Junior Research Group for ”Statistical Learning in Computational Biology”

Nico Pfeifer
Department of Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics
npfeifer@mpi-inf.mpg.de

Group Development

Nico Pfeifer is at the MPI for Informatics since October 2011. He started his group in January 2013 and became a senior researcher in November 2014 having the right to grant Ph.D. titles from Saarland university. Nico Pfeifer supervises five Ph.D. students directly and co-supervises one Ph.D. student together with Thomas Lengauer.

Vision and Research Strategy

Recent advances in high-throughput technologies have led to an exponential increase in biological data (such as genomic, epigenomic and proteomic data). To gain meaningful insights in such large data collections, efficient statistical learning methods are needed that take into account various sources of confounding such as batch effects or population structure, inherent to large biological data sets. We are interested in developing and applying new machine learning / statistical learning methods to solving computational biology problems and answering new biological questions. Application areas include the study of viruses like HIV, Hepatitis C or Influenza as well as the field of epigenetics. Method-wise we are interested in

- integration of heterogeneous data sets
- improving interpretability of non-linear estimators
- efficient learning methods for large data sets.

Due to about two million new HIV infections per year and about 35 million people living with an HIV infection world-wide, the HI virus is still a major threat to mankind. Two areas are of particular importance:

- research towards a vaccine against HIV
- personalized HIV treatment

We are conducting research in both of these areas. Examples include modeling the adaptation of HIV in response to external pressure by the immune system [YKK$^+$13, CBM$^+$12, CLP$^+$12], building better and more interpretable predictors for HIV coreceptor usage and CCR5 antagonist resistance prediction [PL12] as well as the analysis of potent broadly HIV-1 neutralizing antibodies ([PWL14]).
We investigated certain biases in biological data that may have very important implications for the interpretation of results based on this data (see [PWL14] and [DGG$^+$15]).
Additionally, we are also working on methods that can better deal with noisy data. One application scenario is the analysis of molecular measurements on cancer samples. Here, many effects can introduce biases (e.g., population structure, cryptic relatedness, batch effects). If one builds a prediction tool with the assumption that new data to come will be very similar to the data on which the model is trained, standard approaches are applicable. Unfortunately, this is not very often the case. Therefore, we introduced a method that is able to estimate certain differences in the underlying distribution of the training data and the test data and correct for them in the final prediction method. Furthermore, we

provided interpretable results that can be used to understand the underlying causes of the prediction label (see [JP14]).
Another important area is how to best integrate the different measurements (e.g., gene expression, DNA methylation, copy number variation). Here we extended methods for unsupervised multi kernel learning to deal with the different data types ([SP15]).
Additionally, we are interested in developing methods for the analysis of open chromatin regions as well as the three-dimensional organization of chromosomes.

Selected Research Projects

Statistical Learning for Visualizing, Analyzing and Integrating Different Omics Data Sets

Over the past decades, biology has transformed into a high throughput research field both in terms of the number of different measurement techniques as well as the amount of variables measured by each technique (e.g., from Sanger sequencing to deep sequencing) and is more and more targeted to individual cells [SBL13]. This has led to an unprecedented growth of biological information. Consequently, techniques that can help researchers find the important insights of the data are becoming more and more important. Molecular measurements from cancer patients such as gene expression and DNA methylation are usually very noisy. Furthermore, cancer types can be very heterogeneous. Therefore, one of the main assumptions for machine learning, that the underlying unknown distribution is the same for all samples in training and test data, might not be completely fulfilled.

Interpretable per Case Weighted Ensemble Method for Cancer Associations

We introduced a method that is aware of the potential bias regarding different batches of data and utilizes an estimate of the differences during the generation of the final prediction model. For this, we introduced a set of sparse classifiers based on L1-SVMs [BM98], under the constraint of disjoint features used by classifiers. Furthermore, for each feature chosen by one of the classifiers, we introduced a regression model based on Gaussian process regression that uses additional features. For a given test sample we can then use these regression models to estimate for each classifier how well its features are predictable by the corresponding Gaussian process regression model. This information is then used for a confidence-based weighting of the classifiers for the test sample. Schapire and Singer showed that incorporating confidences of classifiers can improve the performance of an ensemble method [SS99]. However, in their setting confidences of classifiers are estimated using the training data and are thus fixed for all test samples, whereas in our setting we estimate confidences of individual classifiers per given test sample.

In our evaluation, the new method achieved state-of-the-art performance on many different cancer data sets with measured DNA methylation or gene expression. Moreover, we developed a method to visualize our learned classifiers to find interesting associations with the target label. Applied to a leukemia data set we found several ribosomal proteins associated with leukemia that might be interesting targets for follow-up studies and support the hypothesis that the ribosomes are a new frontier in gene regulation. This research project was presented at WABI 2014 [JP14].

Integrating Different Data Types by Regularized Unsupervised Multiple Kernel Learning with Application to Cancer Subtype Discovery

Despite ongoing research, cancer remains a major health threat. The identification of subtypes of tumors in certain tissues can guide the decision which treatment may be beneficial for the respective patient. Nowadays established cancer subtypes are mainly based on individual types of molecular data, such as gene expression or DNA methylation. However, the analysis of multidimensional data, consisting of measurements using different platforms, may reveal intrinsic characteristics of the tumor which are based on dependencies between these different data types and can therefore only be detected when integrating the available information. Large-scale projects, such as The Cancer Genome Atlas (TCGA) [TCG] accumulate such heterogeneous data for various cancer types, but we still lack computational methods that are able to reliably integrate the given data.

To enable integrative, exploratory data analysis, we extended an approach to unsupervised multiple kernel learning for dimensionality reduction [LLF11]. In a first step, each input data type is represented by one or several kernel matrices. At this point, a major advantage is the ability of the method to automatically weight the kernel matrices, such that the user is alleviated from the burden of deciding on a kernel function or kernel parameters for each data type, instead, one can simply input a set of kernel matrices for each data type and let the method determine the optimal weighting. In an iterative optimization process, the method then trains a kernel weight vector β, used to calculate the weighted linear combination of the input kernels, and a projection matrix A which allows for dimensionality reduction. Due to the graph embedding framework [YXZ$^+$07] which forms the basis of the method, a large number of dimensionality reduction methods can be applied.

We applied this method to patient data of five different cancer types (glioblastoma multiforme, breast invasive carcinoma, kidney renal clear cell carcinoma, lung squamous cell carcinoma, and colon adenocarcinoma), where for each cancer type three different data types (gene expression, DNA methylation, and miRNA expression) were available. For dimensionality reduction we applied the locality preserving projections algorithm [HN04], which is based on the k-nearest neighborhood of a sample. We used radial basis kernel functions and, in order to investigate the efficacy of the kernel weighting, we compared two different scenarios. In the first one, we represent each input type as one kernel matrix. In Scenario 2, we use five different kernel matrices per data type, obtained by using five different kernel parameters. Our analysis revealed, that uninformative input kernel matrices indeed hardly influence the ensemble matrix. Subsequently, we applied k-means clustering to the integrated patient data to identify integrated cancer subtypes. In order to assess the biological validity of these clusters, we performed a survival analysis evaluating if the potential subtypes differ in prognosis. In Scenario 1 (one kernel per data type), we found significant differences in survival time between the subtypes for all but one cancer type. With Scenario 2, the significance for most data sets increased such that the identified subtypes are at least as significant as those identified by state-of-the-art methods, i.e., the clusters obtained reflect a better separation according to survival time of the patients than the results obtained in Scenario 1. Moreover, a leave-one-out cross validation approach showed, that the identified subtypes are relatively stable, with no decrease in stability when using more than one kernel matrix for a data type. We further looked into the groups identified for glioblastoma multiforme. For this cancer type, we were able to find subtypes that are established for distinct individual data types, but also additional subtypes, potentially based on interaction of the integrated data types. For glioblastoma multiforme, we also investigated how the subtypes respond to different treatments. For the drug Temozolomide, patients from certain subtypes seemed to benefit from that therapy, appearing a significantly increased survival time compared to patients from the same subtype but not treated with Temozolomide. In other clusters, no significant survival time differences between patient treated and not treated with this drug were observed. Overall, our method shows promising results when applied in the field of cancer subtype identification.
A manuscript describing the work was presented at ISMB/ECCB 2015 [SP15].

Projects and Cooperations

We are collaborating with several researchers internationally, nationally and also on campus: David Heckerman, Microsoft Research, Jonathan Carlson, Microsoft Research, Anne-Mieke Vandamme, KU Leuven, Rolf Kaiser, University of Cologne, Jörn Walter, University of the Saarland, Marcel Schulz, MMCI, Saarbrücken, Olga Kalinina, MPI for Informatics

References

[BM98] Paul S Bradley and Olvi L Mangasarian. Feature selection via concave minimization and support vector machines. In *ICML*, volume 98, pages 82–90, 1998.

[CBM$^+$12] Jonathan M Carlson, Chanson J Brumme, Eric Martin, Jennifer Listgarten, Mark A Brockman, Anh Q Le, Celia K S Chui, Laura A Cotton, David J H F Knapp, Sharon A Riddler, Richard Haubrich, George Nelson, Nico Pfeifer, Charles E Deziel, David Heckerman, Richard Apps, Mary Carrington, Simon Mallal, P Richard Harrigan, Mina John, and Zabrina L Brumme. Correlates of protective cellular immunity revealed by analysis of population-level immune escape pathways in HIV-1. *Journal of virology*, 86(24):13202–16, December 2012.

[CLP$^+$12] Jonathan M Carlson, Jennifer Listgarten, Nico Pfeifer, Vincent Tan, Carl Kadie, Bruce D Walker, Thumbi Ndung'u, Roger Shapiro, John Frater, Zabrina L Brumme, Philip J R Goulder, and David Heckerman. Widespread impact of HLA restriction on immune control and escape pathways of HIV-1. *Journal of virology*, 86(9):5230–43, May 2012.

[DGG$^+$15] Matthias Döring, Gilles Gasparoni, Jasmin Gries, Karl Nordström, Pavlo Lutsik, Jörn Walter, and Nico Pfeifer. Identification and Analysis of Methylation Call Differences Between Bisulfite Microarray and Bisulfite Sequencing Data with Statistical Learning Techniques. *BMC Bioinformatics (Proc. ISCB)*, 16(Suppl 3), 2015.

[HN04] Xiaofei He and Partha Niyogi. Locality Preserving Projections. In S. Thrun, L.K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 153–160. MIT Press, 2004.

[JP14] Adrin Jalali and Nico Pfeifer. Interpretable Per Case Weighted Ensemble Method for Cancer Associations. In Dan Brown and Burkhard Morgenstern, editors, *Algorithms in Bioinformatics (WABI 2014)*, volume 8701 of *Lecture Notes in Bioinformatics*, pages 352–353, Wroclaw, Poland, 2014. Springer.

[LLF11] Yen-Yu Lin, Tyng-Luh Liu, and Chiou-Shann Fuh. Multiple Kernel Learning for Dimensionality Reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(6):1147–1160, 2011.

[PL12] Nico Pfeifer and Thomas Lengauer. Improving HIV Coreceptor Usage Prediction in the Clinic Using hints from Next-generation Sequencing Data. *Bioinformatics*, 28(18):i589–i595, 2012.

[PWL14] Nico Pfeifer, Hauke Walter, and Thomas Lengauer. Association Between HIV-1 Coreceptor Usage and Resistance to Broadly Neutralizing Antibodies. *Journal of Acquired Immune Deficiency Syndromes : JAIDS*, 67(2):107–112, 2014.

[SBL13] Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet.*, 14(9):618–30, September 2013.

[SP15] Nora K Speicher and Nico Pfeifer. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics (Oxford, England)*, 31(12):i268–i275, June 2015.

[SS99] Robert E Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37(3):297–336, 1999.

[TCG] The Cancer Genome Atlas, Website, Available from: http://cancergenome.nih.gov/.

[YKK$^+$13] Yuichi Yagita, Nozomi Kuse, Kimiko Kuroki, Hiroyuki Gatanaga, Jonathan M Carlson, Takayuki Chikata, Zabrina L Brumme, Hayato Murakoshi, Tomohiro Akahoshi, Nico Pfeifer, Simon Mallal, Mina John, Toyoyuki Ose, Haruki Matsubara, Ryo Kanda, Yuko Fukunaga, Kazutaka Honda, Yuka Kawashima, Yasuo Ariumi, Shinichi Oka, Katsumi Maenaka, and Masafumi Takiguchi. Distinct HIV-1 Escape Patterns Selected by Cytotoxic T Cells with Identical Epitope Specificity. *Journal of virology*, 87(4):2253–63, February 2013.

[YXZ$^+$07] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin. Graph Embedding and Extensions: A General Framework for Dimensionality Reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):40–51, 2007.