Artificial Intelligence in Medicine (2005) **35**, 1—8

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

GUEST EDITORIAL

Computational intelligence in solving bioinformatics problems

1. Computational intelligence and its challenges

Computational intelligence is a very broad term meaning different things to different people. On the one hand, it can be seen as one encompassing methodologies like artificial neural networks, evolutionary computation, fuzzy systems, ant colony optimization, swarm intelligence, etc., and their hybrids. As the term intelligence implies, it can mean designing computational methods that mimic information processing, organization, and behavior of living organisms. On the other hand, computational intelligence methods, whether more or less inspired by our understanding of how living organisms function, are becoming indispensable tools for modeling and analysis of often vast amounts of biological data. Hybrid approaches are often required to deal with biological data since computational ''intelligence'' cannot be accomplished by any single methodology alone, analogously to human intelligence that can also be seen as a hybrid of several types of ''intelligence'' (to learn, to recognize, to adapt, or to understand a discovered reality).

Important issues in computational intelligence, in spite of great progress made in many of the areas mentioned below, remain as challenging today as they were a decade ago. In no particular order of importance, and far from being exhaustive, but guided by the topics addressed in this special issue, we briefly discuss them below. It seems that the most important issues are those associated with developing more efficient problem-solving techniques. It is the never-ending quest for better search algorithms, and for better constraint satisfaction methods. To illustrate what problems bioinformatics faces, let us suppose that the goal was to simulate activity of an entire single protein at the level of individual atoms. Assuming one had a fast computer operating at the speed of 10^{15} operations/s (an order of magnitude faster than most current computers), it would still take months of processing [1]. We think that more progress will be achieved via design of smarter algorithms than by the progress in hardware.

Another research area is to develop optimal or semi-optimal ways of knowledge representation for a specific problem (or groups of problems) and ways of dealing with uncertain knowledge. An example is a challenging problem of representing time-series stochastic processes. Associated with it are strategies for making valid inferences from stored knowledge, which include logical, probabilistic, and fuzzy inference techniques.

Still another problem is associated with the types of data routinely collected. The easiest by far to collect by biologists and other life scientists are data that are generated by some experiment or a process but without knowing the corresponding output (label/category/diagnosis). In other words, it is very costly to come up with training (also called labeled) data, for which the relationship between input and output, for each data point, is known. An example of training data is an input image of a skin sample and the corresponding diagnosis of, say, skin cancer type. Unfortunately, most of the huge amounts of data collected are of the first type. Making sense of such data is very difficult, and the most important tool at our disposition is clustering, an unsupervised learning technique. However, there are two big problems associated with clustering, namely how to choose an appropriate similarity measure and how to guess a priori the number of clusters one expects to be present in the data. Although there exist few clustering algorithms, like Kohonen's self-organizing feature maps [2] and AutoClass [3], that do not require user to specify the number, they are

0933-3657/$ — see front matter
doi:10.1016/j.artmed.2005.07.001

far from perfect (interpretation of Kohonen's network is very difficult) and thus clustering remains one of the most challenging research areas of computational intelligence. A lot of progress needs to be made in clustering to deal with huge amounts of unsupervised data. Once it becomes possible to find reliable clusters, then experts can possibly label them, i.e. associate with each group a correct label/diagnosis/category [4–6]. Doing that, in turn, would allow for using many of the existing supervised learning techniques [7], which require availability of training data.

An important note: not all supervised learning techniques are equally informative for the user. Models built, say, by using artificial neural networks may work well but since they belong to a group of black-box methods, it is very difficult to make sense of their weights and connections. On the other hand, inductive machine learning techniques, like rule algorithms or decision trees, generate data models in terms of production rules of the form: IF Condition THEN Conclusion. When conditions are specified in terms of original features (see discussion below), then biologists can easily make sense of them, and thus accept or reject such models.

As always, the reality is somewhere in-between of having entirely labeled training data and completely unlabeled data. Quite often we have at our disposal a few labeled training data points and thousands of data points about which we know nothing. People have realized this fact and research is rapidly growing in the area called semi-supervised learning, i.e. development of algorithms that take advantage of few labeled training data points to help learning algorithms to learn from mostly unlabeled data. Obviously, semi-supervised learning methodology requires hybrid approaches. On the one hand, one could utilize a clustering algorithm and modify it in order to take into account a few labeled data points. Such approach is known as partially-supervised clustering [8]. On the other hand, one could use graph based, learning technique, and couple it with standard techniques taking in this way into account both the labeled and unlabeled data [9–12].

Another interesting area of research is to design learning methods that are inspired by our ever increasing knowledge of how living organisms learn and operate. These include learning strategies like ant colony optimization, inspired by observing ant societies, or learning rules inspired by modeling certain brain regions. Such methods are then used for solving practical problems outside of the realm of biology [13–17].

In the area of data mining, which is understood here mainly as a knowledge discovery process [18,19], there is a need for developing intelligent ways of modeling, which should include self-explanatory data modeling, and what is lacking in most of existing systems, an automatic tuning of the model parameters. The visualization techniques start playing ever more important role in data mining, in particular when dealing with biological data [20,21].

Genome sequencing projects changed the approach of researchers involved in molecular biology. Originally, they focused on individual biological molecules in a cell, using relatively manual experiments and thus generating small amounts of data. However, automating of DNA sequencers allowed for sequencing of many short random segments, and, at present post-genomic era researchers use protein and mRNA microarray chips to generate large amounts of data, for allowing systematic understanding of entire biological molecules in a cell. It is interesting to note that it was the interest of molecular biologists to make wet laboratory experiments high-throughput. Due to the nature of these experiments that generate data faster than they can be analyzed, there is a growing need to use bioinformatics tools for their analyses [22–28]. While data mining contributes to automatic relationship discovery and to a detection of new, previously unknown, facts and new relations from data [29–31], it encounters new challenges when applied to microarray data because they are, in general, not validated experimentally. In addition, very often microarray data provide few data points to make valid inferences. The latter leads us to a requirement for much better dimensionality reduction techniques [32–36]. That is of extreme importance to analysis of gene and protein microarray data where the dimension of the search space is dozens of times greater than the number of the measured data points, say on normal versus abnormal patients.

Of particular importance are feature selection techniques, where one selects a subset of the original features, as opposed to feature extraction techniques where the number of features is reduced in a new, transformed, space. An example of the latter is principal components analysis where the new features are linear combinations of the original features. It is obvious that feature selection is preferred by biologists because the selected features must not only be good for prediction or classification purposes but also must make sense to them. Feature selection techniques range from established statistical techniques like chi-square, *F*-ratio, and *t*-statistics [37], through measures from information theory like mutual information and receiver operating characteristics [38,39], to supervised inductive

machine learning techniques. The latter have been successfully used not only for model building (their main purpose) but also for feature selection via feature (and their values) ranking. In short, the features used in the final model of the data (a subset of original features) are ranked according to their strength in discriminating among the known classes (say normal versus abnormal) [40,41]. Another group of feature selection techniques are heuristic methods like branch-and-bound, stepwise forward selection, and correlation-based feature selection [42].

To cope with gene and protein microarray data, in addition to data analysis techniques, there is also a need to use various structured or semi-structured databases [43–51] as well as unstructured information scattered all over the Web. Doing so requires availability of domain-specific ontology [52] in order to utilize this wealth of information. Integrating unstructured information available on the Web with structured databases will certainly help in solving complex bioinformatics problems.

Automatic understanding of data is another area of research with great promise, yet to be taken advantage of, in bioinformatics. The key difference between automatic (or semi-automatic) recognition and automatic understanding is connected with the number of classes considered. In most recognition methods, like those based on support vector machine models, neural networks, etc., the user often knows the number of classes and the features that describe the classes. Sometimes, the templates for each class are known and can be used as a base for class recognition and differentiation. Then, the only thing required is finding an optimal method for identifying appropriate class for each input. In contrast, automatic data understanding can deal with potentially infinite number of classes [53–55]. The classes in this case are called semantic contents of the data. In practical applications, a finite subset of possibly infinite collection of meanings is used, but it is not known a priori which subset of the semantic descriptions will be the most useful. The difference between automatic recognition and automatic understanding can be explained by analyzing at a data flow scheme. In recognition, there is a one-directional flow of data: the input samples are first preprocessed and then recognized. In automatic understanding, there is a two-directional data flow. One, the data stream goes from the input, and the first step in this stream is the same: preprocessing. The second, however, is linguistic description of the data in terms of the data semantic contents. After describing the sense of the data in terms of sentences according to the rules of a selected grammar, the next step of the automatic understanding process is called ''cognitive resonance''. At this step, the two data flows meet. These additional data flow goes in the opposite direction, from (output) background knowledge about the data, and includes ''demands'' describing desired elements of the linguistic description of the input data. Demands describe desired values of the selected features of the data, and every demand is connected with some semantic interpretation of the input data. The process of cognitive resonance can be explained by using analogy with waves coming from two sources (in our case from input and output streams): sometimes they are added and we obtain the resonance peak, in our case correct understanding of the data. For solving complex bioinformatics problems, automatic understanding of data may become a more powerful methodology than automatic recognition.

2. Bioinformatics: our interpretation of the term and its challenges

In minds of many bioinformaticians, including these authors, the terms bioinformatics and computational biology mean about the same. Recently, however, the National Institutes of Health (U.S.A.) [56–58] came up with slightly different definitions, which for the convenience of the reader are repeated below.

Bioinformatics: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioural, or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.
Computational biology: The development and application of data-analytical and theoretical methods, mathematical modeling, and computational simulation techniques to the study of biological, behavioral, and social systems.

Using these definitions, however, it would be rather difficult to categorize a paper as being a bioinformatics or a computational biology paper. Although there are several journals that use both terms in their titles, the top journal in the area uses only the term *Bioinformatics* as its title. To complicate matters even further, there is a growing subfield of bioinformatics called *Neuroinformatics* (there is one paper in this issue that falls into the category), which, again, after the NIH: ''combines neuroscience and informatics research to develop and apply advanced tools and approaches essential

for a major advancement in understanding the structure and function of the brain''.

Another, although by no means new, but growing sub-field of bioinformatics is *Proteoinformatics*, which can be defined as [59]: ''development of computational methods that allow direct prediction of protein structure and function in silico'', that is driven by the availability of completely sequenced genomes, databases of NMR, and X-ray structures of proteins, and compilations of functional information, such as post-translational modifications [60,61]. Several papers in this special issue fall into this category. As a result, after making the above distinctions, we will use the term bioinformatics in a very broad sense that encompasses not only all of the above but also other fields, not mentioned here.

Bioinformatics research has proven to be very successful. Thanks to the development of advanced biochemical and biophysical instrumentation methods, we are able to collect valuable information about genome and proteome sequences, and structures of biological macromolecules. The collected data, however, are often noisy and ambiguous, and thus the need for better techniques to solve complex problems connected with proper interpretation and plausible reconstruction (in terms of models) of the obtained biochemical information. It requires more accurate and faster database and data processing technologies, and better computational intelligence algorithms. Biochemical and biophysical laboratories collect data only about constituent elements that must be combined, analysed, and processed in order to obtain valid bioinformatics models. Fortunately, several of the existing computational intelligence techniques can be adopted for solving bioinformatics problems, and new methods are being developed almost daily. In general, we can use computational intelligence methods providing that they are wisely combined with the bioinformatics domain needs, as illustrated by authors in this special issue.

There are several major issues facing bioinformatics and they are constantly changing. The most basic data being collected in molecular biology are sequences of genomes and proteomes. Thus issues in bioinformatics are related to sequences, like multiple sequence alignment and finding patterns in sequences called motifs. Analyzing protein structures, e.g. predicting the 2D/3D protein structure, or the binding site of a protein, is another important field of research. A protein controls a chemical reaction of biological components that can bind to the protein thus playing a central role in biochemical process. Another issue is experimental data analysis. Molecular biology provides us a wide variety of experimental data. A traditional example is simple 2D-electrophoresis data. The data, however, have increased in both number and size with the advent of gene and protein microarrays, yeast two-hybridization for identifying protein–protein interactions, and RNA interference. Examples of the challenging problems are efficiently searching for ligands binding to some biologically important protein, estimating networks of biological components from a variety of biological experimental data, or finding key biological molecules causing a disease by using different types of data (including microarray).

3. Overview of the papers

This double special issue on Computational Intelligence Techniques in Bioinformatics presents the reader with 14 papers (chosen from 29 submitted) that can be loosely divided into two general groups. The papers in the first group introduce new or improved computational intelligence techniques for solving bioinformatics problems, while the papers in the second group use more established techniques in a novel way for doing the same. Obviously, such a division is highly arbitrary and thus we refrain from labeling the papers. Instead, we briefly describe contributions of each paper from the first group, and then from the second group. By doing so, we hope, we help the reader to identify papers of most interest.

Mamitsuka proposes a new method for efficient finding of the biologically optimal alignment of multiple sequences. A key technique used in his method is ''deterministic annealing'' that attempts to find the global optimum in a parameter space through the annealing process. The author proposes a new simple probabilistic model for the usually time-consuming iterative process of deterministic annealing. Probabilistic parameters of his model are trained from a given sequences based on the deterministic annealing and Expectation Maximization algorithm. When a new sequence is given, this sequence is aligned by parsing it using the trained model. Experimental results show that the proposed method gives a better performance than other competing methods, like a profile Hidden Markov Models, and is time-efficient.

Ruan, *Wang*, *Yang*, *Kurgan*, and *Cios* address a problem of predicting protein secondary structure contents in a given sequence. The key idea of the paper is to use a novel measure that the authors call a composition moment vector. The new measure includes both information about composition of a given sequence and the position of amino acids in

the sequence, in contrast to the often used composition vector that accounts only for the composition. The authors show that the new measure provides functional mapping between primary sequence and the content. The new measure was validated on more than 11,000 protein sequences, using a neural network, and resulted in predicting helix with an average accuracy of 91.5%, and the strand contents with accuracy of 94.5%.

Chen, *Hsu*, *Lee*, and *Ng* address the problem of protein–protein interactions, and note that the interaction data generated in large-scale experimental studies with high throughput technologies have very high error rates. The authors focus on the network consisting of protein–protein interactions and propose the use of a novel measurement, called Interaction Reliability by Alternative Path (IRAP), to computationally assess the reliability of interactions. IRPA is based on the topological properties of the interaction network. The authors further develop an algorithm called Alternative Path Finder to compute the IRAP values efficiently in large, interconnected, and loopy protein interaction networks. Results on real protein interaction networks showed that IRAP is a good measure for discovering reliable protein interactions.

Dvorkin, *Fadok*, and *Cios* present a novel clustering algorithm for analysis of time-series microarray data. The Simple Multilevel Clustering and Linking (SiMCAL1) algorithm presents a complete feature set not found in either Jarvis–Patrick or in other popular clustering methods, such as hierarchical and *k*-means. SiMCAL1 algorithm is multilevel in that it provides a small number of clearly defined hierarchical levels of clusters, but fewer and better-defined than in traditional hierarchical clustering. It offers linking, that is, a temporal ''leader–follower'' relationship between clusters at the same level in each hierarchy. The clusters are formed by finding chains of associations between near neighbors and are thus not necessarily convex, allowing a close fit to existing patterns in the data. The algorithm was used on the data describing activity of the phosphatidylserine receptor (PSR), a crucial molecular switch in the mediation of inflammatory response. Specifically, apoptotic cells appear to suppress inflammatory by expressing PS which engages the macrophage PSR, while lytic cells do not show such activity. By analyzing the behavior of PSR-related genes in mouse macrophages, the authors plan to elucidate the mechanisms involved in this important biological process.

Zhao, *Fanning*, and *Lane* address a problem of finding a minimum set of initiator molecules, short interfering RNAs (siRNAs), for RNAi-based gene family knockdown experiments. RNA interference (RNAi) is a recently discovered genetic technique with widespread therapeutic and genomic applications. They show that the problem of minimizing the number of siRNAs is NP-Hard via a reduction to the set cover problem. They generalize the basic problem by incorporating additional biological constraints and optimality criteria. They provide a branch-and-bound type algorithm for the new problem and show the constraints reduce the search space enough to compute the exact minimal siRNA within reasonable time. They further propose a probabilistic greedy algorithm for this problem for larger cases.

Suzuki, *Tsuji*, and *Ohtake* remind us that living organisms have built-in mechanisms that adapt to various conditions. If these mechanisms could be reproduced on the computer, it may be possible to use biological adaptation methods to engineering of artificial machines. The paper focuses on the nematode (*C. elegans*) that has a relatively simple structure. The authors developed its computer model, artificial *C. elegans*, to analyze control mechanisms with respect to motion by focusing on gentle touch stimulation. Their model consists of a neuronal circuit model for motor control that responds to touch stimuli, and a kinematic model of the body for movement. All parameters included in the neuronal circuit model are adjusted by using the real-coded genetic algorithm. They also use a neuronal oscillator to generate the sinusoidal movement. By using such artificial organism, it may be possible to predict some characteristics that cannot be measured in actual biological experiments.

Arredondo, *Neelakanta*, and *De Groff* address an issue of identifying the delineation/border of separation between codon and non-codon regions in a massive stretch of a DNA chain, when codon and non-codon parts overlap. The authors developed a fuzzy inference engine that scores the extent of nucleic acids to differentiate codon and non-codon regions on a DNA sequence. Their system uses information-theoretic metrics derived from statistical divergence, distance, and discriminant analysis concepts. The authors present simulated studies using human as well as bacterial codon statistics to illustrate the efficiency of their approach.

Narang, *Sung*, and *Mittal* propose a new method for predicting promoter sites in a given sequence. The authors develop a promoter site prediction system, called BayesProm, that infers each sequence fragment in a given sequence as promoter or non-promoter, based on a Bayesian network classifier. BayesProm has a variety of unique features including the use of *oligonucleotide positional density* for modeling promoter sequence characteristics, which they show superior to the often used

"oligonucleotide occurrence frequency analysis". BayesProm directly uses the oligonucleotide positional densities in a continuous variable Bayesian network. BayesProm reports higher sensitivity than any other existing tool at low positive predictive values. The authors claim that their methodology extends the scope of statistical AI models in promoter prediction.

Tang, *Jin*, and *Zhang* address a problem of predicting protein homology between given two proteins. They propose a learning method that combines the idea of association rules with their previous method called Granular Support Vector Machines (GSVM), which systematically combines a SVM with granular computing. Their method, called GSVM-AR, uses association rules with high enough confidence and significant support to find suitable granules to build a GSVM with good performance. The authors compared their method with SVM by KDDCUP04 protein homology prediction data. From the experimental results, GSVM-AR showed significant improvement compared to building one single SVM.

Blazewicz, *Lukasiak*, and *Milostan* focus on the protein folding problem. The HP-model (H—hydrophobic amino acid, P—polar) is a well-studied model that simply deals with a binary (H or P) sequence for investigating protein folding. The HP-model was based on observations that most of hydrophobic amino acids are buried in the core of a protein. Even for such a simple model, it is difficult to formulate fast and realistic folding rules that lead protein to the native state. The authors check the usability of the "tabu search" meta-heuristic strategy for finding low energy structures of the HP-model, while walking efficiently through a conformational space. Experimental results show that the approach is competitive with other methods for a simplified 2D as well as a more complicated 3D HP-model.

Karpenko, *Shi*, and *Dai* focus on a problem of predicting peptides (short protein sequences) that bind to a major histocompatibility complex (MHC) molecule. They propose a method based on Ant Colony System (ACS) to align binders and then construct a position-specific score matrix from the resulting alignment. ACS is a search strategy of ants to find the shortest path from the nest to food sources by communicating via a collective memory consisting of pheromone trails. The performance of their method was evaluated on several benchmark datasets, and demonstrated that: (1) the predictive power of the scoring matrix is comparable to one of the most advanced predictors currently in use, i.e. the Gibbs method, and (2) the method converges to better alignments.

Shim and *Lee* point out that a two-dimensional electrophoresis (2DE) used for identifying target proteins in a tissue had a disadvantage because landmark spots must be annotated manually to correct possible geometric distortions of spots in a protein gel image. The authors propose a method of identifying candidate landmark spots in a gel image automatically. The method identifies the common properties of landmark spots by a clustering algorithm, and summarizes them to a landmark profile. The method then finds landmark spots in a given gel image based on the stored landmark profile by using a A^* search algorithm.

Okada, *Sahara*, *Mitsubayashi*, *Ohgiya*, and *Nagashima* point out that although hierarchical clustering has been extensively used in analyzing expression patterns in microarray gene expression data, their biological interpretations are not easy to determine by analyzing the resulting clusters. The authors propose a novel algorithm that automatically finds biologically interpretable cluster boundaries in hierarchical clustering by referring to gene annotations stored in public genome databases. In addition, the proposed algorithm has a new function of generating a set of clusters that are independent of each other with respect to the distributions of gene functions. The authors claim that this function would enable investigators to efficiently identify non-redundant and biologically-independent clusters.

Huang and *Kecman* present the solution to a difficult problem of using DNA microarrays for cancer diagnosis, where thousands of gene expressions are measured but usually only a small number of patients' samples are available. They use improved recursive feature elimination with support vector machines (RFE-SVMs) to select the most relevant genes. The paper stresses the use of the penalty parameter C, to influence the results, and analyzes how normalizing and scaling operations influence classification results and gene selection. Using both, they showed reduction in a diagnosis error to be about 37%. They also compare gene ranking with other well-known gene selection methods and show that there is a significant consensus among the various algorithms as to which set of genes is relevant.

References

[1] Stewart CA. Bioinformatics: transforming biomedical research and medical care. Commun ACM 2005;47:31—3.
[2] Kohonen T. Self-organizing maps. Berlin: Springer-Verlag, 1995.
[3] AutoClass: http://ic.arc.nasa.gov/projects/bayes-group/autoclass/ (last accessed: 1 September 2004).

[4] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Belmont, CA: Wadsworth, 1984.
[5] Duda RO, Hart PE, Stork DG. Pattern classification. New York: Wiley, 2000.
[6] Duin R. Classifiers in almost empty spaces. In: Proceedings of the 15th international conference on pattern recognition, vol. 2; 2000. p. 1—7.
[7] Simon R. Supervised analysis when the number of candidate features (p) greatly exceeds the number of cases (n). SIGKDD Explorations 2003;5:31—6.
[8] Pedrycz W, Waletzky J. Fuzzy clustering with partial supervision. IEEE Trans SMC Part B 1997;27:787—95.
[9] Zhou D, Bousquet O, Lal TN, Weston J, Schölkopf B. Learning with local and global consistency. In: Thrun S, Saul L, Schölkopf B, editors. Advances in neural information processing systems, vol. 16. Cambridge, MA: MIT Press; 2004 p. 321—8.
[10] Zhu X-J, Ghahramani Z, Lafferty J. Semi-supervised learning using Gaussian fields and harmonic functions. In: Proceedings of the 20th international conference on machine learning (ICML-2003); 2003.
[11] Huang T-M, Kecman V. Semi-supervised learning from unbalanced labeled data an improvement. In: Negotia MG., Howlett RJ, Jain LC, editors. Knowledge based and emergent technologies relied intelligent information and engineering systems Lecture notes on computer science, vol. 3215. Heidelberg: Springer-Verlag; 2004. p. 765—71.
[12] Huang T-M, Kecman V. Performance comparisons of semi-supervised learning algorithms. In: Proceedings of 22nd international conference on machine learning, ICML 2005. New York: ACM Digital Library, 2005.
[13] Sala DM, Cios KJ. Solving graph algorithms with networks of spiking neurons. IEEE Trans Neural Netw 1999;10:953—7.
[14] Sala DM, Cios KJ. Self-organization in networks of spiking neurons. Aust J Intell Inf Process Syst 1998;5:161—70.
[15] Bonabeau E, Dorigo M, Theraulaz G. Swarm intelligence: from natural to artificial systems. Oxford: Oxford University Press, 1999.
[16] Dorigo M, Stutzle T. Ant colony optimization. Cambridge: MIT Press, 2004.
[17] Swiercz W, Cios KJ, Staley K, Kurgan L, Accurso F, Sagel S. New synaptic plasticity rule for networks of spiking neurons. IEEE Trans Neural Netw, in press.
[18] Cios KJ, Pedrycz W, Swiniarski R. Data mining methods for knowledge discovery. Norwell, MA: Kluwer, 1998.
[19] Cios KJ, Teresinska A, Konieczna S, Potocka J, Sharma S. Diagnosing myocardial perfusion SPECT bull's-eye maps—a knowledge discovery approach. IEEE Eng Med Biol Mag 2000;19(4):17—25.
[20] Kovalerchuk B, Delizy F. Visual data mining using monotone Boolean functions. In: Kovalerchuk B, Schwing J, editors. Visual and spatial analysis. Dordrecht: Springer, 2004 [chapter 16].
[21] Vitayev E, Kovalerchuk B. Visual data mining with simultaneous rescaling. In: Kovalerchuk B, Schwing J, editors. Visual and spatial analysis. Dordrecht: Springer, 2004 [chapter 15].
[22] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531—7.
[23] Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci USA 2000;97:262—7.
[24] Csete ME, Doyle JC. Reverse engineering of biological complexity. Science 2002;295:1664—9.
[25] Kitano H. Systems biology: a brief overview. Science 2002; 295:1662—4.
[26] D'haeseleer P, Liang S, Somogyi R. Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics 2000;16:707—26.
[27] Akutsu T, Kuhara S, Maruyama O, Miyano S. A system for identifying genetic networks from gene expression patterns produced by gene disruptions and overexpressions. Genome Inform 1998;9:151—61.
[28] Somorjai R, Dolenko B, Baumgartner R. Class prediction and discovery using gene microarray and proteomics mass spectrometry data: curses, caveats, cautions. Bioinformatics 2003;19:1484—91.
[29] Tuzhilin A, Adomavicius G. Handling very large numbers of association rules in the analysis of microarray data. In: ACM SIGKDD conference on knowledge discovery and data mining. 2002. p. 3960—404.
[30] Jiang XR, Gruenwald L. Microarray gene expression data association rules mining based on JG-tree. In: Proceedings of the 14th international workshop on database and expert systems applications; 2003. p. 27—31.
[31] Liping J, Kian LT. Mining gene expression data for positive and negative co-regulated gene clusters. Bioinformatics 2004;20:2711—8.
[32] Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA 2000;97:10101—6.
[33] Wall M, Dyck E, Brettin TS. SVDMAN—singular value decomposition analysis of microarray data. Bioinformatics 2001;17:566—8.
[34] Raychaudhuri S, Stuart J, Altman RB. Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac Symp Biocomput 2000;455—66.
[35] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learn 2002;46:389—422.
[36] Xia X, Xie Z. AMADA, analysis of microarray data. Bioinformatics 2000;17:570—669.
[37] Liu H, Li J, Wong L. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. Genome Inform 2002;13:51—60.
[38] Cover T, Thomas J. Elements of information theory entropy relative entropy and mutual information, Chichester: Wiley, 1991. p. 12—49 [chapter 2].
[39] Pepe MS. Receiver operating characteristic methodology. J Am Stat Assoc 2000;95:308—11.
[40] Kononenko I. Estimating attributes: analysis and extensions of RELIEF. In: Proceedings of the European conference on machine learning. New York: Springer-Verlag; 1994. p. 171—82.
[41] Cios KJ, Kurgan L. CLIP4: hybrid inductive machine learning algorithm that generates inequality rules. Inf Sci 2004;163:37—83.
[42] Hall M, Holmes G. Benchmarking attribute selection techniques for discrete class data mining. IEEE Trans Knowl Data Eng 2003;15:1437—47.
[43] GenBank: http://www.ncbi.nlm.nih.gov/Genbank/index.html (last accessed: 12 May 2005).
[44] EMBL: http://www.embl-heidelberg.de/ (last accessed: 12 May 2005).
[45] GDB: http://gdbwww.gdb.org/ (last accessed: 12 May 2005).
[46] UniGene: http://www.ncbi.nlm.nih.gov/UniGene/ (last accessed: 12 May 2005).
[47] OMIM: http://www.ncbi.nlm.nih.gov/Omim/ (last accessed: 12 May 2005).

[48] GeneCards: http://bioinformatics.weizmann.ac.il/cards/ (last accessed: 12 May 2005).
[49] NCBI SAGEmap: http://www.ncbi.nlm.nih.gov/SAGE/ (last accessed: 12 May 2005).
[50] Stanford Microarray Database: http://genome-www4.stanford.edu/MicroArray/MDEV/ (last accessed: 12 May 2005).
[51] KEGG: http://www.genome.ad.jp/kegg/kegg2.html (last accessed: 12 May 2005).
[52] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25—9.
[53] Tadeusiewicz R, Ogiela MR. Medical image understanding technology. Berlin/Heidelberg/New York: Springer, 2004.
[54] Tadeusiewicz R, Ogiela MR. Intelligent recognition in medical pattern understanding and cognitive analysis. In: Sarfraz M, editor. Computer-aided intelligent recognition techniques and applications. Hoboken, NJ: John Wiley & Sons Ltd.; 2005. p. 257—74.
[55] Tadeusiewicz R, Ogiela MR. Automatic understanding of medical images—new achievements in syntactic analysis of selected medical images. Biocybern Biomed Eng 2002; 22(4):17—29.
[56] NIH: http://www.bisti.nih.gov (last accessed: September 2004).
[57] NIH: http://grants.nih.gov (last accessed: September 2004).
[58] NIH Roadmap: http://nihroadmap.nih.gov/ (last accessed: September 2004).
[59] Hamady M, Cheung T, Resing K, Cios KJ, Knight R. Key challenges in proteomics and proteoinformatics. IEEE Eng Med Biol Mag 2005;24(3):34—40.
[60] Helmke SM, Yen C, Cios KJ, Nunley K, Bristow MR, Duncan M, et al. Simultaneous quantification of human cardiac α and β-myosin heavy chain protein by MALDI-TOF mass spectrometry. Anal Chem 2004;76:1683—9.
[61] Yen C, Helmke SN, Cios KJ, Perryman MB, Duncan M. Quantitative analysis of proteomics using data mining. IEEE Eng Med Biol Mag 2005;24(3):67—72.

Krzysztof J. Cios*
University of Colorado at Denver and Health Sciences Center, Department of Computer Science & Engineering
Campus Box 109, Denver, CO 80217, USA

Hiroshi Mamitsuka
Kyoto University, Gokasho, Uji 611-0011, Japan
E-mail address: mami@kuicr.kyoto-u.ac.jp

Tomomasa Nagashima
Muroran Institute of Technology and Life Software Laboratory, 27-1, Mizumoto, Muroran, Hokkaido 050-8585, Japan
E-mail address
nagasima@epsilon2.csse.muroran-it.ac.jp

Ryszard Tadeusiewicz
AGH University of Science and Technology
Al. Mickiewicza 30, 30-059 Krakow, Poland
E-mail address: rtad@agh.edu.pl

*Corresponding author. Tel.: +1 303 556 6035
fax: +1 303 556 8369