ESANN'2006 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2006, d-side publi., ISBN 2-930307-06-4.

Neural Networks and Machine Learning in Bioinformatics - Theory and Applications

Udo Seiffert[1], Barbara Hammer[2], Samuel Kaski[3], and Thomas Villmann[4]

1- Leibniz-Institute of Plant Genetics and Crop Plant Research -
Pattern Recognition Group
Corrensstraße 3, D-06466 Gatersleben - Germany

2- Clausthal University of Technology -
Institute of Computer Science
Julius-Albert-Straße 4, D-38678 Clausthal-Zellerfeld - Germany

3- Helsinki University of Technology -
Adaptive Informatics Research Centre
P.O. Box 5400, FI-02015 TKK - Finland

4- University Leipzig - Clinic for Psychotherapy
Karl-Tauchnitz-Straße 25, D-04107 Leipzig - Germany

Abstract. Bioinformatics is a promising and innovative research field. Despite of a high number of techniques specifically dedicated to bioinformatics problems as well as many successful applications, we are in the beginning of a process to massively integrate the aspects and experiences in the different core subjects such as biology, medicine, computer science, engineering, chemistry, physics, and mathematics. Within this rather wide area we focus on neural networks and machine learning related approaches in bioinformatics with particular emphasis on integrative research against the background of the above mentioned scope.

1 Introduction

Completed three years ago, the human genome project (HGP) demonstrates the high standards of technology, algorithms, and tools in bioinformatics for dedicated purposes such as reliable and parallel genome sequencing, fast sequence comparison and search in databases, automated gene identification, efficient modelling and storage of heterogeneous data, etc. Thereby, machine learning has played an indispensable role right from the beginning: gene identification and related tasks are based on various adaptive machine learning tools such as feed-forward networks or decision trees, one promising way to compute multiple alignments is offered by hidden Markov models, dedicated support vector machines constitute one of the most accurate approaches for detecting remote homologies, to name just a few examples. The HGP, however, gave several surprises such as the unexpected sparsity of coding regions of human DNA, pointing out the importance of alternative splicing. These findings have led to new research problems which accompany the not yet satisfactorily solved classical (and mostly NP-hard) problems such as protein structure prediction, multiple alignment, or phylogenetic inference. For all these problems, machine learning offers one promising approach to achieve efficient and reliable heuristic solutions.

Since most proteins arise from post-translational processes, biological networks such as gene interaction networks or metabolic networks play an essential role in the understanding of cell processes. Techniques to infer biological networks from biological data, e.g. gene expression data, as well as electronic databases for biological networks become more and more available, such that the integration of high level biological information into bioinformatics research becomes possible. The continuous development of high quality biotechnology, e.g. micro-array techniques and mass spectrometry, which provide complex patterns for the direct characterization of cell processes, offers further promising opportunities for advanced research in bioinformatics. This way, challenging problems from clinical proteomics, drug design, or design of species can be tackled. However, in all these problems, a variety of different biological information as detailed above plays a central role, and bioinformatics must cross the border towards a massive integration of the aspects and experience in the different core subjects and towards an integrated understanding of relevant processes in systems biology. This puts new challenges not only on appropriate data storage, visualization, and retrieval of heterogeneous information, but also on machine learning tools used in this context, which must adequately process and integrate heterogeneous information into a global picture.

2 Clustering

Clustering a given data set can have different purposes: preprocessing of data to simplify further analysis, identification of typical prototypes, arrangement of data along a dendrogram, identification of closely connected regions of the data, visualization, or data mining. In bioinformatics, several categories of clustering methods are commonly used: *iterative agglomerative hierarchical clustering* provides a dendrogram of a given set of data points based on pairwise distances of the data points. The concrete implementations differ in the way how distances of clusters are computed. Popular methods include unweighted pairs grouping using arithmetic means (UPGMA) or neighbor joining [1, 2, 3]. These methods are used e.g. for phylogenetic inference or, more general, the arrangement of taxa or conditions described by sequence information or more complex patterns such as micro-array data into clusters. One problem of these methods is their usually large sensitivity to noise and the choice of the metric. *Prototype based* methods such as k-means, fuzzy-clustering, neural gas, or the self-organizing map constitute an alternative in bioinformatics which yield a flat fuzzy or crisp decomposition of the given data set into clusters [4, 5]. The prototypes represent the usually a-priori fixed number of clusters by representatives, and the cluster assignment takes place based on the similarity to the cluster prototype. This offers very intuitive and robust results, however, depending on the method, the number of clusters or the topology has to be fixed a priori. A variety of *statistical formulations* provides another interface to very powerful, put often also computationally demanding clustering algorithms. Statistical models have the benefit that hierarchies as well as flat parts can easily be integrated, and the model

assumptions are clearly stated in mathematical terms (although it is usually not guaranteed that the model assumptions meet reality). Popular models applied in bioinformatics include, for example, mixture models, stochastic processes, or latent space models [6, 7, 8, 9]. Model adaptation usually takes place by an optimization of the likelihood or some other analogous objectives. Closely related to statistical models are models stemming from *statistical physics* [10] or *information theory* [11] which also optimize cost terms for clustering. Finally, *graph clustering* plays a role in bioinformatics especially for the analysis or comparison of biological networks, e.g. to determine and compare functionally similar proteins in a given network [12, 13, 14].

Applications of clustering methods in bioinformatics range from clustering of DNA sequences and genes [2, 15], gene expression analysis on the base of microarray data [1, 10, 16, 5, 7, 8, 17, 18, 9], inference of gene interaction networks based on these clusters [19, 20], visualization and mining of proteomics data [21], up to biological networks analysis and the identification of functional groups in protein association networks [12]. Thus, these techniques have an impact on phylogenetics, genomics, proteomics, clinical research, up to the understanding of cell processes and systems biology. The methods differ in their sensitivity to noise, computational complexity, and possibility of online adaptation for new data. Further, clusters might be fuzzy or crisp, the output might be hierarchic or flat, and the input data of the methods ranges from Euclidean vectors, proximity data, up to graph structures.

Often, clustering methods heavily depend on the choice of a *metric* for the given data structures, and the design of appropriate and efficient similarity measures which, in particular, incorporate higher biological information constitutes a topic of ongoing research. DNA sequences or proteins are often compared by some form of pairwise or multiple alignment, whereby several methods which incorporate appropriate statistical information and which make the NP hard multiple alignment problem feasible have been proposed [22]. For protein sequences, not only the primary structure but also the secondary or tertiary structure can be included [23]. Recently, alignment methods have also been proposed for data from mass spectrometry [24, 25] and metabolic pathways [26]. Micro-array data can be processed by means of correlation measures which take the overall shape of gene regulation patterns into account [2]. Thereby, also co-regulation of more than one gene and mutual dependencies of clusters are of particular interest to achieve meaningful results [27, 6, 28]. For more complex structures such as chemical molecules or graphs, a variety of graph kernels which can serve as similarity measure has been developed [29]. Finally, we would like to mention, that kernels can also be defined taking already given cluster information into account, as demonstrated e.g. in [30]. Based on an appropriate similarity measure, similarity based clustering is easily possible. In addition, similarity measures constitute the crucial part for database retrieval and similarity inference, which is quite popular, e.g. to detect remote homologies with similar shape and function for dedicated drug design [31].

Bioinformatics puts a number of challenges towards traditional clustering

methods. Data are usually *very high dimensional*, e.g. long DNA strings, high dimensional spectra, or micro-array data. Thus, the curse of dimensionality must be avoided, e.g. by preprocessing with (nonlinear) principal component analysis, latent semantic indexing, selection of features, or Fourier transform [1, 21, 4, 32]. Gene expression data and clinical spectra can incorporate a *time-dependency* which should be taken into account [1, 16, 9]. For an integrated analysis, *different data types and types of information*, e.g. given by different biological networks must be fused together to a single reliable and meaningful image [33, 34, 15]. Further, additional information such as gene functions from gene interaction networks or other *constraints* should be taken into account for clustering [18, 7, 35]. Unsupervised clustering is generally prone to the 'garbage-in-garbage-out' dilemma: a naive clustering algorithm applied to inappropriate data representations will likely give some random result with only little information. Thus, as much *additional information* as possible, in particular higher biological information, should be integrated to shape the clustering algorithm towards meaningful results. One problem consists in the fact, that different types of data and different clustering algorithms can yield to rather dissimilar results. Because of this fact, methods to *judge the validity* of the output are indispensable. Proposals to achieve this goal range from the automated integration of cluster validity measures [36, 37, 17], bootstrap and consensus methods for several runs [38, 2], up to a direct visual inspection of different outcomes, e.g. multiple dendrograms or results from hierarchical and flat clustering [39, 40].

3 Classification

Unlike unsupervised learning, the objective of supervised classification models is error minimization. Thus, a natural cost function, the number of misclassifications, exists. Nevertheless, several metric-based classification models do not explicitly optimize this cost function, but they are based on intuitive heuristics. Generally, machine learning tools become standard alternative approaches also for classification in bioinformatics [41, 42, 43].

There exists a broad variety of models for classification ranging from traditional statistical approaches to artificial neural networks. In statistics linear and quadratic discriminant analysis or regression models are standard tools with various problem specific extensions and modifications. Naturally, these approaches play an important role in bioinformatics and are widely used [42]. Yet, the assumptions on the data in terms of their applicability to a particular method are often not fulfilled. In particular, in standard statistics usually normal distributions within data are assumed. Other statistical classification models base on probabilistic approaches [44]. Thereby, Bayesian inference models as complement to significance statistics claim an increasing impact [45] and were successfully applied in splice site recognition, for example [46]. Yet, the underlying assumptions about normal data distribution are frequently a substantial restriction for application. This must be seen in the context of the frequently occurring problems concerning the data in bioinformatics, which are sparseness, noise, un-

balanced data (ethic problems to get data from healthy volunteers in medical investigations), fuzziness, and missing values, which often make additional techniques in machine learning necessary to account for these facts [32, 47]. Further, the structure of data may be high-dimensional (curse of dimensionality) and highly structured as in case of spectral (functional) data.

Traditionally, trees and tree classification schemes are of great interest in medicine and biology. Thus, decision trees are the natural choice for many bioinformatics classification problems like in taxonomy or phylogenetic dependency representations. Further, they can be used in medical decision systems. *Induction of decision trees* constitutes a standard symbolic machine learning tool for classification of data [48]. A decision tree consists of a tree which interior nodes are labelled by a dimension number and the connections to the node's children are labelled by real values which split the dimension into intervals. The leaves contain class labels. Given a datum, a decision is based on consecutive decisions provided by the interior nodes until the class information is reached at the leaves. For a given training set, a decision tree can recursively be induced by the choice of an interior node and an induced split of the training set until a widely uniform classification is possible at the leaves. Thereby, an appropriate measure such as the entropy guides the choice of the splitting dimension [49]. However, decision tree learners consider only one attribute at a time, such that relevance distributed among several attributes cannot be detected. They provide explicit rules and an ordering of the dimensions with respect to their importance for the decision tree by means of their depth within the tree. Yet, decision trees are sensitive with respect to disturbed data, which lead to instable solutions, i.e. different resulting tree structures. One possible solution is to combine tree generating systems with robust classification schemes like neural networks. One approach based on prototype based classifiers are BB-trees which can be used for decision system generation [50].

Artificial neural networks offer new possibilities for machine learning approaches in biomedical applications. The robust behaviour according to noisy data, the high adaptability provides several of the above mentioned features which are required in bioinformatics.

Again, applications exist for many areas in bioinformatics such as micro-array analysis [51, 52], analysis of mass spectra and biomarker fishing in proteomics [53, 41, 54]. An overview can be found in [55]. Besides the classic multiple-layer perceptron (MLP), support vector machines (SVM) provide a powerful utility for classification, which are able to handle complex data structures, nonlinearities, and high-dimensionality [56, 51]. A principle alternative are prototype based classifiers, the root of which is the family of learning vector quantization (LVQ), introduced in [57]. In dependence on the class distribution representatives are generated which act as characteristic prototypes for the several classes. Prototype based methods have the advantage that the classification scheme is easy to verify. Thus, the intuitive understanding of the decision scheme gives hints for the classification process in contradiction to the black box decision of an MLP-network, and prototypes often allow visualization of the classification

behavior [58]. Several extensions exist, e.g. neighborhood cooperativeness for stability and improved convergence [59], fuzzy classification schemes [54, 60]. In this context, also strategies for optimal data selection for learning (active learning) can be applied [61].

Further, for all classification methods the underlying metric plays a crucial role: the metric can be chosen in agreement with the classification task or may be contradictious in the worst case. Therefore, adaptive, non-standard metrics are required for optimum classification [62]. Whereas MLPs inherently weight the data streams during learning, prototype based classifiers and SVMs can be extended to deal with metric adaptation and non-standard metrics as demonstrated in splice site recognition, mass spectroscopy or gene expression analysis [63, 56, 64].

4 Visualization and mining

One of the most difficult and central topics in bioinformatics is how to best infer systemic properties of cells and organisms from data, model them, and take them into account in data analysis. It has been widely appreciated that most of the earlier biological research has focused on studying only parts of the systems, and understanding how the parts interact will be the next big challenge. The field studying the integration of the parts, and more generally systemic properties of cells and ultimately organisms, has been coined *systems biology*. Although the name and its scope have received criticism, it is clear that modelling of systemic properties is a key to understanding functioning of cellular systems.

On the cellular level, biological systems have so far been conceptualized in terms of several interacting systems: gene regulatory networks, metabolic pathways, signaling networks, and more generally interaction networks of proteins. Two recent advances in studying such networks are particularly important for machine learning: (i) New so-called high-throughput measurement techniques have been and are being developed to measure different aspects of the functioning of cells. Gene expression micro-arrays that measure genome-wide gene activity are perhaps the best-known examples, but for instance gene regulation and protein-protein interactions can be measured on a massive scale as well. (ii) The measurement data and insights derived from them are being collected into databases, of which many are publicly available. Item number one above poses interesting new challenges to modelling methods, and item number two makes it practically feasible to apply machine learning [65, 66].

A lot is already known of the various interaction networks, but very little of the knowledge is available in a quantitative form, ready to be incorporated into statistical models. So far, the modelling problems are typically underconstrained and even ill-defined, and usually modelling is interleaved with the use of data-driven methods to "look at the data." Particularly the noisy high-throughput measurement data need careful exploratory analysis before they can be incorporated into quantitative models. Hence, the field needs machine learning both as flexible statistical models and for visualization and data mining.

The earliest studies of high-throughput gene expression data used clustering to study hints of gene regulation [67]. Later, more sophisticated methods for deriving regulatory interactions from data with, for instance, Bayesian networks were developed [68]. It has recently turned out that the problem of deriving interaction graphs from data is very hard, which of course makes it even more intriguing to modelers. Moreover, the fact that inference of interaction networks from data is not trivial, leaves room for other kinds of innovative machine learning methods. Again, inference from data and existing data bases needs to be interleaved with data-driven exploratory methods. Visualization is particularly useful in an interactive modelling process. One of the papers [69] in this session introduces a new machine learning method for the general task of visualizing interaction graphs, applicable and timely in systems-biological analyses of cellular interaction networks.

5 High-performance computing

Many biomedical problems (e.g., micro-array gene expression data analysis, image based pattern recognition, genetic and biochemical network analysis, protein-protein interactions, phylogenetic reconstruction, genetic linkage analysis, protein structure prediction, etc.) require either *computationally expensive numerical operations*, or operations on *large-scale data sets*, or, the presence of both characteristics leads to even more computational challenges. It is often unsuitable or even impossible to solve these problems on conventional single processor computers due to their enormous amount of computation time.

Despite of constantly rising clock rates and hardware design improvements, the speed of single processor machines is limited by some fundamental physical constraints of about three to five billion floating point operations per second. Parallel computing overcomes this asymptotic behaviour by joining several processors to compute the different parts of a complex problem independently and quasi simultaneously. Following Amdahl's law [70] the speed-up that can be gained by parallelism generally depends on the granularity of – or better said the potential to cleverly decompose – the problem at hand (see [71] for an illustration). Many bioinformatics applications possess exactly this feature. Examples can be given for parallel clustering [72] and Bayesian inference [73], parallel sequence and string processing [74, 75], parallel image processing [76], and finally even distributed data bases [77].

While these considerations define rather the requirements, the equipment in modern bioinformatics labs assures the technical basis for high-performance computing (HPC) with compute farms, computer clusters, and shared-memory systems. These different types of computer hardware provide a more or less good and suitable ground for bioinformatics applications of different fields along with their different hardware demands.

Artificial neural networks offer a very high potential in terms of parallel processing as well, due to their inherent parallelism [78, 79, 80]. Their demands regarding the underlying computer hardware is variable, but tends to

prefer closely and fast connected architectures. This comes from the relatively tight connections between artificial neurons (nodes) within a network. Thus, a clear preference is given to shared-memory computers or at least to very fast (e.g. Myrinet) interconnected Beowulf clusters. This way, ANN based high-performance computing paves the way for a number of biomedical applications and investigations, which were not conceivable otherwise. Meanwhile a rather large number of very successful applications have evolved in this wide field. For a representative survey refer to [55].

Also some more or less recent trends, such as DNA computing [81], evolvable hardware [82, 83], or organic computing [84], made their way into this field. Especially organic computing seems to be particularly interesting for ANN based bioinformatics. Since many labs are equipped with computer clusters and compute farms, and applications are usually not running exclusively on the available computers, hardware adaptive implemented algorithms are particularly desired [85], most suitably those systems being able to adapt themselves at run-time [86, 87].

Acknowledgement

This work was supported by a grant of the German Federal Ministry of Education and Research (No. 0312706A), and by the Academy of Finland (No. 207467). In addition, the first author wishes to thank Andrea Matros for the valuable support.

References

[1] R. Amato, A. Ciaramella, N. Deniskina, C. Del Mondo, D. di Bernardo, C. Donalek, G. Longo, G. Mangano, G. Miele, G. Raiconi, A. Staiano, and R. Tagliaferri. A multi-step approach to time series analysis and gene expression clustering. *Bioinformatics*, Advance Access, 2006.

[2] R. Jothi, E. Zotenko, A. Tasneem, and T.M. Przytycka. COCO-CL: hierarchical clustering of homology relations based on evolutionary correlations. *Bioinformatics*, Advance Access, 2006.

[3] D. Ashok Reddy, B.V.L.S. Prasad, and C.K. Mitra. Functional classification of transcription factor binding sites: information content as a metric. *Journal of Integrative Bioinformatics*, 2006-02-06, 2006.

[4] D. Chen, H. Bensmail, and Y. Xu. Clustering gene expression data with kernel principal components. *Journal of Bioinformatics and Computational Biology*, 3(2):303–316, 2005.

[5] F. Luo, L. Khan, F. Bastani, I. Yen, and J. Zhou. A dynamically growing self-organizing tree (dgsot) for hierarchical clustering gene expression profiles. *Bioinformatics*, 20:2605–2617, 2004.

[6] S. Kaski, J. Nikkila, J. Sinkkonen, L. Lahti, J.E.A. Knuuttila, and C. Roos. Associative clustering for exploring dependencies between functional genomics data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(2):203–216, 2005.

[7] W. Pan. Incorporating gene functions as priors in model-based clustering of microarray gene expression data. *Bioinformatics*, Advance Access, 2006.

[8] S. Rogers, M. Girolami, C. Campbell, and R. Breitling. The latent process decomposition of cDNA microarray data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(2):143–156, 2005.

[9] F.-X. Wu, W.J. Zhang, and A.J. Kusalik. Dynamic model-based clustering for time-course gene expression data. *Journal of Bioinformatics and Computational Biology*, 3(4):821–836, 2005.

[10] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3/4):281–297, 1999.

[11] Z. Dawy, B. Goebel, J. Hagenauer, C. Andreoli, T. Meitinger, and J.C. Mueller. Gene mapping and marker clustering using shannon's mutual information. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(3):47–56, 2006.

[12] B. Adamcsek, G. Palla, I. Derényi I.J. Farkas, and T. Vicsek. Cfinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, Advance Access, 2006.

[13] J.K. Choi, U. Yu, O.J. Yoo, and S. Kim. Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics*, 22:4348–4355, 2006.

[14] A.D. King, N. Przulj, and I. Jurisica. Protein complex prediction via cost-based clustering. *Bioinformatics*, 20:3013–3020, 2004.

[15] J. Kasturi and R. Acharya. Clustering of diverse genomic data using information fusion. *Bioinformatics*, 21:423–429, 2005.

[16] K. Hakamada, M. Okamoto, and T. Hanai. Novel technique for preprocessing high dimensional time-course data from DNA microarray: mathematical model-based clustering. *Bioinformatics*, Advance Access, 2006.

[17] V.S. Tseng and C. Kao. Efficiently mining gene expression data via a novel parameterless clustering method. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(2):355–365, 2005.

[18] H.L. Turner, T.C. Bailey, W.J. Krzanowski, and C.A. Hemingway. Biclustering models for structured microarray data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(2):316–329, 2005.

[19] Z. Jiang and Y. Zhou. Using gene networks to drug target identification. *Journal of Integrative Bioinformatics*, 2005-12-07, 2006.

[20] M. Quach, P. Geurts, and F. d'Alché Buc. Elucidating the structure of genetic regulatory networks: a study of a second order dynamical model on artificial data. In this volume.

[21] H. Bensmail, J. Golek, M.M. Moody, J.O. Semmes, and A. Haoudi. A novel approach for clustering proteomics data using bayesian fast fourier transform. *Bioinformatics*, 21:2210–2224, 2005.

[22] S.S. Adi and C.E. Ferreira. Gene prediction by multiple syntenic alignment. *Journal of Integrative Bioinformatics*, 2005-11-18, 2006.

[23] J. Ebert and D. Brutlag. Development and validation of a consistency based multiple structure alignment algorithm. *Bioinformatics*, Advance Access, 2006.

[24] N. Jeffries. Algorithms for alignment of mass spectrometry proteomic data. *Bioinformatics*, 21:3066–3073, 2005.

[25] M. Strickert, T. Czauderna, S. Peterek, A. Matros, H.-P. Mock, and U. Seiffert. Full-length HPLC signal clustering and biomarker identificationin tomato plants. In *Proceedings of the 7th International FLINS Conference on Applied Artificial Intelligence*, 2006. To appear.

[26] R.Y. Pinter, O. Rokhlenko, E.Yeger-Lotem, and M. Ziv-Ukelson. Alignment of metabolic pathways. *Bioinformatics*, 21:3401–3408, 2005.

[27] L. Ji and K. Tan. Mining gene expression data for positive and negative co-regulated gene clusters. *Bioinformatics*, 20:2711–2718, 2004.

[28] M. Strickert, U. Seiffert, N. Sreenivasulu, W. Weschke, Th. Villmann, and B. Hammer. Generalized Relevance LVQ (GRLVQ) with correlation measures for gene expression data. *Neurocomputing*, 69:651–659, 2006.

[29] L. Ralaivola, S.J. Swamidass, H. Saigo, and P. Baldi. Graph kernels for chemical informatics. *Neural Networks*, 8:1093–1110, 2005.

[30] J. Weston, C. Leslie, E. Ie, D. Zhou, A. Elisseeff, and W.S. Noble. Semi-supervised protein classification using cluster kernels. *Bioinformatics*, 21:3241–3244, 2005.

[31] K. Rother, M. Dunkel, E. Michalsky, S. Trissl, A. Goede, U. Leser, and R. Preissner. A structural keystone for drug design. *Journal of Integrative Bioinformatics*, 2006-01-19, 2006.

[32] O. Okun, N. Zagoruiko, A. Alves, O. Kutnenko, and I. Borisova. Selection of more than one gene at a time for cancer prediction from gene expression data. In this volume.

[33] J.A. Berger, S. Hautaniemi, S.K. Mitra, and J. Astola. Jointly analyzing gene expression and copy number data in breast cancer using data reduction models. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(3):2–16, 2006.

[34] F. Boyer, A. Morgat, L. Labarre, J. Pothier, and A. Viari. Syntons, metabolons and interactons: an exact graph-theoretical approach for exploring neighbourhood between genomic and functional data. *Bioinformatics*, 21:4209–4215, 2005.

[35] D. Zhu, A.O. Hero, H. Cheng, R. Khanna, and A. Swaroop. Network constrained clustering for gene microarray data. *Bioinformatics*, 21:4014–4020, 2005.

[36] N. Bolshakova, F. Azuaje, and P. Cunningham. An integrated tool for microarray data clustering and cluster validity assessment. *Bioinformatics*, 21:451–455, 2005.

[37] L. Ji and K. Tan. Identifying time-lagged gene clusters using gene expression data. *Bioinformatics*, 21:509–516, 2005.

[38] T. Grotkjaer, O. Winther, B. Regenberg, J. Nielsen, and L.K. Hansen. Robust multi-scale clustering of large DNA microarray datasets with the consensus algorithm. *Bioinformatics*, 22:58–67, 2006.

[39] T.M.W. Nye, P. Lió, and W.R. Gilks. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. *Bioinformatics*, 22:117–119, 2006.

[40] A. Torrente, M. Kapushesky, and A. Brazma. A new algorithm for comparing and visualizing relationships between hierarchical and flat gene expression data clusterings. *Bioinformatics*, 21:3993–3999, 2005.

[41] M. Hilario, A. Kalousis, M. Müller, and C. Pelligrini. Machine learning approaches to lung cancer prediction from mass spectra. *Proteomics*, 3:1716–1719, 2003.

[42] J.C. Lindon, E. Holmes, and J.K. Nicholson. Pattern recognition methods and application in biomedical magnetic resonance. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 39:1–40, 2001.

[43] J. Li, Z. Zhang, J. Rosenzweig, Y.Y. Wang, and D.W. Chan. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clinical Chemistry*, 48(8):1296–1304, 2002.

[44] R.H. Lilien, H. Farid, and B.R. Donald. Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. *Journal of Computational Biology*, 10(6):925–946, 2003.

[45] P. Congdon. *Bayesian Statistical Modelling*. Wiley, 2001.

[46] D. Cai, A. Delcher, B. Kao, and S. Kasif. Modeling splice sites with bayes networks. *Bioinformatics*, 16(2):152–158, 2000.

[47] Y. Sun, M. Robinson, R. Adams, R. te Boeckhorst, A.G. Rust, and N. Davey. Using sampling methods to improve binding site predictions. In this volume.

[48] J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[49] I.H. Witten and E. Frank. *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, 2000.

[50] B. Hammer, A. Rechtien, M. Strickert, and Th. Villmann. Relevance learning for mental disease classification. In M. Verleysen, editor, *Proc. Of European Symposium on Artificial Neural Networks (ESANN'2005)*, pages 139–144, Brussels, Belgium, 2005. d-side publications.

[51] M.P.S. Brown, W.N. Grundy, D. Lin, N. Christianini, C.W. Sugnet, T.S. Furey, M. Ares Jr., and D. Haussler. Knowledge-based analysis of microarray gene expression data using support vector machines. *PNAS*, 97(1):262–267, 2000.

[52] R.L. Somorjai, B. Dolenko, and R. Baumgartner. Class prediction and discovery using gene microarray and proteomics mass spectrosopy data: curses, caveats, cautions. *Bioinformatics*, 19(12):1484–1491, 2003.

[53] G. Ball, S. Mian, F. Holding, R.O. Allibone, J. Lowe, S. Ali, G. Li, S. McCardle, I.O. Ellis, C. Creaser, and R.C. Rees. An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. *Bioinformatics*, 18(3):395–404, 2002.

[54] Th. Villmann, F.-M. Schleif, and B. Hammer. Local metric adaptation for soft nearest prototype classification to classify proteomic data. *Neurocomputing*, page to appear, 2006.

[55] U. Seiffert, L.C. Jain, and P. Schweizer, editors. *Bioinformatics using Computational Intelligence Paradigms*, volume 176 of *Studies in Fuzziness and Soft Computing*. Springer-Verlag, Heidelberg, 2005.

[56] F.-M. Schleif, U. Clauss, Th. Villmann, and B. Hammer. Supervised relevance neural gas and unified maximum separability analysis for classification of mass spectrometric data. In *Proceedings of the International Conference of Machine Learning Applications (ICMLA'2004)*, pages 374–379. IEEE Press, 2004.

[57] T. Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995. (Second Extended Edition 1997).

[58] M. Biehl, P. Pasma, M. Pijl, L. Sanchez, and N. Petkov. Classification of boar spermatozoid head images using learning vector quantization. In this volume.

[59] B. Hammer, M. Strickert, and Th. Villmann. Supervised neural gas with general similarity measure. *Neural Processing Letters*, 21(1):21–44, 2005.

[60] C. Brüß, F. Bollenbeck, F.-M. Schleif, W. Weschke, Th. Villmann, and U. Seiffert. Fuzzy image segmentation with fuzzy labeled neural gas. In this volume.

[61] F.-M. Schleif, B. Hammer, and Th. Villmann. Margin based active learning for LVQ networks. In this volume.

[62] B. Hammer and Th. Villmann. Classification using non-standard metrics. In M. Verleysen, editor, *Proc. off the 13. European Symposium on Artificial Neural Networks (ESANN'2005)*, pages 303–316, Brussels, Belgium, 2005. d-side publications.

[63] M. Strickert, N. Sreenivasulu, W. Weschke, U. Seiffert, and Th. Villmann. Generalized relevance LVQ with correlation measure for biological data. In M. Verleysen, editor, *Proc. of the 13. European Symposium on Artificial Neural Networks (ESANN'2005)*, pages 331–338, Brussels, Belgium, 2005. d-side publications.

[64] B. Hammer, M. Strickert, and Th. Villmann. Prototype based recognition of splice sites. In U. Seiffert, L.C. Jain, and P. Schweizer, editors, *Bioinformatics using Computational Intelligence Paradigms*, pages 25–56. Springer-Verlag, 2005.

[65] U. Seiffert. Biologically inspired image compression in biomedical High-Throughput Screening. In A.J. Ijspeert, M. Murata, and N. Wakamiya, editors, *Biologically Inspired Approaches to Advanced Information Technology*, volume 3141 of *Lecture Notes in Computer Science*, pages 428–440. Springer-Verlag, Heidelberg, Oct 2004.

[66] M. Strickert, S. Teichmann, N. Sreenivasulu, and U. Seiffert. High-throughput multi-dimensional scaling (Hit-MDS) for cDNA-array expression data. In W. Duch, J. Kacprzyk, E. Oja, and S. Zadrozny, editors, *Artificial Neural Networks: Biological Inspirations (ICANN 2005)*, volume 3696 of *Lecture Notes in Computer Science*, pages 625–634. Springer-Verlag, Heidelberg, 2005.

[67] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95:14863–14868, 1998.

[68] N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303:799–805, 2004.

[69] J. Venna and S. Kaski. Visualizing gene interaction graphs with local multidimensional scaling. In this volume.

[70] G. Amdahl. Validity of the single processor approach to achieving large-scale computing capabilities. In *Proceedings of the American Federation of Information Processing Societies (AFIPS) Conference*, volume 30, pages 483–485, 1967.

[71] Y.-S. Hwang and J.H. Saltz. Identifying parallelism in programs with cyclic graphs. *Journal of Parallel and Distributed Computing*, 63(3):337–355, 2003.

[72] S. Rajasekaran. Efficient parallel hierarchical clustering algorithms. *IEEE Transactions on Parallel and Distributed Systems*, 16(6):497–502, 2005.

[73] X. Feng, D.A. Buell, J.R. Rose, and P.J. Waddell. Parallel algorithms for Bayesian phylogenetic inference. *Journal of Parallel and Distributed Computing*, 63(7-8):707–718, 2003.

[74] F. Gebali and A.N.M. Ehtesham Rafiq. Processor array architectures for deep packet classification. *IEEE Transactions on Parallel and Distributed Systems*, 17(3):241–252, 2006.

[75] S. Rajko and S Aluru. Space and time optimal parallel sequence alignments. *IEEE Transactions on Parallel and Distributed Systems*, 15(12):1070–1081, 2004.

[76] J. Fernandez, J. Carazo, and I. Garcia. Three-dimensional reconstruction of cellular structures by electron microscope tomography and parallel computing. *Journal of Parallel and Distributed Computing*, 64(2):285–300, 2004.

[77] S. Menon. Allocating fragments in distributed databases. *IEEE Transactions on Parallel and Distributed Systems*, 16(7):577–585, 2005.

[78] U. Seiffert. Artificial neural networks on massively parallel computer hardware. *Neurocomputing*, 57:135–150, March 2004.

[79] T. Czauderna and U. Seiffert. Implementation of MLP networks running Backpropagation on various parallel computer hardware using MPI. In *Proceedings of the 5th International Conference on Recent Advances in Soft Computing (RASC)*, pages 116–121, 2004.

[80] U. Seiffert and B. Michaelis. Multi-dimensional Self-Organizing Maps on massively parallel hardware. In Nigel Allinson, Hujun Yin, Lesley Allinson, and Jon Slack, editors, *Advances in Self-Organizing Maps: Proceedings of the 3. Workshop on Self-Organizing Maps WSOM 2001*, pages 160–166, London, U.K., 2001. Springer-Verlag.

[81] M. Amos. *Theoretical and Experimental DNA Computation*. Natural Computing. Springer, Berlin, 2005.

[82] M. Murakawa, S. Yoshizawa, I. Kajitani, and T. Higuchi. Evolvable hardware for generalized neural networks. In Tetsuya Higuchi, editor, *Evolvable Systems, Proceedings of the 15th International Joint Conference on Artificial Intelligence, IJCAI-97*, pages 1146–1155, San Francisco, 1997. Morgan Kaufmann.

[83] Y. Liu, K. Tanaka, M. Iwata, T. Higuchi, and M. Yasunaga, editors. *Evolvable Systems: From Biology to Hardware, 4th International Conference on Evolvable Systems, ICES 2001 Tokyo, Japan*, volume 2210 of *Lecture Notes in Computer Science*. Springer-Verlag, Tokyo, 2001.

[84] C. Müller-Schloer, T. Ungerer, and B. Bauer, editors. *Organic and Pervasive Computing*, volume 2981 of *Lecture Notes in Computer Science*. Springer, Berlin, 2004.

[85] J. Dongarra and V. Eijkhout. Self-adapting numerical software for next generation applications. *International Journal of High Performance Computing and Applications*, 17(2):125–131, 2003.

[86] U. Seiffert. Adaptive implementation of artificial neural networks reflecting changing hardware resources at run-time. In M.H. Hamza, editor, *Proceedings of the 23rd International Conference on Artificial Intelligence and Applications (AIA)*, pages 733–737, Anaheim, 2005. IASTED, ACTA Press.

[87] Th. Villmann, B. Hammer, and U. Seiffert. Perspectives of self-adapted self-organizing clustering in organic computing. In A.J. Ijspeert, T. Masuzawa, and S. Kusumoto, editors, *Biologically Inspired Approaches to Advanced Information Technology*, pages 141–159. Springer-Verlag, Heidelberg, 2006.