Report from Dagstuhl Seminar 16261

Integration of Expert Knowledge for Interpretable Models in Biomedical Data Analysis

Edited by

Gyan Bhanot[1], Michael Biehl[2], Thomas Villmann[3], and Dietlind Zühlke[4]

1 **Rutgers University – Piscataway, US,** gbhanot@rutgers.edu
2 **University of Groningen, NL,** m.biehl@rug.nl
3 **Hochschule Mittweida, DE,** thomas.villmann@hs-mittweida.de
4 **Seven Principles AG – Köln, DE,** dietlind.zuehlke@doc-z.de

Abstract

This report documents the talks, discussions and outcome of the Dagstuhl seminar 16261 "Integration of Expert Knowledge for Interpretable Models in Biomedical Data Analysis". The seminar brought together 37 participants from three diverse disciplines, who would normally not have opportunities to meet in such a forum, let alone discuss common interests and plan joint projects.

Seminar June 26 to July 1, 2016 – http://www.dagstuhl.de/16261
1998 ACM Subject Classification J.3 [Life and Medical Sciences] Biology, E.1 Data Structures, F.2 Analysis of Algorithms and Problem Complexity, I.6.4 Model Validation and Analysis, I.6.5 Model Development
Keywords and phrases Biomedical data analysis, Data visualization, Expert interactions, Feature selection and dimensionality reduction, Knowledge integration, Modeling
Digital Object Identifier 10.4230/DagRep.6.6.88
Edited in cooperation with Friedrich Melchert

1 Executive Summary

Gyan Bhanot
Michael Biehl
Thomas Villmann
Dietlind Zühlke

License Ⓒ Creative Commons BY 3.0 Unported license
© Gyan Bhanot, Michael Biehl, Thomas Villmann, and Dietlind Zühlke

The participants were drawn from three distinct disciplines: Biomedical Research, Machine Learning and Visualizations. On the first day, three overview talks on different aspects of bio-medical research were presented, including an overview of omics and clinical data and databases, a summary of current problems in cancer prognosis and metastasis, and steroid metabolomics and its relevance to disease. On the next two days, there were four overview talks on computer science topics, including machine learning, modeling and visualization. Participants also had the opportunity to give shorter presentations of their current research areas and describe open problems, as well as introduce new and relevant datasets and methods. In total, 16 such short talks were presented, covering various areas of biomedical research and computer science. All talks served as starting points for extensive plenary and individual

Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

Integration of Expert Knowledge for Interpretable Models in Biomedical Data Analysis, *Dagstuhl Reports*, Vol. 6, Issue 6, pp. 88–110
Editors: Gyan Bhanot, Michael Biehl, Thomas Villmann, and Dietlind Zühlke
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

evening and after dinner discussions about the integration of expert knowledge into data analysis and modeling, specifically targeted to cancer informatics. From these discussions, it was clear that there was an urgent need for interactive collaboration to foster successful analysis and interpretation of biomedical data and the success of such collaboration would hinge on active participation from domain experts from biomedical research, data mining and visualization.

Motivated by this conclusion, we identified a joint project in cancer genomics, which would exploit the expertise represented by the seminar participants. On the fourth day, participants discussed the interactive methodology we will follow in the project. Following this, first results obtained by analysis of cancer data from The Cancer Genome Atlas was presented in a joint talk by representatives from all three disciplines (biology, machine learning, visualization). We will extend this project further in the coming months with active participation from the clinicians and computer scientists. The goal of this effort is not just to solve a relevant and outstanding problem in cancer biology but also to work towards publication of our findings in a high-impact journal authored by all participants. To foster this project, we will establish a Wiki, which will serve as a platform for collaboration and communication.

The participants gave feedback on Friday on the organization and content of the seminar. All participants were appreciative of the open, friendly and constructive atmosphere that made learning and insight possible for experts from very diverse disciplines. Getting to know the basic methods used in each field was seen as the perfect starting point for future collaborations. The idea of a joined wiki page as a collaboration platform as well as the already started joined project were highlighted as especially important. Follow-up-meetings of newly formed interdisciplinary teams were initiated and planned e.g. one in Copenhagen. The participants were very enthusiastic about having a further meeting after about a year to discuss results and new directions resulting from the joint project initiated here. Apart from working on a specific project in cancer biology, the goal of the collaboration is to establish a methodology for interactions, disseminate ideas and protocols among the disciplines and establish a common language to foster understanding.

In summary, biologists, both medical and computational experts in the seminar are enthusiastic about joining forces to solve outstanding problems in understanding biological processes. Many of the machine learning methods presented by participants are ready to be applied in real environments such as in clinical use or in research laboratories, after proper technology transfer. Such technology transfer requires targeted funding and agreed upon protocols to ensure adequate resources and necessary quality control, for subsequent release to the community.

The participants felt that influential members in each community should seek opportunities and avenues to urge the appropriate agencies (NIH, NFS, EU Scientific bodies) to establish a targeted program for technology transfer of computational solutions to challenges in the interpretation of biomedical data. Such a program would solicit competitive funding proposals from groups consisting of both biomedical and computational experts, and require products that are rigorously demonstrated on real problems, as well as satisfy appropriate coding and user interface standards, and where appropriate, satisfy requirements of interfacing or integration with existing established systems currently in use by the community.

In medicine the data is treasure
Whose value's beyond any measure
But it is not surprising
That without analysing
Acquisition is meaningless pleasure

(Michael Biehl and Gyan Bhanot)

2 Table of Contents

Executive Summary

Overview of Talks

Working groups

3 Overview of Talks

3.1 Mutations and Immune Therapy

Gyan Bhanot (Rutgers University – Piscataway, US)

License 🅭 Creative Commons BY 3.0 Unported license
© Gyan Bhanot

In this talk, I presented an overview and results from collaborative work I am involved in in an exciting new area of cancer research, which is in the treatment of patients using immune checkpoint blockade therapy. Using data from The Cancer Genome Atlas (TCGA), we find that it is possible to identify a robust mutation burden threshold, which we call the immune Checkpoint Activating Mutation (iCAM) threshold, which can identify subsets of patients likely to respond to immune checkpoint therapy in melanoma, lung, colon, endometrial, stomach, ovarian, cervical, bladder, and breast cancers. Further, We also find that iCAM+ (responsive) patients can be identified with good accuracy using commercially available clinical-grade tumor sequencing assays. Finally, we find that iCAM+ and iCAM- tumors have different underlying mutation patterns, suggesting distinct underlying mechanisms of mutagenesis. The goal of the talk was to explain an exciting new field in cancer biology in simple language to the community of computational experts in the audience so that they could apply some of their expertise to the analysis of diverse and multimodal cancer datasets that are available in the public domain.

3.2 Molecular Landscapes in Renal Cell Carcinoma

Aguirre de Cubas (Vanderbilt University – Nashville, US) and W. Kimryn Rathmell (Vanderbilt University School of Medicine, US)

License 🅭 Creative Commons BY 3.0 Unported license
© Aguirre de Cubas and W. Kimryn Rathmell

A common urological malignancy, renal cell carcinoma is a heterogeneous disease, consisting of a number of different cancers, characterized by different histologies, genetic drivers, clinical course, and response to therapy. The majority of RCC's are defined by three major histologic subtypes: 75% are clear cell RCC (CCRCC, KIRC), 15-20% Type 1 or Type 2 papillary RCC (PRCC, KIRP), and 5% chromophobe RCC (ChRCC, KICH). This study evaluated gene expression by RNA-sequencing in 843 TCGA-RCCs (488 KIRC, 274 KIRP, 81 KICH) across histologic subtypes to identify features unique to each subtype and common to multiple subtypes, and included normal kidney tissue samples (n=129). For this, we applied weighted gene coexpression network analysis (WGCNA), a systems biology approach, and identified modules of closely connected co-expressed genes, which may act in a network and may serve as molecular signatures for an underlying phenotype.

WGCNA of mRNA expression data revealed strong immune and vasculature-related gene signatures associated with KIRC, and a significant loss of an ion transport and baso-lateral gene signatures and gain of mitochondrial membrane gene signatures in KIRP and KICH, but not KIRC. All RCCs had upregulation of the ribosomal gene signature compared to normal kidney tissue. These comparisons demonstrate clear genomic differences between the RCC histologic subtypes, as well as common features that unify some or all of the subtypes, which could provide the evidences for novel directions for developing targeted therapeutic interventions that could be effective across subtypes.

3.3 Information richness and data missingness in clinical studies for Heart Failure

Gert-Jan de Vries (Philips Research Lab. – Eindhoven, NL)

License Ⓒ Creative Commons BY 3.0 Unported license
© Gert-Jan de Vries

Prospective clinical studies typically yield a rich set of patient parameters. This, however, also increases chances of missing some parameters for some patients. Our study into chronic heart failure also shows various patterns of data missingness. Preliminary results from our structured study of the efficacy of various data imputation techniques seem to indicate that there is limited added value of more sophisticated imputation techniques over the relatively simple techniques with respect to the performance of classifiers of 30-day readmission. Other challenges for successful application of such predictive methods in clinical practice include the inherent selection bias that occurs in prospective studies. We showed that this bias has large effect on the generalizability of predictive models and should not be overlooked in model development.

3.4 Feature extraction for x-ray scattering from biomolecules and nanoparticles.

Sebastian Doniach (Stanford University, US)

License Ⓒ Creative Commons BY 3.0 Unported license
© Sebastian Doniach

To obtain information about the structure of biological molecules in terms of atomic positions in the molecule, x-ray crystallography is the principal tool. But what happens if the circumstances are such that a crystal of the molecules of interest is not available? Here, x-ray scattering from a solution of the molecule of interest gives a one-dimensional scattering intensity $I(q)$, as an average scattering over all orientations of the molecule, expressed as a function of the scattering wave vector q defined in terms of the scattering angle, θ.

Features of the molecular structure may be extracted from this function by a variety of methods of which the simplest is the Guinier method which measures the radius of gyration, R_g of the molecule in terms of the Gaussian character of the scattering function at small scattering angles $I(q) \sim \exp(-q^2 R_g^2/3)$. This provides a model-free method for extraction of the feature parameter R_g by fitting the data to the Guinier function at sufficiently small scattering angles. The non-linear nature of this function makes this a "super-resolution" method in which the spatial resolution of the resulting R_g is 2-3 orders of magnitude more accurate than would be given by linear Fourier analysis of the data [1].

In the last few years, the advent of x-ray free electron lasers (xFELs) has now made possible a method to obtain 3-dimensional atomic structures from scattering data of a solution of molecules in the xFEL x-ray beam. In this technique, each flash of the laser beam provides a very high intensity of x-rays in a few 10s of femtoseconds. The scattered x-rays are captured on a 2-dimensional detector as an image which is azimuthally isotropic around the direction of the incoming x-ray beam.

Angular correlators of this image as a function of the azimuthal angle, ϕ, around scattering rings of fixed theta, define a 3-dimensional correlation function $C(q_1, q_2, \psi)$. Here q_1 and

q_2 denote two possible scattering angles, with $\cos(\psi)$ being the angular projection between scattering vectors defined by two points on the scattered image [2].

It may be shown that the average of $C(q_1, q_2, \psi)$ over all orientations of the molecules provides a function from which the atomic structure of the molecules in the solution may in principle be extracted as feature parameters [2]. In practice the experimental determination of the average correlator requires averaging over many thousands of images in order to reduce the background noise due to uncorrelated scattering events [3, 4].

The next step in feature extraction also involves determination of scattering phases as in x-ray crystallography. As is fundamental to image analysis in general, the phase retrieval problem is NP complete and relies on prior knowledge of the chemistry of the molecule, such as is used in molecular replacement, to determine a final structure.

Our group has published a first proof of principle for this method, applied to the example of a disordered suspension of 60 nanometer gold nanoparticles (gold NPs) [3]. A detailed report of xFEL scattering measurements from suspensions of gold NPs has been submitted for publication and is in the refereeing process [4].

References

1 Lipfert, J., Doniach, S. (2007). Small-Angle X-Ray Scattering From RNA, Proteins, And Protein Complexes. Annual Rev. Of Biophysics And Biomolecular Structure 36, 307–327.

2 Kam, Z. (1977) Determination of macromolecular structure in solution by spatial correlation of scattering fluctuations. Macromolecules 10, 927–934.

3 Mendez, D. et al. (2014). Observation of correlated x-ray scattering at atomic resolution. Phil. Trans.R. Soc. B 369, 20130315.

4 Mendez, D. et al. (2016) Angular correlations of photons from solution diffraction at an x-ray free electron laser encode molecular structure. submitted for publication.

3.5 Feature selection, relevance learning, and causality

Barbara Hammer (Universität Bielefeld, DE)

License 🅭 Creative Commons BY 3.0 Unported license
© Barbara Hammer

Machine learning offers valuable methods to infer a regularity from given data, such as a classification prescription for a disease based on characteristic patterns from gene expression levels, SNPs, or any other signal. Formally, given a sufficient amount of training data, classification models can be accompanied by strong guarantees as concerns their validity, such as its performance as evaluated in a cross validation. In practical applications, however, this is often not enough: the models typically act as black boxes and they do not allow an insight into why a classification has been taken. We would like to take a look at a few techniques which enable practitioners to partially open the black box.

Feature selection

Feature selection aims for a selection of features from the given data which are particularly relevant for the given task at hand. Typically, feature relevance is measured by the impact of a feature on the given classifier as concerns its overall classification accuracy, i.e. the features which contribute most to the given classification are selected. There exists a large variety of feature selection techniques together with different software suites [1]. Typically, one distinguishes filter methods (the features are selected by adding relevant ones or deleting irrelevant ones from the full set of features as measured by some general criterion such as e.g. mutual information), wrapper methods (features are added / deleted depending on the performance of a classifier based on the remaining feature set), and embedded methods (feature selection is embedded into the classifier learning e.g. by adding sparsity constraints.)

While feature selection constitutes an ubiquitous technology in classification learning, it faces a few drawbacks: feature selection constitutes a so-called NP-hard problem, i.e. it is not possible to efficiently and reliably select all relevant features from a given data set in case of a large number of features. One particular challenge is faced when features are correlated, and important information is only revealed when a set of features is considered. Further, feature selection typically takes a binary decision, i.e. a feature is selected or discarded, but it is not weighted according to its relevance. Hence ambiguities or different degrees of feature relevance cannot easily be distinguished.

Relevance learning

Relevance learning aims for a solution of these problems by explicitly assigning a real-valued relevance to every feature. This takes into account correlated features and different degrees of feature relevance, since their weighting can be simultaneously and smoothly be adjusted. One particularly elegant way for an efficient relevance learning is offered by its direct embedding into metric based algorithms, such as proposed in [2]. Linear feature weighting essentially corresponds to a change of the data metric, the Euclidean metric is substituted by a general positive semidefinite quadratic form. Hence every metric learning methods can be used for relevance learning, such as successfully demonstrated for so-called generalised matrix learning vector quantisation in challenging biomedical applications [3].

While relevance learning is capable of simultaneously weighting several correlated features, it does not solve the challenges of feature redundancy: for redundant features, practitioners can pick from a set of equal features based on other criteria such as the feature costs. Recently, a very elegant scheme how to judge such settings and how to offer practitioners an interactive tool for an appropriate feature choice has been presented based on a suitable L-1 regularisation [4].

Causality

While relevance learning is capable of efficiently dealing with feature correlations, it restricts the analysis to a mere statistical one: it measures correlation of features rather than causality. While correlated features can be interwoven just based on coincidence, a bias in the measurement, or a hidden cause, causal feature relation characterises the fact that a measurement A causes measurement B, e.g. a genetic marker A can cause a certain disease B. Causality is typically inferred from data gathered in experiments, by manipulating probabilities [5]: A is cause of B if the probability of B is a different one when A is manipulated (but everything else is kept constant) in an experiment. Correlation, on the contrary, only refers to the fact that the probability of B is observed differently, given different instantiations of A, but B itself might be causes by yet another, possibly unobserved C.

Interestingly, under some assumptions, there do exist techniques to infer causality solely based on given measurement data, thus weakening the burden of an additional (possibly impossible or not ethical) experiment [7]. Given two correlated features, it is provably possible to determine the direction of the cause under special conditions such as independence of noise, or nonlinearity of the signals [6, 8, 9, 10]. Given a number of measurements, their overall causality can be determined by an efficient investigation of so-called Markov blankets, such as popular for Bayes network inference [12]. Interestingly, the resulting structure can be beneficial for an increased classification accuracy in biomedical applications, besides their increased interpretability [13]. Still, causal inference constitutes an ongoing research topic with quite a number of challenges ahead, such as cyclic relations [11], and their potential for biomedical applications is not yet fully explored.

Acknowledgement. Funding by the CITEC center of excellence is gratefully acknowledged.

References

1 Yvan Saeys, Inaki Inza, Pedro Larranaga: A review of feature selection techniques in bioinformatics. Bioinformatics 23(19): 2507–2517 (2007)

2 Petra Schneider, Michael Biehl, Barbara Hammer: Adaptive Relevance Matrices in Learning Vector Quantization. Neural Computation 21(12): 3532–3561 (2009)

3 W. Arlt, M. Biehl, A. E. Taylor, S. Hahner, R. Libe, B. A. Hughes, P. Schneider, D. J. Smith, H. Stiekema, N. Krone E. Porfiri, G. Opocher, J. Bertherat, F. Mantero, B. Allolio, M. Terzolo, P. Nightingale, C. H. L. Shackleton, X. Bertagna, M. Fassnacht, P. M. Stewart, Urine steroid metabolomics as a biomarker tool for detecting malignancy in adrenal tumors, J. of Clinical Endocrinology & Metabolism 96:3775–3784 (2011)

4 Alexander Schulz, Bassam Mokbel, Michael Biehl, Barbara Hammer: Inferring Feature Relevances From Metric Learning. SSCI 2015:1599–1606

5 Constantin F. Aliferis, Alexander R. Statnikov, Ioannis Tsamardinos, Subramani Mani, Xenofon D. Koutsoukos: Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation. Journal of Machine Learning Research 11:171–234 (2010)

6 Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, Bernhard Schölkopf: Distinguishing cause from effect using observational data: methods and benchmarks. CoRR abs/1412.3773 (2014)

7 Peter Spirtes: Introduction to Causal Inference. Journal of Machine Learning Research 11: 1643-1662 (2010)

8 Jonas Peters, Joris M. Mooij, Dominik Janzing, Bernhard Schölkopf: Causal discovery with continuous additive noise models. Journal of Machine Learning Research 15(1):2009–2053 (2014)

9 P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In Advances in Neural Information Processing Systems 21 (NIPS 2008), pp. 689–696, 2009.

10 Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, Kenneth Bollen: DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model. Journal of Machine Learning Research 12: 1225–1248 (2011)

11 Gustavo Lacerda, Peter Spirtes, Joseph Ramsey, Patrik O. Hoyer: Discovering Cyclic Causal Models by Independent Components Analysis. CoRR abs/1206.3273 (2012)

12 Markus Kalisch, Peter Bühlmann: Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. Journal of Machine Learning Research 8:613–636 (2007)

13 Mei Liu, Ruichu Cai, Yong Hu, Michael E. Matheny, Jingchun Sun, Jun Hu, Hua Xu: Determining molecular predictors of adverse drug reactions with causality analysis based on structure learning. JAMIA 21(2):245–251 (2014)

3.6 Statistical inference and dimensionality reduction in evolutionary tree spaces

Hossein Khiabanian (Rutgers Cancer Inst. of New Jersey – New Brunswick, US)

License 🅭 Creative Commons BY 3.0 Unported license
© Hossein Khiabanian

Joint work of Zairis, Khiabanian, Blumberg, Rabadan
Main reference S. Zairis, H. Khiabanian, A. Blumberg, R. Rabadan, "Moduli Spaces of Phylogenetic Trees Describing Tumor Evolutionary Patterns", in Proc. of the 2014 Int'l Conf. on Brain Informatics and Health (BIH'14), LNCS, Vol. 8609, pp. 528–539, Springer, 2014; pre-print available as arXiv:1410.0980 [q-bio.QM].
URL http://dx.doi.org/10.1007/978-3-319-09891-3_48
URL http://arxiv.org/abs/1410.0980

Phylogenetic trees are arguably the most common representation of evolutionary processes; they have been frequently used to characterize pathogen spread, the relationship between different species, and the evolution of cancers. Comparison between different trees is a key part of the analysis of evolutionary phenomena, for instance in comparing the evolutionary trajectories of tumors in different patients in relation to their differential response to therapy.

Recently, an elegant approach has been introduced by Billera-Holmes-Vogtmann that allows a systematic comparison of different evolutionary histories using the metric geometry of tree spaces. In this manuscript, we begin by reviewing in detail the relevant mathematical and computational foundations for applying standard techniques from machine learning and statistical inference in these spaces, which we refer to as evolutionary moduli spaces.

The solutions of biological problems often deal with heavily populated phylogenetic trees with many leaves, which are very cumbersome to visualize and analyze in their relevant evolutionary moduli spaces. To address this issue, we introduce *tree dimensionality reduction*, a structured approach to reduce large and complex phylogenetic trees to a distribution of smaller trees. We prove a stability theorem ensuring that small perturbations of the large trees are taken to small perturbations of the resulting distributions.

We then present a series biologically motivated applications to the analysis of genomic data, spanning cancer and infectious disease. The first quantifies how chemotherapy can disrupt the evolution of common leukemias. The second examines a link between geometric information and the histologic grade in relapsed gliomas, where longer relapse branches were specific to high grade glioma. The third concerns genetic stability of xenograft models of cancer, where heterogeneity at the single cell level increased with later mouse passages. The last studies genetic diversity in seasonal influenza A virus. We apply tree dimensionality reduction to project 24 years of longitudinally collected H3N2 hemagglutinin sequences into distributions of smaller trees spanning between three and five seasons. A negative correlation is observed between the influenza vaccine effectiveness during a season and the variance of the distributions produced using preceding seasons' sequence data. We also show how tree distributions relate to antigenic clusters and choice of influenza vaccine. These results provide compelling evidence that our formalism exposes links between genomic data of influenza A and important clinical observables, namely vaccine selection and efficacy.

3.7 Integrative Genomics Analysis Identify Molecular Subtypes of Lung Carinoids

Saurabh V. Laddha (Rutgers Cancer Inst. of New Jersey – New Brunswick, US)

License Ⓒ Creative Commons BY 3.0 Unported license
© Saurabh V. Laddha

Joint work of Saurabh V. Laddha, Edaise Da Silva, Kenneth Robyzk, Brian R. Untch, Laura H.Tang, Chang S. Chan

Lung carcinoids are a rare and slow growing type of primary lung neoplasms. Recent genomics studies provided more detailed biomarker/molecular subtype classifications of cancers, which bring us to a possibility to accurately classify based on molecular alterations and cell type. Here, we investigated the genomic and molecular alterations to discover subtypes of lung carcinoids (includes 13 atypical and 17 typical samples only) with distinct biological characteristics and identify a set of novel gene signature and biomarkers. We perform targeted sequencing of 354-cancer gene panel (n=29), Transcriptome (n=30) and methylome (n=18) on Lung carcinoids and identified recurrent mutated genes enriched in Histone covalent modifier/Chromatin remodeler, DNA repair and protein kinases pathways. We found 10 samples (30%) with mutations in covalent Histone modifier and SWI complexes with MEN1 and ARID1A being recurrently mutated. Unsupervised clustering and dimensionality reduction were used on expression and methylation level data resulted in 3 robust subtypes from both dataset: Subtype 1(SB1), Subtype 2(SB2) and Subtype 3(SB3). We speculate this subtyping may originate from different cell types and cellular mechanisms. Integrative multilayered data revealed SB2 group samples are enriched for MEN1 gene mutations. Samples in SB1 and SB2 are combination of both typical and atypical carcinoids but interestingly SB3 is enriched for typical carcinoids. Also SB3 is enriched (pval=0.001) for endobronchial lung carcinoids and has better recurrence free survival (Pval=0.003). We found gene and methylation signatures, which alone selectively discriminate subtypes. Immunohistochemistry on selected four key biomarkers stratified subtypes clearly at the protein level (ASCL1 biomarker for SB1, S100 for SB2, ALDH1A1 for SB2 and 3, TTF1 for SB1 and 3) and matches with integrative mutations, gene expression and methylation level classifications for novel Lung carcinoids subtypes.

3.8 Automatic organ delineation of medical images with atlases based on machine learning

John A. Lee (UC Louvain-la-Neuve, BE)

License Ⓒ Creative Commons BY 3.0 Unported license
© John A. Lee

Radiation oncology (RO) treats cancer patient with beams of ionising particles like photons, protons, or even heavier ions. One or several beams are aimed at the tumor (or target volumes), in order to kill cancerous cells. To reach the tumor, these beams typically have to pass through surrounding healthy tissues, including organs at risk, the irradiation of which is the source of secondary effects. Very often, RO is thus compared to a ballistic problem: a tradeoff must be reached between tumor local control and side effects. To this end, medical images are acquired, in order to focus the beams and concentrate the irradiation dose primarily on the target. This evaluation requires to delineate both the target volumes and the organs at risk. While some medical expertise is required for the former ones, delineation of the latter could be automated, at least in theory, in order to spare time during treatment planning. Currently, the main approach of the problem is based on so-called atlases. The atlas consists of one or several template images of some patients, which are delineated beforehand. When the image of a new patient comes in and must be delineated, the atlas is deformed in a non-rigid way to match the new image and the deformed contours are then propagated. This approach is not entirely successful, mainly because it is difficult to model complex deformation between the anatomies of two different patients. Another approach that we have developed consists in training classifiers that learns to label pixels with the correct organ label. Preliminary experiments have been conducted, relying on pre-segmentation of the image into superpixels and computation of features. Then organs are tagged in an iterative/incremental way, one by one. First results show that this method works and can compete with atlases, although it requires more organ labels than these. A new project will start with a single classifier, in the form of a deep convolutional neural network.

3.9 Insight generation from biomedical data with flexible models

Paulo J. Lisboa (John Moores University – Liverpool, GB)

License 🅭 Creative Commons BY 3.0 Unported license
© Paulo J. Lisboa

The talk focused on the need to interpret biomedical data models to gain insights and ascertain validity. To this end, specific methodologies were considered. First, flexible models of time-to-event data were shown to model the hazard ratios without recourse to assumptions of proportionality that are typically made in linear models. The resulting survival models give insights into progression of disease over time, with single and competing risks. Second, pn data typical of bioinformatics are susceptible to significant false positive errors and so require the use of specific statistical methodologies for false detection rate control. This is important both for classification and for clustering. Clustering models can also be sensitive to parameter choices, requiring the use of stability measures to ensure that the resulting clusters are robust. These methodologies were outlined and complemented by linear supervised visualization methods which are particularly suitable visualized high-dimensional clusters in low dimensions with little mixing. This was followed by a brief summary of convex non-negative matrix factorization methods to separate linear mixed signals, which was shown to have important medical applications fundamental signal processing method. Finally, the use of information geometry was shown to open-up non-liner classifiers by mapping them onto data structures, which provides a principled approach to patients-like-you representations for access and retrieval of relevant reference cases.

3.10 The stories told by breast cancer whole genome sequencing

John Martens (Erasmus University – Rotterdam, NL)

License Ⓒ Creative Commons BY 3.0 Unported license
© John Martens

Joint work of On behalf of the FP7-BASIS consortium
Main reference S. Nik-Zainal, H. Davies, J. Staaf, M. Ramakrishna, D. Glodzik, X. Zou, I. Martincorena, L. B. Alexandrov, S. Martin, D. C. Wedge, P. Van Loo, Y. S. Ju, M. Smid, A. B. Brinkman, S. Morganella, M. R. Aure, O. C. Lingjærde, A. Langerød, M. Ringnér, S.-M. Ahn, S. Boyault, J. E. Brock, A. Broeks, A. Butler, C. Desmedt, L. Dirix, S. Dronov, A. Fatima, J. A. Foekens, M. Gerstung, G. K. J. Hooijer, S. J. Jang, D. R. Jones, H.-Y. Kim, T. A. King, S. Krishnamurthy, H. J. Lee, J.-Y. Lee, Y. Li, S. McLaren, A. Menzies, V. Mustonen, S. O'Meara, I. Pauporté, X. Pivot, C. A. Purdie, K. Raine, K. Ramakrishnan, F. G. Rodríguez-González, G. Romieu, A. M. Sieuwerts, P. T. Simpson, R. Shepherd, L. Stebbings, O. A. Stefansson, J. Teague, S. Tommasi, I. Treilleux, G. G. Van den Eynden, P. Vermeulen, A. Vincent-Salomon, L. Yates, C. Caldas, L. van't Veer, A. Tutt, S. Knappskog, B. K. T. Tan, J. Jonkers, Å. Borg, N. T. Ueno, C. Sotiriou, A. Viari, P. A. Futreal, P. J. Campbell, P. N. Span, S. Van Laere, S. R. Lakhani, J. E. Eyfjord, A. M. Thompson, E. Birney, H. G. Stunnenberg, M. J. van de Vijver, J. W. M. Martens, A.-L. Børresen-Dale, A. L. Richardson, G. Kong, G. Thomas, and M. R. Stratton, "Landscape of somatic mutations in 560 breast cancer whole-genome sequences", in Nature, Vol. 534, pp. 47–54, 2016.
URL http://dx.doi.org/10.1038/nature17676

The analysis of over 1100 breast cancer exome and over 500 whole-genome sequences has advanced our understanding of breast cancer in several ways. First, we now know that at least 93 protein-coding cancer genes carry probable somatically acquired driver mutations. And every breast cancer carries a median of about 3-4 driver gene mutations (range 1-8) in multiple signaling pathways relevant for the genesis and/or progression of breast cancer. Second, twelve different base substitution patterns are observed likely representing 12 different mechanisms responsible for the somatically acquired base substitutions in the breast cancer genome. The most prominent forces include 2 different age-related and 2 different APOBEC-driven mutational substitution patterns. Age-related patterns have been ascribed to unavoidable replication errors occurring in mammary epithelial cells during tissue maintenance; APOBEC-driven patterns are most likely due APOBEC enzymes (APOBEC3B and at least one other family member). The natural role of APOBEC enzymes is to protect tissue against e.g. viral infection mutagenizing the viral genome but for currently unknown reasons these enzymes are erroneously expressed in breast cancer harassing their own genome. Luminal breast cancer overexpressing APOBEC3B, and as a result having a higher mutational load, have an adverse outcome and are less likely to respond to endocrine therapy. Third, and on top of the somatically acquired substitutions, the breast cancer genome is characterized by several different types of rearrangements. Three rearrangement types, characterized by tandem duplications or deletions, appear associated with defective homologous-recombination-based DNA repair: one with deficient BRCA1 function, another with mostly deficient BRCA2 function, and a third of which the genetic cause remains unknown. Finally, intra-tumoral mutational heterogeneity is present in almost every primary breast cancer and usually comprises one dominant tumor clone but next to that several clonally related tumor subclones. Sequencial analysis of samples subsequently revealed that (pre-existent) tumor subclones can be resistant to neo-adjuvant chemotherapy while the dominant clone of the primary tumor is highly responsive. Also, rather than the dominant clone, a hardly detectable subclone can be responsible for the local or distant metastases. The clinical significance of most of these novel observations remains to be understood but we can speculate on it during the discussion.

3.11 Visual Reasoning with High-Dimensional Data

Klaus Mueller (Stony Brook University, US)

License Ⓒⓒ Creative Commons BY 3.0 Unported license
© Klaus Mueller

The growth of digital data is tremendous. Any aspect of life and matter is being recorded and stored on cheap disks, either in the cloud, in businesses, or in research labs. We can now afford to explore very complex relationships with many variables playing a part. But for this we need powerful tools that allow us to be creative, to sculpt this intricate insight formulated as models from the raw block of data. High-quality visual feedback plays a decisive role here. In this talk I will discuss various platforms we have developed over the years to make the exploration of large multivariate data more intuitive and direct. These platforms were conceived in tight collaborations with domain experts in the fields of climate science, health informatics, and computer systems.

3.12 Integrating and interpreting -omic features to reveal and define the complexity of human cancer

W. Kimryn Rathmell (Vanderbilt University School of Medicine, US)

License Ⓒⓒ Creative Commons BY 3.0 Unported license
© W. Kimryn Rathmell

Using renal cell carcinoma (RCC) as an example, recent high throughput sequencing efforts and analysis of gene expression datasets has revealed high levels of heterogeneity within this class of tumors. It has become clear that the renal cell carcinomas encompass multiple discrete diseases, originating from discrete regions of the kidney nephron, and with unique mutation and genomic attributes. Three recent publications of The Cancer Genome Atlas define the core features of these diseases: clear cell RCC, papillary RCC, and chromophobe RCC. A recent review summarizes these findings. Using expression data, we can subclassify these diseases further, for example the clear cell subtype can be assigned into clear cell A and B (ccA and ccB) subtypes. These subgroupings were developed as a part of an ongoing biology-informatics collaboration, ultimately leading to the development of meaningful subgroups that have important biological attributes. The ccA subgroup applies classical RCC expression features of angiogenesis and beta oxidation metabolism, and is a better prognosis group. The ccB subgroup, by comparison expresses genes associated with invasive growth, such as wnt and TGFbeta signaling. Not surprisingly, this set is independently associated with poor outcome. Many groups have validated these findings in multiple independent datasets, making this one of the only robust expression-based signatures useful in developing the prognosis of cancer. Work with these and other genomic datasets has begun to reveal the full complexity of this cancer, as subsampling has demonstrated that there is heterogeneity of the mutation profile within individual tumors, and has shown us that within one tumor there can co-exist numerous clones that have separate evolutionary phylogeny, and more disturbingly, that metastatic tumors can have even more divergency. Our exploration of the ccA and ccB subgroups shows that of these subsets, one often dominates, but that it is clear both subtypes can co-exist in a single tumor. Functional imaging using glucose uptake positron emission tomography can correlate metabolic signals with the ccB subtype feature set, allowing us to consider imaging-based strategies for knowing the state of disease in the whole individual. More recently, some of the major genetic features that have been linked with RCC include mutations in chromatin modifier genes. These defects play poorly understood roles in promoting tumorigenesis. Our recent work using alignment of shortread sequencing of non-protein bound fractional chromatin shows that major remodeling of the chromatin can occur, and may contribute to the overall evolution of the tumor. In particular, the loss of a single histone methyltransferase leads to the loss of one critical mark of actively transcribed genes, Histone H3 lysine 36 trimethyation. The ultimate impact of this feature disruption, which we have shown promotes tumorigenic phenotypes, on tumor progression molecularly is not certain. In summary, massive sequencing efforts have recently redefined the landscape of renal cell carcinoma, distinguishing between disparate disease types, shedding light on the intratumor heterogeneity and evolutionary strategies at play, and providing us with a new set of features that contribute to the development and progression of this set of cancers.

These findings leave us with an emerging list of questions:

- What additional features are deregulated in response to chromatin remodeling?
- Can modeling help us resolve the connections between histone marks, allowing us to "read" the sentences of the histone code, rather than merely the individual words?
- Recent work with immune checkpoint inhibitor therapy suggests this is effective in a significant proportion of patient. Is there any way to predict this response, or the durability of response?
- Can we use machine learning to understand the evolution of renal tumors?
- Can modeling help us predict the heterogeneity of a tumor, or the pattern of heterogeneity?
- How can we better quantify and use imaging parameters to help guide disease staging, prognosis, or therapeutic decision making?

References

1 Haake SM, Weyandt JD, Rathmell WK. *Insights into the Genetic Basis of the Renal Cell Carcinomas from the Cancer Genome Atlas (TCGA)*. Mol Cancer Res. 2016 Jun 21. pii: molcanres.0115.2016. [Epub ahead of print]

2 PMID: 27330105 Brooks SA, Khandani A, Fielding J, Lin W, Sills T, Lee Y, Arreola A, Milowsky MI, Wallen EM, Woods ME, Smith AB, Nielsen ME, Parker JS, Lalush DS, Rathmell WK. *Alternate metabolic programs define regional variation of relevant biological features in renal cell carcinoma progression*. Clin Cancer Res. 2016 Jun 15;22(12):2950–9. PMID: 26787754. PMCID: PMC4911278.

3 Haake SM, Brooks SA, Welsh E, Fulp WJ, Chen DT, Dhillon J, Haura E, Sexton W, Spiess PE, Pow-Sang J, Rathmell WK, Fishman M. *Patients with ClearCode34-identified molecular subtypes of clear cell renal cell carcinoma represent unique populations with distinct comorbidities*. Urol Oncol. 2016 Mar;34(3):122. PMID: 26546482 PMCID: PMC4761468

4 Linehan WM, Spellman PT, Ricketts CJ, Creighton CJ, Fei SS, Davis C, Wheeler DA, Murray BA, Schmidt L, Vocke CD, Peto M, Al Mamun AA, Shinbrot E, Sethi A, Brooks S, Rathmell WK, Brooks AN, Hoadley KA, Robertson AG, Brooks D, Bowlby R, Sadeghi S, Shen H, Weisenberger DJ, Bootwalla M, Baylin SB, Laird PW, Cherniack AD, Saksena G, Haake S, Li J, Liang H, Lu Y, Mills GB, Akbani R, Leiserson MD, Raphael BJ, Anur P, Bottaro D, Albiges L, Barnabas N, Choueiri TK, Czerniak B, Godwin AK, Hakimi AA, Ho TH, Hsieh J, Ittmann M, Kim WY, Krishnan B, Merino MJ, Shaw KR, Reuter VE, Reznik E, Shelley CS, Shuch B, Signoretti S, Srinivasan R, Tamboli P, Thomas G, Tickoo S, Burnett K, Crain D, Gardner J, Lau K, Mallery D, Morris S, Paulauskis JD, Penny RJ, Shelton C, Shelton WT, Sherman M, Thompson E, Yena P, Avedon MT, Bowen J, Gastier-Foster JM, Gerken M, Leraas KM, Lichtenberg TM, Ramirez NC, Santos T, Wise L, Zmuda E, Demchok JA, Felau I, Hutter CM, Sheth M, Sofia HJ, Tarnuzzer R, Wang

Z, Yang L, Zenklusen JC, Zhang J, Ayala B, Baboud J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Ally A, Balasundaram M, Balu S, Beroukhim R, Bodenheimer T, Buhay C, Butterfield YS, Carlsen R, Carter SL, Chao H, Chuah E, Clarke A, Covington KR, Dahdouli M, Dewal N, Dhalla N, Doddapaneni HV, Drummond JA, Gabriel SB, Gibbs RA, Guin R, Hale W, Hawes A, Hayes DN, Holt RA, Hoyle AP, Jefferys SR, Jones SJ, Jones CD, Kalra D, Kovar C, Lewis L, Li J, Ma Y, Marra MA, Mayo M, Meng S, Meyerson M, Mieczkowski PA, Moore RA, Morton D, Mose LE, Mungall AJ, Muzny D, Parker JS, Perou CM, Roach J, Schein JE, Schumacher SE, Shi Y, Simons JV, Sipahimalani P, Skelly T, Soloway MG, Sougnez C, Tam A, Tan D, Thiessen N, Veluvolu U, Wang M, Wilkerson MD, Wong T, Wu J, Xi L, Zhou J, Bedford J, Chen F, Fu Y, Gerstein M, Haussler D, Kasaian K, Lai P, Ling S, Radenbaugh A, Van Den Berg D, Weinstein JN, Zhu J, Albert M, Alexopoulou I, Andersen JJ, Auman JT, Bartlett J, Bastacky S, Bergsten J, Blute ML, Boice L, Bollag RJ, Boyd J, Castle E, Chen YB, Cheville JC, Curley E, Davies B, DeVolk A, Dhir R, Dike L, Eckman J, Engel J, Harr J, Hrebinko R, Huang M, Huelsenbeck-Dill L, Iacocca M, Jacobs B, Lobis M, Maranchie JK, McMeekin S, Myers J, Nelson J, Parfitt J, Parwani A, Petrelli N, Rabeno B, Roy S, Salner AL, Slaton J, Stanton M, Thompson RH, Thorne L, Tucker K, Weinberger PM, Winemiller C, Zach LA, Zuna R; Cancer Genome Atlas Research Network. *Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma*. N Engl J Med. 2015 Nov 4. [Epub ahead of print] PMID: 26536169

5 Begg CB, Seshan VE, Zabor EC, Furberg H, Arora A, Shen R, Maranchie JK, Nielsen ME, Rathmell WK, Signoretti S, Tamboli P, Karam JA, Choueiri TK, Hakimi AA, Hsieh JJ. *Genomic investigation of etiologic heterogeneity: methodologic challenges*. BMC Med Res Methodol. 2014 Dec 22;14(1):138. PMID: 25532962

6 Davis, CF, Ricketts, C, Wang, M, Yang, L, Cherniack, AD, Shen, H, Buhay, C, Kim, SC, Fahey, CC, Hacker, KE, Bhanot, G, Gordenin, DA, Chu, A, Gunaratne, PH, Biehl, M, Seth, S, Kaipparettu, BA, Bristow, CA, Donehower, LA, Wallen, EM, Smith, AB, Tickoo, SK, Tamboli, P, Reuter, V, Schmidt, LS, Hsieh, JJ, Choueiri, TK, Hakimi, AA, The Cancer Genome Atlas Research Network, Chin, L, Meyerson, ML, Kucherlapati, R, Park, W-Y, Robertson, AG, Laird, PW, Henske, EP, Kwiatkowski, DJ, Park, PJ, Morgan, M, Shuch, B, Muzny, D, Wheeler, DA, Linehan, WM, Gibbs, RA, Rathmell, WK, Creighton, CJ. *The somatic genomic landscape of chromophobe renal cell carcinoma*. Cancer Cell. 2014 26(3):319-330. PMCID PMC4160352.

7 Brooks SA, Brannon AR, Parker JS, Fisher JC, Sen O, Kattan MW, Hakimi AA, Hsieh JJ, Choueiri TK, Tamboli P, Maranchie JK, Hinds P, Miller CR, Nielsen ME, Rathmell WK. *ClearCode34: A Prognostic Risk Predictor for Localized Clear Cell Renal Cell Carcinoma*. Eur Urol. 2014 66(1):77–84 PMC4058355

8 Simon, J, Hacker, KE, Brannon AR, Parker JS, Weiser M, Ho TH, Kuan PF, Jonasch E, Furey TS, Prins JF, Lieb, JD, Rathmell, WK, Davis, IJ. *Variation in chromatin accessibility in human kidney cancer links H3K36 methyltransferase loss with widespread RNA processing defects*. Genome Research. 2014 24(2):241-50. PMCID: PMC3912414

9 The Cancer Genome Atlas Network. *Comprehensive molecular characterization of clear cell renal cell carcinoma*. Nature, 499(7456):43-9, 2013. PMID:23792563.

3.13 Integrating Expert Knowledge through Interactive Visualization

Timo Ropinski (Universität Ulm, DE)

License ⓒ Creative Commons BY 3.0 Unported license
© Timo Ropinski

Analyzing and interpreting imaging data is crucial for many disciplines in science, such as medicine or biology. In the past, two trends could be observed which helped to deal with the challenges involved in the image analysis process. On the one hand, advanced computer vision and image processing algorithms have been applied in order to extract information directly from the data. On the other hand, visualization algorithms have been exploited, which generate expressive representations of the data that allow the user to extract the relevant information by forming a mental model. Both approaches benefit from different capabilities. When applying computer vision based techniques, high throughput computing can be facilitated, while visualizations allow to exploit the robust pattern recognition capabilities of the human observer. While computing is performed with high accuracy and speed, visualization allows to better cope with noise as well as uncertainty and can also facilitate domain knowledge in a more direct way. In my talk I have presented synergies between the two approaches, and delineated compute capabilities and observe capabilities. By taking into account both, a more effective knowledge extraction from imaging data becomes possible. The discussed concepts are demonstrated with several examples from different compute intensive disciplines.

3.14 Learning in the space of models

Peter Tino (University of Birmingham, GB)

License ⓒ Creative Commons BY 3.0 Unported license
© Peter Tino

I will first introduce the general concept of the emerging field of "learning in the model space". The talk will then focus on time series data. After reviewing some of the existing model based time series kernels, I will introduce a framework for building new kernels based on temporal filters inspired by a class of parametrized state space models known as "reservoir" models. I will briefly outline the key theoretical concepts of their analysis and design. The methodology will be demonstrated in a series of sequence classification tasks and in an incremental fault detection setting. Finally, the framework will be extended to the "parametric case", where curial aspects of the data gene ratio process are known and the observations are sparse and irregularly sampled.

3.15 Visual Exploration and Analysis of Brain Connectivity: Approaches and Challenges

Gunther H. Weber (Lawrence Berkeley National Laboratory, US)

License 🅭 Creative Commons BY 3.0 Unported license
© Gunther H. Weber

Joint work of Sugeerth Murugesan, Kristofer Bouchard, Jesse A. Brown, Edward Chang, Peter Denes, Max Dougherty, Bernd Hamann, William W. Seeley, Andrew Trujillo, Gunther H. Weber, Associate Member
Main reference S. Murugesan, K. Bouchard, J. A. Brown, B. Hamann, W. W. Seeley, A. Trujillo, G. H. Weber, "Brain Modulyzer: Interactive visual analysis of functional brain connectivity", in IEEE Transactions on Computational Biology and Bioinformatics, 2016, in press.
URL http://dx.doi.org/10.1109/TCBB.2016.2564970

The brain is a highly connected, dynamic system of specialized brain regions interacting in complex ways. Neuroscientists use a variety of technologies to explore connectivity in the human brain, including resting state functional magnetic resonance imaging (fMRI) brain scans and e lectrocorticography (ECoG). In this talk I present a combination of visualization techniques and methods from graph theory and analysis—such as community detection—for analyzing fMRI and ECoG data sets. Providing immediate feedback by displaying analysis results instantaneously while changing parameters gives neuroscientists a powerful means to comprehend complex brain structure more effectively and efficiently and supports forming hypotheses that can then be validated via statistical analysis. Furthermore, I discuss the challenges presented by visualizing clustering results and improved ECoG data acquisition technologies with a large number of recorded channels as well as higher sampling rates.

3.16 Cardiovascularvascular Diseases: From Data Generation to Analysis

Thomas Wischgoll (Wright State University – Dayton, US)

License 🅭 Creative Commons BY 3.0 Unported license
© Thomas Wischgoll

Cardiovascular diseases remain the leading cause of death in Western societies. This presentation will provide insight into the use of expert knowledge and models to derive diagnostic tools that have the potential to aid in the diagnosis of diffuse cardiovascular diseases that tend to be more difficult to detect in CT angiograms. In order to develop these methods more basic research is needed to prove the validity of the approach, including validation of accuracy as well as approach itself. For that, specimens of porcine hearts were prepared and then analyzed followed by a statistical comparison between computed and optical measurements. Similarly, a database of healthy and diseased patient data was used to showcase the effectiveness of the methodology.

3.17 Choice from Noise: Modelling Biology

Röbbe Wünschiers (Hochschule Mittweida, DE)

License 🅭 Creative Commons BY 3.0 Unported license
© Röbbe Wünschiers

From a methodological point of view, gene technology is shifting towards engineering. While in the past, the elucidation of gene and gene product function was dominated by trial and error, the new paradigm in genetic engineering (in the framework of synthetic biology) is based on educated predictions. The necessary data come from systems biology and its accompanying methods, as well as from new high throughput methods in nucleic acid sequencing and protein/metabolite detection. These layers of information are used in modelling and simulation. Our research picks up this paradigm in the field of fermentative biogas production and photo-biological hydrogen evolution. During the seminar I tried to draw the attention towards pitfalls in data generation. In heterogeneous biological samples the extraction of biomolecules might be biased due to extraction techniques. This has to be accounted for in downstream data processing and analyses. Furthermore, most investigations limit themselves on transcriptomic data and assume that the abundance of transcripts is proportional to the activity of the encoded proteins, respectively. Post transcriptionally processing and regulations as well as other regulatory processes, e.g. transcript or genome editing, are usually not included. It can be expected that the collaborative interaction of scientists from different fields will help to overcome this shortcomings.

3.18 Integration and relevance analysis of heterogeneous factors for stratification in biology and medicine

Dietlind Zühlke (Seven Principles AG – Köln, DE)

License 🅭 Creative Commons BY 3.0 Unported license
© Dietlind Zühlke

In the talk a research support system is presented, that helps medical researchers to identify diagnostic result patterns that characterise pertinent patient groups for personalized medicine. Example disease is breast cancer. The approach integrates established clinical findings with systems biology analyses. In this respect it is related to personalised medicine as well as translational research. Technically the system is a computer based support environment that links machine learning algorithms for classification with an interface for the medical domain expert. The involvement of the clinician has two reasons. On the one hand the intention is to impart an in-depth understanding of potentially relevant 'omics' findings from systems biology (e.g. genomics, transcriptomics, proteomics, and metabolomics) for actual patients in the context of clinical diagnoses. On the other hand the medical expert is indispensable for the process to rationally constrict the pertinent features towards a manageable selection of diagnostic findings. Without the suitable incorporation of domain expert knowledge machine based selections are often polluted by noise or irrelevant but massive variations. Selecting a subset of features is necessary in order to tackle the problem that for statistical reasons the amount of features has to be in an appropriate relationship to the number of cases that are available in a study (curse of dimensionality). The cooperative selection process is iterative. Interim results of analyses based on automatic temporary feature selections have to be graspable and criticisable by the medical expert. In order to support the understanding of machine learning results a prototype based approach is followed. The case type related documentation is in accordance with the way the human expert is cognitively structuring experienced cases. As the features for patient description are heterogeneous in their type and nature, the machine learning based feature selection has to handle different kinds of pertinent dissimilarities for the features and integrate them into a holistic representation.

4 Working groups

4.1 Wiki "Experts Integrated"

Michael Biehl (University of Groningen, NL) and Friedrich Melchert (Fraunhofer Institut – Magdeburg, DE)

License Creative Commons BY 3.0 Unported license
© Michael Biehl and Friedrich Melchert
URL https://xprt.webhosting.rug.nl/

The wiki is meant to provide a platform for discussions, exchanging ideas, suggesting concrete problems, initializing collaborations etc. It offers discussion forums, file repositories, event calendar etc.

All participants of the Dagstuhl Seminar have been provisionally registered and should have received a corresponding notification. For completion of the registration process please visit the provided URL and make use of the *"User Wizard"* to set up account and profile.

In case of questions or if the notification was not received please contact Michael Biehl (`m.biehl@rug.nl`) or Friedrich Melchert (`friedrich.melchert@gmail.com`).

4.2 Matlab code available: Learning Vector Quantization, relevance learning and matrix adaptation

Kerstin Bunte (University of Birmingham, GB) and Michael Biehl (University of Groningen, NL)

License Creative Commons BY 3.0 Unported license
© Kerstin Bunte and Michael Biehl

Learning Vector Quantization [1] (LVQ) constitutes a particularly intuitive approach to prototype-based classification. An important conceptual extension is the use of adaptive distance measures, as for instance in the framework of Generalized Matrix Relevance LVQ (GMLVQ) [2, 3, 4].

An easy-to-use Matlab (TM) code collection for the simplest variants of GMLVQ is provided at http://www.cs.rug.nl/~biehl/gmlvq. The downloadable archive contains code for the basic training process, visualization of learning curves, and validation procedures. A brief documentation and example data sets are also provided.

A more comprehensive collection is available at http://matlabserver.cs.rug.nl/gmlvqweb/web/. It comprises several variants of LVQ and GMLVQ schemes, including supervised dimension reduction and local relevance matrices.

References

1 T. Kohonen. Self-Organizing Maps. 2nd ed. Berlin, Heidelberg: Springer (1997)

2 P. Schneider, M. Biehl and B.Hammer. Adaptive Relevance Matrices in Learning Vector Quantization. Neural Computation 21(12):3532–3561 (2009)

3 P. Schneider, K. Bunte, B. Hammer and M. Biehl. Regularization in Matrix Relevance Learning. IEEE Trans. on Neural Networks 21(5):831–840 (2010)

4 K. Bunte, P. Schneider, B. Hammer, F.-M. Schleif, T. Villmann and M. Biehl. Limited Rank Matrix Learning – Discriminative Dimension Reduction and Visualization. Neural Networks 26(4):159–173 (2012)

Participants

- Gyan Bhanot
Rutgers Univ. – Piscataway, US
- Michael Biehl
University of Groningen, NL
- Kerstin Bunte
University of Birmingham, GB
- Aguirre de Cubas
Vanderbilt Univ. – Nashville, US
- Gert-Jan de Vries
Philips Research Lab. – Eindhoven, NL
- Sebastian Doniach
Stanford University, US
- Shridar Ganesan
Rutgers Cancer Inst. of New Jersey – New Brunswick, US
- Tina Geweniger
Hochschule Mittweida, DE
- Gernoth Grunst
Hennef, DE
- Barbara Hammer
Universität Bielefeld, DE
- Marika Kaden
Hochschule Mittweida, DE
- Hossein Khiabanian
Rutgers Cancer Inst. of New Jersey – New Brunswick, US
- Saurabh V. Laddha
Rutgers Cancer Inst. of New Jersey – New Brunswick, US
- John A. Lee
UC Louvain-la-Neuve, BE
- Pietro Lio'
University of Cambridge, GB
- Paulo J. Lisboa
John Moores University – Liverpool, GB
- Markus Lux
Universität Bielefeld, DE
- Elke K. Markert
The Beatson Inst. f. Cancer Research – Glasgow, GB
- John Martens
Erasmus Univ. – Rotterdam, NL
- Thomas Martinetz
Universität Lübeck, DE
- Friedrich Melchert
Fraunhofer Institut – Magdeburg, DE
- Erzsébet Merényi
Rice University – Houston, US
- Klaus Mueller
Stony Brook University, US
- David Nebel
Hochschule Mittweida, DE
- Jeffrey Rathmell
Vanderbilt University School of Medicine, US
- W. Kimryn Rathmell
Vanderbilt University School of Medicine, US
- Anupama Reddy
Duke University Medical Center – Durham, US
- Timo Ropinski
Universität Ulm, DE
- Joshua T. Taylor
Rice University – Houston, US
- Peter Tino
University of Birmingham, GB
- Alexei Vazquez
The Beatson Inst. f. Cancer Research – Glasgow, GB
- Thomas Villmann
Hochschule Mittweida, DE
- Gunther H. Weber
Lawrence Berkeley National Laboratory, US
- Ole Winther
University of Copenhagen, DK
- Thomas Wischgoll
Wright State University – Dayton, US
- Röbbe Wünschiers
Hochschule Mittweida, DE
- Dietlind Zühlke
Seven Principles AG – Köln, DE