

Rough Sets in Bioinformatics

Torgeir R. Hvidsten and Jan Komorowski

The Linnaeus Centre for Bioinformatics, Uppsala University, Uppsala, Sweden
hvidsten@lcb.uu.se, janko@lcb.uu.se

Abstract. Rough set-based rule induction allows easily interpretable descriptions of complex biological systems. Here, we review a number of applications of rough sets to problems in bioinformatics, including cancer classification, gene and protein function prediction, gene regulation, protein-drug interaction and drug resistance.

1 Introduction

Molecular biology represents a fascinating and important application area for machine learning techniques in general and rough set-based methods in particular. Although biology traditionally has been a reductionistic discipline focusing on breaking living systems into increasingly smaller parts, and on studying these parts separately, the discovery of the remarkable order and structure of these systems at the molecular level has suggested the possibility of studying their holistic molecular operation. However, it is only in the last 10 years that technological breakthroughs have made it possible to obtain large scale data that can facilitate such research. The first complete genome was sequenced in 1995 (the bacteria *Haemophilus influenzae* Rd [1]) and has been followed by many others (including the human genome, see <http://www.genomesonline.org>, [2]). Although important, sequence information only gives us the static code inherited from individual to individual. Other insights such as identifying the functional elements (i.e. genes) of the genomic sequence, understanding how and under which conditions genes are transcribed and translated into protein(s) (i.e. gene regulation) and determining the tasks/interactions carried out by each protein (i.e. protein function) require data on the dynamic operation of biological systems under different conditions. One example of a technology that provides this type of data is DNA microarrays that can measure the transcription levels of thousands of genes in parallel [3,4]. Moreover, developing technology will soon be able to directly perform similar large-scale measurements of proteins (i.e. proteomics, [5]).

High-throughput experimental technologies have created the need for computer programs and techniques to analyze the resulting data. The field of bioinformatics has thus developed from being a discipline mainly associated with sequence databases and sequence analysis to a computational science that uses different types of data to describe biology [6]. The ultimate goal of this research is to allow computational simulations of complex living systems. This will presumably require that we determine the function of all sequenced proteins (functional

genomics) and that we understand the general principles orchestrating protein regulation and interaction (systems biology).

Over the years, biologists have accumulated a large amount of knowledge about the specific functions of individual proteins. This has been accomplished through carefully chosen experimental strategies that often start out with a hypothetical function that is then confirmed or rejected in the laboratory. As the sequence databases grew larger, experimental biology was revolutionized by computational sequence similarity search methods such as BLAST [7]. These programs can align functionally uncharacterized protein sequences with protein sequences of known function, and identify statistically significant sequence similarities. Such similarities indicate that the two proteins have a common evolutionary ancestor and that, although their amino acid sequences have diverged over time, their functions have remained similar. The relationship between growth proteins and cancer was discovered in this way (by Doolittle in the early 1980s). This first success story of bioinformatics further suggested a general strategy for using machine learning in molecular biology, that is, to take advantage of the assumed relationship between high-throughput data such as sequence data and available knowledge of, for example, protein function to induce general models that represent this relationship. These models may then be used to predict function for uncharacterized proteins and thus provide experimentalists with novel hypotheses that can be tested in the laboratory. Furthermore, these models may give us valuable biological insight such as, in the case of function prediction, the location of the functional site of a protein.

One problem with the machine learning strategy to functional genomics is that most knowledge of protein function only exists as plain text in scientific publications. For this reason, text mining has been and will continue to be an important part of bioinformatics [8,9]. Furthermore, large efforts have been put into developing controlled vocabularies and data structures for representing knowledge in a computer readable form [10]. A prominent example is Gene Ontology (GO) [11]. GO consists of three sub-ontologies that describe three different aspects of protein function. Molecular functions are tasks performed by single proteins, biological processes are ordered assemblies of molecular functions that together carry out broad biological goals in the cell and cellular components are subcellular locations where proteins are active. Each ontology is a directed acyclic graph (DAG) where nodes describe, for example, molecular functions at different levels of specificity (called GO terms) and edges represent the relationships between different GO terms. For example, GO tells us that the *cell cycle* is a part-of *cell proliferation* and that the *mitotic cell cycle* is-a *cell cycle*. The advantage of GO is that the current knowledge about the function of a gene or a protein may be represented by associating it with one or more terms in GO, and that the structure of GO makes it easy to write computer programs that can compare and organize these annotations for many or all genes in a genome.

Given a set of training examples, e.g. sequences with GO annotations, the application of supervised machine learning is far from trivial. High-throughput experimental data is inevitably obscured by a relatively large amount of noise. In

addition, the training examples are reflections of the currently available knowledge about the function of a protein and may thus be incomplete or, in the worst case, wrong. This problem is made worse by the fact that the biological knowledge used to build the training examples is often automatically retrieved from text or even computationally inferred from e.g. high sequence similarity. Finally, functional genomics presents us with particularly difficult challenges related to learning, including an often large number of functional classes to discriminate, examples belonging to several different functional classes and classes with few examples. These challenges make it especially important to choose good methods for validating the statistical and biological significance of the induced models. Furthermore, it demands a lot from the applied machine learning method.

Rough set theory [12,13,14] is founded on the concept of discernibility, i.e. that data may be described only in terms of what differentiates relevant classes of observations. From the concept of discernibility, decision rules are constructed by extracting minimal information needed to uphold the discernibility structure in the data set [15,16]. The fact that the framework does not attempt to discern objects that are equal or objects that are from the same class (e.g. have the same function), makes it possible to describe incomplete and conflicting data in terms of easily interpretable decision rules. In this article, we will review some of the successful studies in which rough set-based rule induction has been used to describe biological systems at the molecular level. These studies include

- cancer classification using gene expression data,
- prediction of the participation of genes in biological processes based on temporal gene expression profiles,
- modeling of the combinatorial regulation of gene expression,
- prediction of molecular function from protein structure,
- prediction of protein-ligand interactions in drug discovery, and
- modeling of drug resistance in HIV-1

and are modeled using rules such as

- **IF** Gene A is up-regulated **AND** Gene D is down-regulated
THEN Tissue is healthy
- **IF** Transcription factor F binds **AND** Transcription factor V binds
THEN Gene is co-regulated with Gene H
- **IF** Protein contains motif J
THEN Function is magnesium ion binding **OR** copper ion binding
- **IF** Protein contain motif D **AND** Ligand water-octanol coeff. $> c_1$
THEN Binding affinity is high
- **IF** Change in frequency of alpha-helix at position X $> c_3$
THEN Resistant to drug W

In particular, we will focus on how these application areas have been coded in a discrete manner to facilitate rule induction, how biological knowledge can be incorporated into this representation process and what can be read out of the rule model in terms of biological insight. Technical details will not be discussed here and may be found in the respective publications.

2 Gene Expression Analysis

The complementary nature of the DNA double helix is of great importance to replication and transcription in living organisms, and may also be utilized for the large-scale measurement of mRNA levels in cells. Two complementary nucleic acid molecules (i.e. strands) will combine under the right conditions to form double stranded helices. In a reaction vessel this is referred to as hybridization. Hence, it is possible to use identified DNA strands (probes) to query complex populations of unidentified, complementary strands (targets) by checking for hybridization. Microarrays are glass slides or wafers populated with large numbers of strands derived from identified genes. By applying a target sample of unidentified mRNA to the array, the expression level of each gene probe may be quantified from the extent of hybridization between the probes and the targets. Since one slide may contain probes from thousands of genes, one microarray experiment may determine the genome-wide expression state of a cell sample. Furthermore, systematic series of microarray experiments may reveal the specific changes in cellular gene expression associated with different physiological or pathophysiological responses. A microarray study comprises a number of steps including experimental design [17], filtering and normalization of the data [18] and high-level computational data analysis. The last step was in the early phase of microarray analysis mostly restricted to clustering analysis, and in particular, hierarchical clustering [19]. However, the limitations of clustering both in terms of interpretation and evaluation soon saw a shift in focus from unsupervised learning (i.e. clustering) to supervised learning [20,21].

2.1 Cancer Classification

Standard medical classification systems for cancer tumors are based on clinical observations and the microscopical appearance of the tumor. These systems fail to recognize the molecular characteristics of the cancer that often corresponds to subtypes that need different treatment. Studying the expression levels of genes in tumor tissue may reveal such subtypes and may also diagnose the disease before it manifests itself on a clinical level. Thus, the goal of data analysis of cancer microarray data is to develop models for earlier detection and better understanding and treatment of cancer.

Gastric Carcinoma. Midelfart *et al.* [22,23] used rough set-based classifiers to identify molecular markers that allow classification of gastric carcinoma. Gastric carcinoma is often not detected until at an advanced stage, which is one of the reasons why this is the second most frequent cause of cancer death world-wide. The study developed classifiers for six different clinical parameters; intestinal or diffuse types (also known as the Lauren classification), site of primary tumor (cardia, corpus or antrum), penetration of the stomach wall or not, lymph node metastasis or not, remote metastasis or not, and high or normal serum gastrin. The expression levels of 2504 genes were measured in tumor samples taken from only 17 patients. Rule models were induced in a

leave-one-out cross-validation procedure for each of the six clinical parameters. In each iteration, the 10 to 40 most differentially expressed genes were identified using a bootstrap t-test [24]. By differentially expressed genes, we here mean genes that showed a consistently higher expression in e.g. intestinal samples compared to the diffuse samples as measured by the bootstrap t-test. The expression of these genes was discretized into e. g. low, medium and high expression and rules were induced. Classification accuracy and area under the receiver operating characteristics curve (AUC) [25] were reported for all six clinical parameters ranging from 0.79 to 1.00 (average 91.5) and 0.66 to 1.00 (average 0.89), respectively.

A particularly difficult challenge in cancer classification from microarray data is the large number of measured genes compared to the number of cancer patients. This is a problem because one is faced with a huge search space (i.e. subsets of 2504 genes) and only a few data points to restrict the search. A possible consequence could be overfitting, that is, decision rules that explain the training set, but fail to generalize to the test set. In this study, reduct computation was limited to a low number of differentially expressed genes. However, the number of genes compared to the number of patients still makes it difficult to exclude the possibility that some of these genes are discriminatory by chance. Thus, to further add robustness to the identification of gastric carcinoma markers, the study reported as a measure of strength the number of cross validations in which a particular gene was part of at least one decision rule in the rule model. This resulted in the identification of several genes known to be highly expressed in gastric carcinomas as well as several interesting new genes.

The rule induction process offer a number of algorithms for discretization and reduct computation. Combined with a low number of training examples, these options constitute a real risk that even cross validation estimates may be optimistic in the sense that they do not reflect a true ability to correctly classify unseen samples. The authors of the study realized this, and consecutively repeated the cross validation procedure for each of the six clinical parameters on 2000 dataset where the clinical parameter values were randomly shuffled. By recording the fraction of randomized data sets that resulted in a higher AUC value than the real data set, they obtained a p-value for each clinical parameter reflecting the probability that the reported AUC value could be obtained by chance [26]. Even though the initial cross validation estimates looked impressive, this careful analysis showed that the AUC value of three of the six clinical parameters were not statistically significant at p-value 0.05. However, location of tumor ($p < 0.031$), lymph node metastasis ($p < 0.007$) and the Lauren classification ($p < 0.007$) were shown to be adequately described by the rule model.

Adenocarcinoma. Dennis *et al.* [27] used rough sets to build a classification system for identifying the primary site of cancer based on expression levels in a sample taken from a secondary tumor. While it is the primary tumor that causes symptoms in most patients, about 10-15% of cancers are discovered as metastases in solid organs, body cavities or lymph nodes. Most of these secondary tumors are adenocarcinomas, for which the seven commonest primary sites are

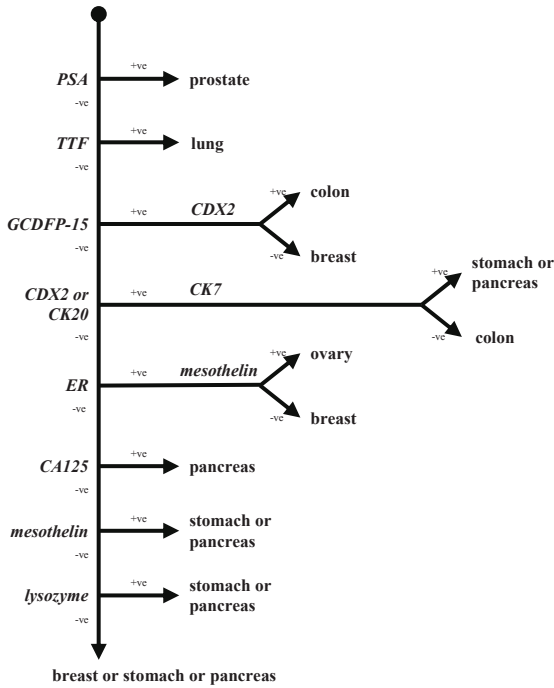


Fig. 1. Decision tree used to predict the site of origin of metastatic cancer from 10 molecular markers [27]

breast, colon, lung, ovary, pancreas, prostate and stomach. Because prognosis and therapy are linked to the site of origin, and because histologically such tumors appear similar, finding molecular markers for these sites could greatly improve treatment.

The study assessed the expression patterns of 27 markers in 452 adenocarcinoma patients. 12 markers were scored as either present or absent (+ or -), while the remaining markers were scored as absent, weak, intermediate or strong (0, 1, 2 or 3). Decision rules were induced from 352 adenocarcinomas and used to build a decision tree of 10 markers (see Figure 1). This tree was then used to predict the site of origin of 100 unseen adenocarcinomas with a success rate of 88%. This is a very high accuracy considering there were seven different sites to predict, and indicate a huge potential for molecular markers in identifying the primary site of these cancers.

2.2 Predicting Participation of Gene Products in Biological Processes

Hvidsten *et al.* [28,29,30] developed a method for modeling the participation of gene products in GO biological process from temporal expression profiles. Several publications had earlier used hierarchical clustering to illustrate the correspondence between expression similarity and gene function [31,32,19]. However, none

of these studies actually quantified the relationship. Furthermore, it is known that functionally related genes often are anti-coregulated and that genes usually are associated with more than one function. These aspects are not well modeled by a set of broad, non-overlapping expression clusters. Brown et al. [21] was the first to approach the problem in a supervised manner by using support vector machines to predict a limited set of six functional categories from expression data.

In the 2001 paper [28], a template language was proposed to describe the discrete changes in expression over subsets of time points in an expression time profile. The idea behind this language was that the relative change in mRNA levels over limited periods of time is more important to distinguish one biological process from another than the absolute mRNA levels given by each time point. Furthermore, rough set-based rule induction was used to associate combinations of discrete changes in expression with one or a small number of GO biological processes. For example, the rule

```
IF      0 - 4(constant) AND 0 - 10(increasing)
THEN GO(protein metabolism and modification) OR
        GO(mesoderm development) OR GO(protein biosynthesis)
```

and the rule

```
IF      0h-4h (increasing) AND 6h-10h (decreasing)
        AND 14h-18h (constant)
THEN GO (cell proliferation) OR GO (cell-cell signaling) OR
        GO (intracellular signaling cascade) OR GO (oncogenesis)
```

describe the limited set of biological processes (THEN-part) associated with particular expression profile constraints (IF-part, e.g. 0h-4h (increasing) means increasing expression level from 0 to 4 hours). The first rule has a support of five genes, four of which are annotated to *protein metabolism and modification*. The second rule has support of four, three of which were annotated to *cell proliferation*. Thus the main reason for indeterministic rules is that genes are annotated to several different GO terms.

The predictive performance of the approach was tested using cross validation on all annotated genes in two expression time profile data sets with human genes [19,33]. Thus the correspondence between expression similarity and GO biological process was properly evaluated and quantified for 23 and 27 biological process, respectively. Each biological process was subjected to a permutation test that showed that most of the classes indeed could be predicted with a statistically significant AUC value not obtainable by chance.

The cross validation results may in general be considered estimates of the prediction quality one can expect when predicting functionally uncharacterized genes using a model induced from *all* training examples. However, predictions to uncharacterized genes were also evaluated directly by searching for homology information that could be used to make assumptions about the biological processes of these genes [30]. Of the 24 genes where such assumptions could be made, 11 genes had one or more classifications that matched this assumption.

In addition to predicting the biological process of uncharacterized genes, a model induced from all examples was also used to re-classify characterized genes [30]. The resulting false positives were then used to guide a second literature search for possible missing annotations (i.e. information on biological process annotations existing in the literature, but overlooked during the initial literature search). Of the 14 genes with a false positive re-classification to DNA metabolism, four were found to actually participate in this process. Furthermore, it was revealed that 12 of the 24 false positive re-classifications to oncogenesis also represented missing annotations. Thus, it was shown that computational models could be used directly both to guide new literature searches for partially characterized genes and to propose new functional hypotheses for unseen genes.

The studies described here all used a set of predefined biological processes as basis for learning. Midelfart *et al.* [71-73] later introduced rough set-based rule classifiers that actively learn in the Gene Ontology graph, dynamically selecting biological processes with the best predictive performance.

2.3 Gene Regulation

One of the major challenges faced by molecular biology is to dissect the regulatory circuitry of living cells. The ability of transcription factors to selectively bind specific DNA motifs (i.e. transcription factor binding sites) in the regulatory regions of genes is essential for the complex regulation systems observed in living organisms. The assumption that genes regulated by the same transcription factors (i.e. co-regulated) should contain common binding sites and exhibit similar expression (i.e. co-expressed) enables the study of gene regulation at a genome-wide scale using sequence and expression data.

Pilpel *et al.* [34] found that genes sharing pairs of binding sites are significantly more likely to be co-expressed than genes with only single binding sites in common. This result is in agreement with the hypothesis that a limited number of transcription factors combine in various ways in order to respond to a large number of various stress conditions.

Hvidsten *et al.* [35] used rough set-based rule induction to perform a comprehensive analysis of the combinatorial nature of gene regulation in yeast. The method extracted IF-THEN rules of minimal binding site combinations or modules (IF-part) shared by genes with a common expression profile (THEN-part). The rules hence described general, underlying relationships in an easily understandable format, providing hypotheses on combinatorial co-regulation that may later be experimentally validated.

The approach was tested on a database of known and putative regulatory sequence motifs in yeast [36] using six expression data sets including one cell cycle study and five studies including different stress conditions [34]. The rule learning framework was subsequently applied to each gene to obtain rules that associate the expression profile of that gene with a minimal binding site combination shared by similarly expressed genes. Rules were then discarded if they did not provide a clear and general pattern in terms of modules associated with several

genes where a majority had similar expression. Only in these cases the evidence for actual co-regulation was considered sufficiently strong.

The discovered binding site modules were evaluated using transcription factor binding interactions provided by a genome-wide location analysis [37] and Gene Ontology annotations. The evaluation clearly showed that the retrieved binding site modules reflected actual co-regulation and furthermore showed that genes associated with these modules very often share biological roles in terms of biological process, molecular function and cellular component. The results were statistically significant compared to genes either associated with a randomly chosen set of binding sites, similar expression or neither of these constraints.

Two rules were discussed as a case study and had support in the literature. As an example, the rule

IF RAP1 AND MCM1 AND SWI5 THEN Similar expression

describes eight genes and suggests that the transcription factor RAP1 (that regulates genes that encode ribosomal proteins in growing yeast cells, but also other non-ribosomal genes) requires the cell cycle regulating transcription factors MCM1 and SWI5 to be present when specifically targeting ribosomal genes in growing yeast. That is, the ribosomal genes targeted by RAP1 are only regulated when the cell is in the cell cycle (i.e. growing) which is when MCM1 and SWI5 are present. RAP1 presumably combines with other transcription factors when regulating other non-ribosomal genes.

By applying the method to expression data obtained under several different conditions the authors were able to discover a number of binding site modules common to several of these responses in addition to modules that seem to be exclusive to a particular stress response. The overlap between modules clearly shows the large extent to which relatively few transcription factors combine to facilitate a much large number of expression outcomes.

A later follow-up study [38] used expression similarity restricted to subintervals of cell cycle time profiles (similar to the template language discussed in section 2.2), and showed that this improvement greatly increased the biological significance of the retrieved modules as well as making it possible to retrieve modules that were not detectable using expression similarity over the whole time profile. A second follow-up study [39] refrained from using expression similarity altogether. Instead, this study used prior knowledge of the cell cycle period time to detect different classes of periodically expressed genes in three different synchronization studies, and then used rough set-based rule learning to describe the regulatory mechanisms behind these classes. These mechanisms were then shown to be much more specific towards the cell cycle machinery than mechanisms discovered from expression clusters, and thus showed the advantage of incorporating biological knowledge into the data analysis process whenever it is possible.

3 Protein Analysis - Function and Interaction

It is believed that sequence similarity search methods can identify functionally characterized homologues for less than 50% of the proteins predicted from

genome sequencing projects. However, even though global sequence similarity between distantly related proteins may be virtually undetectable, similarities may still be present in terms of conserved amino acids in the functional sites (functional sites are known to be more conserved than the overall sequence), conserved global structure (structure is known to be more conserved than sequence) or conserved local structure related to the functional site (again, functional sites are more conserved also in terms of structure). Thus more advanced sequence similarity methods and methods using structural similarities may represent a solution for proteins where functional hypotheses cannot be obtained from global sequence similarity [40]. Unfortunately, the Protein Data Bank (PDB) [2] only contains around 30 thousands protein structures while there are about 30 million protein sequences in UniProt (Universal Protein Resource) [41]. To remedy this situation, structural genomics projects systematically aim at solving protein structures for new protein families [42], using these structures as templates for in silico structure prediction methods (i.e. homology modeling) [43], and then applying the solved and predicted structures to infer function [44]. However, to be successful this strategy requires new and improved methods that utilize structure to predict function and interactions.

Here we will review research using rough set-based rule induction to model protein function and interaction. Two of these studies describe protein structure in terms of local descriptors of protein structure. A local descriptor is defined by A. Kryshatovych and K. Fidelis as a set of short backbone fragments centered in three dimensional space around a particular amino acid [45]. By generating local descriptors for all amino acids and all proteins in PDB, and by clustering these descriptors into groups of structurally similar descriptors, it is possible to build a library of a few thousand local building blocks from which virtually all proteins in PDB may be assembled (see Figure 2). This library of recurring local substructures may then be used for representing and comparing protein structure.

3.1 Function Prediction from Structure

Although global structural similarity is often a sign of function similarities [46], many folds such as the TIM barrel and the Rossmann fold are found in proteins with many different functions. Thus local similarity methods are more powerful in these cases [47]. Recently, researchers have started building tools that use a large number of different features including both local and global structure [48,49]. These so-called meta-servers obtain functional predictions by allowing a large number of different evidence to vote, and then selecting the most likely function. However, such approaches do not construct explicit models that are often very useful in further analysis.

Hvidsten *et al.* [50] proposed a change in this paradigm by inducing IF-THEN rules that associate combinations of local substructures with specific protein functions (Figure 3). This approach differs from other studies in that the applied library of local substructures encompasses *all* recurring motifs and *all* annotated proteins using no prior knowledge of functional sites or any sequence information, and in that the structure-function relationship is explicitly represented in

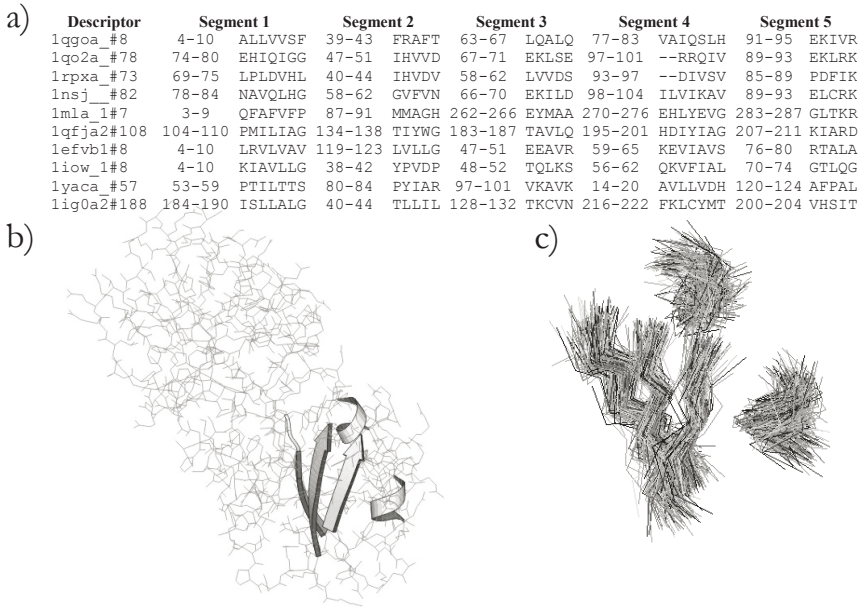


Fig. 2. An example of a local descriptor of protein structure (b), its structural neighbors (c) and the resulting sequence alignment (a)

a descriptive model. Moreover, the structure-function relationship in proteins was quantified by assessing the predictive performance of the model using cross validation and AUC analysis. The main conclusions that could be drawn from this study were as follows:

- A majority of the 113 molecular functions could be predicted with a statistically significant accuracy as assessed by a permutation study.
- GO molecular functions were better predicted than GO biological processes or GO cellular components.
- Combinations of local similarities allowed discerning proteins with different functions, but similar global structure (i.e. fold), e.g. the TIM barrel and the Rossmann fold.
- Catalytic activities were better predicted than most functions involving binding.
- Structure-based predictions complemented sequence-based predictions and also provided correct predictions when no significant sequence similarities to characterized proteins existed.

It has previously been observed that GO biological processes are better explained by expression data than are GO molecular functions [21,35]. This is intuitive, since genes participating in the same biological process need to be transcribed at the same point in time. However, it is interesting to observe in this study that molecular functions are better explained by specific structural

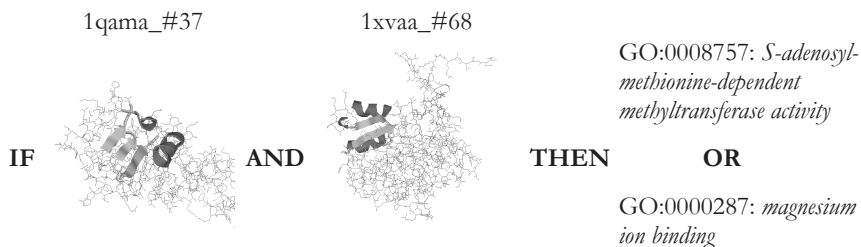


Fig. 3. The rule combines two substructure (i.e. 1qama_#37 and 1xvaa_#68) to describe 12 proteins annotated with GO:0008757. Two of these proteins are additionally annotated with GO:0000287.

shapes than are biological processes. Again, this is intuitive since a molecular function will require a protein to interact with a specific type of molecules, while proteins participating in the same biological process may interact with a wide variety of different molecules. Thus, this shows that different data is needed to predict different aspects of the molecular activity of proteins.

3.2 Protein – Ligand Interactions

An important goal of modern drug discovery is to develop computational models that can predict the interactions between drug targets (i.e. proteins) and ligands (i.e. drugs). A common approach in this research is QSAR (Quantitative Structure Activity Relationship), where the interaction between one protein and a series of ligands is modeled, and docking, where the three-dimensional structure of the protein is used to model the protein-ligand complex [51]. Proteochemometrics (PCM) [52] takes a different approach to molecular recognition in which the protein-ligand interaction space is modeled using series of *both* proteins and ligands. This approach greatly reduces the number of known interactions needed for modeling and may predict cross-interactions between drugs and other proteins in the proteome.

PCM uses machine learning methods to model the degree to which proteins and ligands interact (i.e. experimentally measured binding affinity) using chemical and structural descriptors to represent the proteins and the ligands. Strömbergsson *et al.* [53] used rough set-based rule learning to model interactions between G-Protein-coupled receptors (GPCR) and ligands. GPCRs are membrane-bound proteins that share a conserved structural topology of seven transmembrane helices. GPCRs are of particular interest, since about 50% of all recently launched drugs are targeted towards these receptors. The main novel result of this study was that rules allowed direct interpretation of the model, something that is not possible with the commonly used linear regression approach. For example, the rule model suggested that helix 2 was determinative for high and low binding affinity in three different data sets.

Previous approaches to PCM modeling have used protein descriptors that are calculated from a multiple alignment of the studied proteins. This limits

modeling to closely related proteins in terms of sequence or structure. In Strömbergsson *et al.* [54], the authors showed that using local descriptors of protein structure one can model vastly different proteins both in terms of sequence and structure. It was shown that the induced rule model combined local substructures and ligand descriptors to generalize beyond the enzyme-ligand interactions present in the training set. An interesting interpretation from the rules was that strongly bound enzyme-ligand complexes were described in terms of the presence of specific local substructures, while weakly bound complexes were described by the absence of certain local substructures. This is intuitive, since there may be only one or a few ligands that geometrically fit the active site of a specific enzyme and form a strongly bound complex, while there may be many ligands that only form weakly bound complexes with the same enzyme. The preferred description of the latter is to point to the absence of the local substructure that, if present, would have resulted in a strongly bound complex.

3.3 HIV-1 Modeling

The HIV virus has a high rate of replication leading to mutations and the development of drug resistance. The HIV-1 protease plays an essential role in replication by cleaving the viral precursor Gag and Gag-Pol polyproteins into structural and functional elements. For this reason the HIV-1 protease has been an attractive targets for the design of drugs that inhibit the protease and thus stop the replication of the HIV virus.

The HIV-1 protease cleaves the viral polyprotein by recognizing a sequence represented by four amino acids on each side of the actual position of cleavage. Kontijevskis *et al.* [55] collected all the experimental data on cleavable and non-cleavable sites from 16 years of HIV research (374 cleavable and 1251 non-cleavable substrates). Decision rules were induced based on the physico-chemical properties of the amino acids in these substrates, and cross validation demonstrated high predictive performance (accuracy and AUC well above 0.90). While previous studies based on less comprehensive data sets have revealed some patterns of limited predictive ability, analysis of this model showed that the rules encompassed properties from at least three substrate positions indicating a more complex relation than previously assumed. Nonetheless, as the cross validation evaluation showed, the rough set-based approach did recover general patterns determining HIV-1 protease cleavage specificity and several novel patterns were reported in the paper.

The HIV-1 reverse transcriptase (RT) transforms the viral RNA into DNA that can be incorporated into the genetic material of the host cell. Kierczak *et al.* [56] used rough set-based rule induction to predict drug resistance to six different drugs for a large number of mutated RTs. Existing biochemical knowledge related to the sequence and structure of RT was used to build descriptors from 19 known resistance-related positions. Cross validation accuracy and AUC values were in the ranges of 0.82-0.94 and 0.70-0.97, respectively, for the different drugs. As for the study of Kontijevskis *et al.*, the rules were pruned and inspection revealed general and novel patterns important for drug resistance.

4 Discussion

In this paper, we have reviewed a number of publications where rough set-based rule learning has been used to predict and describe molecular properties of biological systems. And we have seen how discrete representations and legible rules allow interpretations that gain new insight into molecular biology. The ability to describe data in terms of legible rules is particularly important in biology where biologists are interested in understanding the mechanisms underlying the data just as much as they are interested in the predictions themselves. Moreover, discrete representations add to this readability and allow the models to combine different heterogenous data sources containing both continuous and categorical data. Furthermore, the elegant representation of indeterministic data in terms of disjunctions of decisions in rules makes otherwise difficult problems, such as proteins annotated to several GO terms, easy to handle. Finally, developments in rough set-based rule induction such as dynamic reducts [57] and approximate reducts [16,58] allow the description of noisy data. We believe that it is due to these properties that rough sets now have gained a wide acceptance as a powerful tool for data analysis in life sciences. Pawlak's ideas were simple, yet powerful and rich enough to be of outstanding practical use in biology, and also continue to stimulate theoretical research in computer science.

Several challenges are particularly interesting in the context of rough sets and molecular biology. The first challenge is that of developing methods for illustrating and pruning rules [59,60] in order to allow interpretation. Some methods were reviewed in this article. In Dennis *et al.* [27], rules were represented as a decision tree, something which is very familiar to physicians. In Hvidsten *et al.* [35], predicted regulatory mechanisms with inconclusive evidence or low support were not considered. This is a simple yet powerful approach to rule filtering, but is also dangerous since potentially important discoveries may be lost [61,62]. Finally, the HIV-1 studies [55,56] used a group generalization method for rule pruning, where groups of rules with overlapping IF-parts and identical decisions are merged into generalized rules if the accuracy do not fall below a predefined threshold [60]. The second challenge is that of feature selection in order to avoid overfitting. In this review we saw a statistical method for selecting differentially expressed genes in a cancer classification study [22,23]. However, this procedure will exclude genes that individually are not significant, but that posses a significant discriminatory power in combination with other genes. Sampling methods such as random forests [63] might offer a solution to this problem in feature selection since they investigate the classification power of more than one gene at a time using subsets of all features. The third challenge is that of representing biological systems in a way that allows effective machine learning (i.e. feature synthesis). Examples discussed here were a template language for representing expression time profiles and local descriptors to represent protein structure. More than the development of computational methods themselves, we believe that the development of new ways to represent biological systems is the most important in order to successfully solve the puzzles of molecular biology. This also includes a final challenge, namely that of combining various sources of data in the

representation process such as, for example, using both molecular markers and clinical data in cancer classification. The significance of the second and third challenges was already recognized by A. Skowron in 1995 [64].

PubMed (<http://www.ncbi.nlm.nih.gov/PubMed>) is the main database providing access to all published biomedical literature. Searching for "rough set(s)" in titles and abstracts of articles in this database gives 69 hits since 1988 and reveal a large number of application areas beyond those described here. The true number of articles using rough sets in life sciences, however, is probably much higher since this search was limited to title and abstract and since only four of the papers reviewed in this article were retrieved by the search. Google Scholar (<http://scholar.google.com>), which searches through the whole text of all available scientific publications online, returned 290 articles with "rough set(s)" and "bioinformatics", and 11900 articles with "rough set(s)".

The published studies reviewed here all used the ROSETTA system, which is a user friendly, freely available software package for rough set-based rule induction and model evaluations [65] (<http://rosetta.lcb.uu.se/>).

Acknowledgements

We would like to thank co-authors of the reviewed articles for a stimulating collaboration. In particular Astrid Lægreid for her continuous help with all issues related to biology. The ROSETTA system has been an essential aid in this research. It was mainly developed by Alexander Øhrn under the supervision of Jan Komorowski, Trondheim, and in collaboration with Andrzej Skowron's group in Warsaw.

This research was supported by grants from the Knut and Alice Wallenberg Foundation (in part through the Wallenberg Consortium North), the Swedish Research Council, and the Swedish Foundation for Strategic Research.

References

1. Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M.: Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269** (1995) 496–512
2. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.: The protein data bank. *Nucleic Acids Research* **28** (2000) 235–242
3. Schena, M., Shalon, D., Davis, R., Brown, P.O.: Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science* **270** (1995) 467–470
4. Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P., Trent, J.M.: Expression profiling using cDNA microarrays. *Nat Genet* **21** (1999) 10–14
5. Patterson, S.D., Aebersold, R.H.: Proteomics: the first decade and beyond. *Nat Genet* **33 Suppl** (2003) 311–323
6. Kanehisa, M., Bork, P.: Bioinformatics in the post-sequence era. *Nat Genet* **33 Suppl** (2003) 305–310

7. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research* **25** (1997) 3389–3402
8. Shatkay, H., Feldman, R.: Mining the biomedical literature in the genomic era: an overview. *J Comput Biol* **10** (2003) 821–855
9. Jenssen, T.K., Lægreid, A., Komorowski, J., Hovig, E.: A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet* **28** (2001) 21–28
10. Brazma, A., Krestyaninova, M., Sarkans, U.: Standards for systems biology. *Nat Rev Genet* **7** (2006) 593–605
11. The Gene Ontology Consortium: Gene ontology: tool for the unification of biology. *Nature Genetics* **25** (2000) 25–29
12. Pawlak, Z.: Rough sets. *International Journal of Information and Computer Science* **11** (1982) 341–356
13. Pawlak, Z.: *Rough Sets: Theoretical Aspects of Reasoning about Data*. Volume 9 of Series D: System Theory, Knowledge Engineering and Problem Solving. Kluwer Academic Publishers, Dordrecht, The Netherlands (1991)
14. Komorowski, J., Pawlak, Z., Polkowski, L., Skowron, A.: Rough sets: A tutorial. In Pal, S.K., Skowron, A., eds.: *Rough Fuzzy Hybridization: A New Trend in Decision-Making*. Springer, Singapore (1999) 3–98
15. Skowron, A., Rauszer, C.: The discernibility matrices and functions in information systems. In Słowiński, R., ed.: *Intelligent Decision Support: Handbook of Applications and Advances in Rough Sets Theory*. Volume 11 of Series D: System Theory, Knowledge Engineering and Problem Solving. Kluwer Academic Publishers, Dordrecht, The Netherlands (1992) 331–362
16. Skowron, A., Nguyen, H.S.: Boolean reasoning scheme with some applications in data mining. In [66] 107–115
17. Churchill, G.A.: Fundamentals of experimental design for cDNA microarrays. *Nat Genet* **32 Suppl** (2002) 490–495
18. Quackenbush, J.: Microarray data normalization and transformation. *Nat Genet* **32 Suppl** (2002) 496–501
19. Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C.F., Trent, J.M., Staudt, L.M., Jr., J.D., Boguski, M.S., Lashkari, D., Shalon, D., Botstein, D., Brown, P.O.: The transcriptional program in the response of human fibroblasts to serum. *Science* **283** (1999) 83–87
20. Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., Lander, E.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286** (1999) 531–537
21. Brown, M.P.S., Grundy, W.N., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M., Haussler, D.: Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA* **97** (2000) 262–267
22. Midelfart, H., Komorowski, J., Nørsett, K., Yadetie, F., Sandvik, A., Lægreid, A.: Learning rough set classifiers from gene expression and clinical data. *Fundamenta Informaticae* **53** (2002) 155–183
23. Nørsett, K.G., Lægreid, A., Midelfart, H., Yadetie, F., Erlandsen, S.E., Falkmer, S., Grønbech, J.E., Waldum, H.L., Komorowski, J., Sandvik, A.K.: Gene expression based classification of gastric carcinoma. *Cancer Lett* **210** (2004) 227–237
24. Efron, B., Tibshirani, R.J.: *An Introduction to the Bootstrap*. Chapman & Hall, London (1993)

25. Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143** (1982) 29–36
26. Manley, B.F.J.: *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman & Hall (2002)
27. Dennis, J.L., Hvidsten, T.R., Wit, E.C., Komorowski, J., Bell, A.K., Downie, I., Mooney, J., Verbeke, C., Bellamy, C., Keith, W.N., Oien, K.A.: Markers of adenocarcinoma characteristic of the site of origin: Development of a diagnostic algorithm. *Clin Cancer Res* **11** (2005) 3766–3772
28. Hvidsten, T.R., Komorowski, J., Sandvik, A.K., Læg Reid, A.: Predicting gene function from gene expressions and ontologies. In Altman, R.B., Dunker, A.K., Hunter, L., Lauderdale, K., Klein, T.E., eds.: *Pacific Symposium on Biocomputing*, Mauna Lani, Hawai'i, World Scientific Publishing Co. (2001) 299–310
29. Hvidsten, T.R., Læg Reid, A., Komorowski, J.: Learning rule-based models of biological process from gene expression time profiles using gene ontology. *Bioinformatics* **19** (2003) 1116–1123
30. Læg Reid, A., Hvidsten, T.R., Midelfart, H., Komorowski, J., Sandvik, A.K.: Predicting gene ontology biological process from temporal gene expression patterns. *Genome Res* **13** (2003) 965–979
31. Eisen, M., Spellman, P., Brown, P., Botstein, D.: Cluster analysis and display of genome-wide expression pattern. *Proc. Natl. Acad. Sci. USA* **95** (1998) 14863–14868
32. Brown, P.O., Botstein, D.: Exploring the new world of the genome with DNA microarrays. *Nat Genet* **21** (1999) 33–37
33. Cho, R.J., Huang, M., Campbell, M.J., Dong, H., Steinmetz, L., Sapinoso, L., Hampton, G., Elledge, S.J., Davis, R.W., Lockhart, D.J.: Transcriptional regulation and function during the human cell cycle. *Nature Genetics* **27** (2001) 48–54
34. Pilpel, Y., Sudarsanam, P., Church, G.M.: Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature genetics* **29** (2001) 153–159
35. Hvidsten, T.R., Wilczyński, B., Kryshatovych, A., Tiuryn, J., Komorowski, J., Fidelis, K.: Discovering regulatory binding-site modules using rule-based learning. *Genome Res* **15** (2005) 856–866
36. Hughes, J.D., Estep, P.W., Tavazoie, S., Church, G.M.: Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J Mol Biol* **296** (2000) 1205–1214
37. Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K., Young, R.A.: Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298** (2002) 799–804
38. Wilczyński, B., Hvidsten, T.R., Kryshatovych, A., Tiuryn, J., Komorowski, J., Fidelis, K.: Using local gene expression similarities to discover regulatory binding site modules. Accepted in *BMC Bioinformatics* (2006)
39. Andersson, C.R., Hvidsten, T.R., Isaksson, A., Gustafsson, M.G., Komorowski, J.: Revealing cell cycle control by combining model-based detection of periodic expression with novel *cis*-regulatory descriptors. Submitted (2006)
40. Skolnick, J., Fetrow, J.S.: From genes to protein structure and function: Novel applications of computational approaches in the genomic era. *Trends Biotechnol* **18** (2000) 34–39

41. Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., Yeh, L.S.L.: UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **32** (2004) D115–D119
42. Chandonia, J.M., Brenner, S.E.: The impact of structural genomics: Expectations and outcomes. *Science* **311** (2006) 347–351
43. Tress, M., Ezkurdia, I., Graña, O., López, G., Valencia, A.: Assessment of predictions submitted for the CASP6 comparative modeling category. *Proteins* **61 Suppl 7** (2005) 27–45
44. Zhang, C., Kim, S.H.: Overview of structural genomics: from structure to function. *Curr Opin Chem Biol* **7** (2003) 28–32
45. Hvidsten, T.R., Kryshtafovych, A., Komorowski, J., Fidelis, K.: A novel approach to fold recognition using sequence-derived properties from sets of structurally similar local fragments of proteins. *Bioinformatics* **19 Suppl 2** (2003) II81–II91
46. Pazos, F., Sternberg, M.J.E.: Automated prediction of protein function and detection of functional sites from structure. *Proc Natl Acad Sci U S A* **101** (2004) 14754–14759
47. Orengo, C.A., Todd, A.E., Thornton, J.M.: From protein structure to function. *Curr Opin Struct Biol* **9** (1999) 374–382
48. Laskowski, R.A., Watson, J.D., Thornton, J.M.: ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res* **33** (2005) W89–W93
49. Pal, D., Eisenberg, D.: Inference of protein function from protein structure. *Structure* **13** (2005) 121–130
50. Hvidsten, T.R., Lægread, A., Kryshtafovych, A., Andersson, G., Fidelis, K., Komorowski, J.: High through-put protein function prediction using local substructures. Submitted (2006)
51. Terfloth, L.: Drug design. In Gasteiger, J., Engel, T., eds.: *Cheminformatics*. Wiley-VCH, Weinheim (2003) 497–618
52. Wikberg, J.E.S., Maris, L., Peteris, P.: Proteochemometrics: A tool for modelling the molecular interaction space. In Kubinyi, H., Mller, G., eds.: *Chemogenomics in Drug Discovery - A Medicinal Chemistry Perspective*. Wiley-VCH, Weinheim (2004) 289–309
53. Strömbergsson, H., Prusis, P., Midelfart, H., Lapinsh, M., Wikberg, J.E.S., Komorowski, J.: Rough set-based proteochemometrics modeling of G-protein-coupled receptor-ligand interactions. *Proteins* **63** (2006) 24–34
54. Strömbergsson, H., Kryshtafovych, A., Prusis, P., Fidelis, K., Wikberg, J.E.S., Komorowski, J., Hvidsten, T.R.: Generalized modeling of enzyme-ligand interactions using proteochemometrics and local protein substructures. Accepted to *Proteins* (2006)
55. Kontijevskis, A., Wikberg, J.E.S., Komorowski, J.: Computational proteomics analysis of HIV-1 protease interactome. Submitted (2006)
56. Kierczak, M., Rudnicki, W.R., Komorowski, J.: Construction of rough set-based classifiers for predicting HIV resistance to non-nucleoside reverse transcriptase inhibitors. Manuscript (2006)
57. Bazan, J.G., Skowron, A., Synak, P.: Dynamic reducts as a tool for extracting laws from decision tables. In: *Proc. International Symposium on Methodologies for Intelligent Systems*. Volume 869 of *Lecture Notes in Artificial Intelligence*., Springer-Verlag (1994) 346–355
58. Vinterbo, S., Øhrn, A.: Minimal approximate hitting sets and rule templates. *International Journal of Approximate Reasoning* **25** (2000) 123–143

59. Ågotnes, T., Komorowski, J., Løken, T.: Taming large rule models in rough set approaches. In [66] 193–203
60. Makosa, E.: Rule tuning. Master thesis. The Linnaeus Centre for Bioinformatics, Uppsala University (2005)
61. Düntsch, I.: Statistical evaluation of rough set dependency analysis. *Int. J. Human-Computer Studies* **46** (1997) 589–604
62. Düntsch, I., Gediga, G.: Uncertainty measures of rough set prediction. *Artificial Intelligence* **106** (1998) 109–137
63. Breiman, L.: Random forests. *Machine Learning* **45** (2001) 5–32
64. Skowron, A.: Synthesis of adaptive decision systems from experimental data. In Aamodt, A., Komorowski, J., eds.: *Fifth Scandinavian Conference on Artificial Intelligence*, Trondheim, Norway, IOS Press (1995) 220–238
65. Komorowski, J., Øhrn, A., Skowron, A.: ROSETTA rough sets. In Klösgen, W., Żytkow, J., eds.: *Handbook of Data Mining and Knowledge Discovery*. Oxford University Press (2002) 554–559
66. Żytkow, J.M., Rauch, J., eds.: *Proceedings of the Third European Symposium on Principles and Practice of Knowledge Discovery in Databases (PKDD'99)*. Volume 1704 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag, Prague, Czech Republic (1999)