of features were common to top 3 ranked submissions. The challenge outcome and lessons learned will be shared with the computational scientific community.

29 ARBA: Association-Rule-Based Annotator

Rabie Saidi, Imane Boudellioua, Robert Hoehndorf, Victor Solovyev and Maria Martin

The widening gap between known proteins and their functions has encouraged the development of methods to automatically infer annotations. Automatic functional annotation of proteins is expected to meet the conflicting requirements of maximizing annotation coverage while minimizing erroneous functional assignments. This trade-off imposes a great challenge in designing intelligent automatic annotations systems. In the scope of this work, we suggest that association rule mining and selection techniques can be used effectively as computational methods for functional prediction. We introduce our automatic annotation system, ARBA (Association-Rule-Based Annotator) that can be used to enhance the quality of automatically generated annotations as well as annotating proteins with unknown functions. ARBA learns on data from UniProtKB/Swiss-Prot (1) and uses InterPro signatures and organism taxonomy as attributes to predict most of the protein functional annotations including Gene Ontology terms, metabolic pathways, EC numbers, etc. With respect to certain quality measures, we find all rules which would define significant relationships between attributes and functional annotations in UniProtKB/Swiss-Prot entries. The set of extracted rules represent the comprehensive knowledge which could explain protein functions. However, these rules comprise redundant information and their high number makes it infeasible to apply them on large sets of data such as UniProtKB/TrEMBL (1). To address this issue, ARBA puts these rules into a fast competition process called SkyRule (2) based on two concepts, namely dominance and comparability (2). Rules are then elegantly and considerably reduced in number and aggregated to form concise prediction models that assign functional annotations to UniProtKB entries.

To give a picture of the efficiency of ARBA in this paper, we briefly report its performance in the case of prediction of metabolic pathway involvement for prokaryotes.

30 Fast metabolite identification using Input Output Kernel Regression

Céline Brouard, Huibin Shen, Kai Dührkop, Florence D'Alché-buc, Sebastian Böcker and Juho Rousu

An important problematic of metabolomics is to identify metabolites using tandem mass spectrometry data. In this work we propose to address the metabolite identification problem using a structured output prediction approach, called Input Output Kernel Regression. We show that our method achieves state-of-the-art accuracy in metabolite identification rates and results in vast improvements in running times.

31 Simple enough biomarkers predict a complex disease phenotype

Iryna Nikolayeva, Kevin Bleakley, Anavaj Sakuntabhai and Benno Schwikowski

During dengue virus outbreaks, many hospitals are overcrowded with patients due to the possibility of complications that occur several days after hospital admission. Little is known about mechanisms triggering this severe reaction. On the molecular level, single molecular biomarkers are associated to dengue severity. But none of them has any predictive power. Predicting at admission which patients will develop complications from molecular data may allow to have a new insight on dengue etiology.

Based transcriptomic data from blood serum in 42 patients at hospital admission, we find simple, yet biologically powerful predictors of dengue complications from omics data.

Specifically, we use a generalization of linear models that describe the disease severity using an ensemble of monotonic functions of pairwise transcript measurements. Our implementation allows, for the first time to our knowledge, genome-wide screening and goes beyond classical linear and logistic models; it allows to model relations such as “AND” and “OR” between genes. Features are easier to interpret. And our ensemble model allows us to control the complexity of our predictor.

Our best-performing classifier is composed of 66 pairs of gene coding transcripts and has an accuracy of 83,3%. Transcripts in this biomarker are enriched in various innate and adaptive immune pathways. These pathways include processes known to play important roles in severe dengue: Cytokine-cytokine receptor interaction, TNF signalling pathways, regulation of cell-cell adhesion, regulation of epithelial cell differentiation.... Moreover some known single molecular feature biomarkers appear in our ensemble biomarker such as the endothelial permeability activation marker VEGFA or the metalloproteinase MMP9 related to vascular leakage.

We present the methodology, results from our genome-wide screen for biomarkers for dengue severity, and compare its predictive performance to the state-of-the art biomarker prediction methods.

32 Non-Stationary Gaussian Process Regression with Hamiltonian Monte Carlo

Markus Heinonen, Henrik Mannerstrom, Juho Roussu, Kaski Samuel and Harri Lähdesmäki

We present a novel approach for non-stationary Gaussian process regression (GPR), where the three key parameters -- noise variance, signal variance and lengthscale -- can be simultaneously input-dependent. We develop gradient-based inference methods to learn the unknown function and the non-stationary model parameters, without requiring any model approximations. For inferring the full posterior distribution we use Hamiltonian Monte Carlo (HMC), which conveniently extends the analytical gradient-based GPR learning by guiding the sampling with the gradients. The MAP solution can also be learned with gradient ascent. In experiments on several synthetic datasets and in modelling of temporal gene expression, the non-stationary GPR is shown to give major improvement when modeling realistic input-dependent dynamics.

33 Deciphering gene regulatory network from kinetic gene expression with an unfavorable data-to-variables ratio

Lise Pomiès, Mélanie Decourteix, Justin Bedo, Nathalie Leblanc-Fournier, Bruno Moulia and Florence d'Alché-buc

Inference of gene regulatory networks is central to analyze complex dynamical biological processes. For processes without prior information, global approaches such as DNA-microarray or RNAseq are generally used to acquire gene expression measurements. Those approaches allow to obtain a huge number of genes (or variables) but for few expressions conditions (or data). However inference of gene regulatory networks usually based on mathematical modeling (linear autoregressive models, dynamic Bayesian networks, state-space models) requires at least as many as datapoints as variables to be relevant and in general even more. To tackle this problem, we reduce the network inference problem to module inference problem and investigate, in a real problem, three strategic choices: (1) clustering, (2) data smoothing and (3) a modular approach with linear modeling. We will not insist on the first step, clustering, which has been widely used to reduce the number of variables, allowing to focus on a few representative genes instead of a whole genome. Once this subset of representative genes is defined from DNAchips data, a qPCR experiment allows to acquire additional datapoints for this reduced set. The second step, Data Smoothing, consists in the artificial addition of datapoints using a smoothing approach, here a Gaussian process modeling, on replicate data. The two first steps are supposed to provide a more favorable ratio between datapoints and number of variables. Eventually in the third step, to reduce the number of parameters, we assume that the remaining genes can be gathered in different modules, each of them enjoying the same regulatory program. To implement this idea, we define a mixture of linear dynamical models, each of one applying on a subset of genes. The originality of this third step relies on the estimation procedure, E.-M. like, that maximizes the log-likelihood under a constraint of Hilbert-Schmidt Independence criterion between modules.

In this contribution, we develop this methodology and present work in progress on a real biological problem, the accommodation process of poplar to repeated mechanical loads. Actually, in Europe, 5 % of the annual timber harvest is lost due to strong winds. Due to global climatic changes, increases of strong wind episode frequency is expected while chronic wind speed during the vegetation period - useful for the acclimation of plant to wind - may decrease. It is important to understand how trees may or may not acclimate to these new wind regimes. Wind induces stem flexions perceived by trees, that trigger a growth acclimation response during several days [2]. Up to date, only a few studies analyzed the underlying mechanisms of the tree responses to wind at the transcriptomic level, rather focusing on few genes or at a single time after the mechanical solicitation. In this study, we analyzed the kinetic responses of the whole transcriptome of Populus after one controlled bending of the stem, with DNA-microarray. We obtained around 3,000 candidates genes for the network with a kinetics composed of 4 measurement points, which means a really unfavorable data-to-variables ratio.

34 AGO-sRNA affinity to improve in silico sRNA detection and classification in plants

Lionel Morgado and Frank Johannes

Small RNAs (sRNA) have an important role in the regulation of gene expression, either through post-transcriptional silencing or the recruitment of repressive epigenetic marks such as DNA methylation. In plants, the mode of action of a given sRNA is tightly related with the Argonaute protein (AGO) to which it binds. High throughput sequencing in combination with immunoprecipitation techniques have made it possible to determine the sequences of sRNA that are bound to different families of AGO. Here we apply Support Vector Machines (SVM) to recent AGO-sRNA sequencing data of A. thaliana to learn which sRNA sequence features govern their differential association with certain AGOs. Our SVM classifiers show good sensitivity and specificity and provide a framework for accurate in silico sRNA detection and classification in plants.

35 Inferring dynamically evolving regulatory networks using mechanistic modeling approach

Jukka Intosalmi, Kari Nousiainen, Juho Timonen, Helena Ahlfors and Harri Lähdesmäki

Mechanistic ordinary differential equation based models are able to describe molecular interactions involved in gene regulation. Using statistical techniques, such models can be calibrated to experimental data. Furthermore, in some cases the model structure can be inferred from time-course measurements. However, mechanistic models assume network structure to be static and do not support well changes in the network structure. However, these kinds of changes in a network may be cause by unobserved changes in epigenome or signaling pathways and are important to consider.

In a recent study [1], we introduced the so-called latent effect mechanistic (LEM) model that can be used to model regulatory networks in the presence of rewiring effects. The LEM model is accompanied with a statistical framework as well as computational techniques that make it possible to calibrate the model structure and parameters in a data-driven manner. In this poster presentation, we outline the key idea behind the LEM modeling and present how the modeling approach can be successfully applied to study the regulatory interactions during T helper 17 (Th17) cell differentiation using time-course RNA sequencing data. Further, we discuss possible extensions to the methodology that is used to calibrate the model structure.

[1] Intosalmi J, Nousiainen K, Ahlfors H, Lähdesmäki H, Data-driven mechanistic analysis method to reveal dynamically evolving regulatory networks, Bioinformatics (ISMB2016), Vol. 32, No. 12, pp. i288-i296, 2016.

36 Predicting flowering time using a combined statistical-mathematical model

Aalt-Jan Van Dijk and Jaap Molenaar

Background: Plants integrate various signals in order to flower under optimal conditions. We recently published a dynamic model for the network involved in this signal integration. It describes how a set of eight transcription factors regulate each other and how perturbations in their expression change flowering time. To fit model parameters we exploited time course expression levels.

Method: In the present project we connect the earlier developed integration network model to a network model for the various upstream genes that are involved in receiving environmental and

endogenous signals. Data available for these genes consists of large amounts of gene expression data, typically of a static nature. Our approach is to use for the upstream network a statistical model, whereas the integration network is formulated in terms of ODEs. We demonstrate how to connect the Bayesian Network model with the dynamical ODE model.

Results: By connecting a large statistical model to a detailed smaller mathematical model we are able to tackle the complexity of the system underlying the regulation of flowering time. The Bayesian Network predicts the effect of perturbations of any of the upstream genes on genes included in the integration network model, which in turn predicts the resulting change in flowering time. We demonstrate the predictive power of this approach by comparison with mutant data.

Conclusion: We show that the intricate regulation of flowering time can be successfully modelled if one combines a statistical model that responds to the environmental inputs with an ODE model that triggers the flowering.