HARD DATA ANALYTICS PROBLEMS MAKE FOR BETTER DATA ANALYSIS ALGORITHMS:

Bioinformatics as an Example

Jaume Bacardit, Paweł Widera,
Nicola Lazzarini, and Natalio Krasnogor

Interdisciplinary Computing and Complex BioSystems Research
Group, School of Computing Science, Newcastle University

Abstract

Data mining and knowledge discovery techniques have greatly progressed in the last decade. They are now able to handle larger and larger datasets, process heterogeneous information, integrate complex metadata, and extract and visualize new knowledge. Often these advances were driven by new challenges arising from real-world domains, with biology and biotechnology a prime source of diverse and hard (e.g., high volume, high throughput, high variety, and high noise) data analytics problems. The aim of this article is to show the broad spectrum of data mining tasks and challenges present in biological data, and how these challenges have driven us over the years to design new data mining and knowledge discovery procedures for biodata. This is illustrated with the help of two kinds of case studies. The first kind is focused on the field of protein structure prediction, where we have contributed in several areas: by designing, through regression, functions that can distinguish between good and bad models of a protein's predicted structure; by creating new measures to characterize aspects of a protein's structure associated with individual positions in a protein's sequence, measures containing information that might be useful for protein structure prediction; and by creating accurate estimators of these structural aspects. The second kind of case study is focused on omics data analytics, a class of biological data characterized for having extremely high dimensionalities. Our methods were able not only to generate very accurate classification models, but also to discover new biological knowledge that was later ratified by experimentalists. Finally, we describe several strategies to tightly integrate knowledge extraction and data mining in order to create a new class of biodata mining algorithms that can natively embrace the complexity of biological data, efficiently generate accurate information in the form of classification/regression models, and extract valuable new knowledge. Thus, a complete data-to-information-to-knowledge pipeline is presented.

Introduction

OUR CURRENT CAPACITY TO GENERATE high volumes of biological data through the use of experimental technologies is changing the way in which research is performed across all the biosciences.¹ The impact that these new technologies might have is, ultimately, limited by the ability to analyze and understand the vast volumes of data thus generated. While great progress has been achieved in what entails biodata analysis, most of the research effort tends to focus almost exclusively on the core data mining tasks (e.g., classification

and regression), and far less effort has gone into the crucial task of knowledge discovery itself. This constrains the process of uncovering new biological understanding that, after all, is the *ultimate goal* of data-driven biology.

For the last decade we have developed methodologies^{2–11} for the analysis of a range of biological data with knowledge discovery playing a central role in these works. The aim of this article is to show how bioinformatics has been a driver for innovation in data mining and knowledge discovery. We first describe in the section The Ever-Growing Spectrum of

Biological Data the large volumes and complexity of the existing biological data. Next, we illustrate this idea with two kinds of case studies of biological data.

The first one (in the section Case Study: Protein Structure Prediction) is focused on protein structure prediction (PSP), the task of estimating the three-dimensional (3D) shape of a protein from its amino acid sequence. State-of-the-art methods tackle this problem through sophisticated analysis pipelines that decompose the problem into several challenges involving classification and regression. We can characterize aspects of a protein's structure associated with individual positions of its amino acid sequence through a variety of measures, which in turn are estimated as classification/regression tasks. Moreover, regression is also used to design the so-called *energy functions* that can distinguish between good and bad models of a protein's predicted structure. We have (1) designed symbolic regression methods to create energy functions that are better than the state of the art (in the section Evolving Energy Functions for PSP Methods), (2) proposed new measures to characterize the parts of a protein's structure associated with individual positions of its sequence that have proven to contain useful information complementary to standard measures (in the section Predicting Structural Aspects of a Protein), and (3) created accurate classification models for several of these smaller challenges (section Protein Contact Map Prediction).

The second kind of case study (section Case Study: Analysis of Omics Data) is the analysis of omics data. Omics refers to a family of biotechnologies that can detect and measure a very large number of molecules or molecular states at once from a biological sample, and create datasets with extremely high dimensionality. We have created data mining and knowledge discovery methods that generate accurate classification models equal or better than the state of the art and that have been able to discover the functions of unknown genes, as proven by experimental validation in the biological lab.

Finally, we describe in the section Towards Knowledge Intensive Biodata Mining, our vision of the next generation of knowledge-intensive biodata mining methods, where knowledge discovery not only is one of the outcomes, but also becomes an integral part of both the mining process and the algorithmic development itself.

The Ever-Growing Spectrum of Biological Data

Data derived from living organisms are complex and heterogeneous, reflecting the multiple scales involved in the

characterization of living entities. The genome of an organism contains its blueprint in the form of DNA sequences, organized in smaller units called genes. The DNA sequence of a gene is transcribed (in a complex process called transcriptional regulation) into an RNA sequence, which later is translated into a protein (a sequence or chain of amino acids).¹² Moreover, proteins control chemical processes within a cell, such as metabolism and signal transduction, and are also used as the building block for structural material (e.g., collagen and hair). There are many more processes and subsystems in a living organism, but above we have illustrated four of the main ones, namely, its complement of DNA, RNA, protein, and metabolic processes. Some of these subsystems can be considered as static—for example, humans (and any other species) have a predefined genome size at birth, and each gene or protein will present a certain set of constant properties and functions—while other subsystems or processes are highly dynamic—for example, the specific RNA sequences and the amount of each of them present at any given point of time and in any given cell/tissue/organ of an organism differ from each other by time and location.

The amount of biological information stored in public databases about biological entities is enormous and ever growing. This information is sometimes the object of the mining process, while in other times it provides crucial domain knowledge that can enrich the process of mining other types of biological data. GenBank is the public database from the U.S. National Institutes of Health that stores the *static* information on gene sequences, and currently contains 173 million gene sequences. Uniprot (the world's largest database holding the *static* information on protein sequences) contains information on almost 80 million proteins. Both databases continue to grow exponentially.^{13,14} Moreover, in the last two decades many biotechnologies have been developed to measure and quantify the *dynamic* activity of living samples at a broad spectrum of scales, in what is globally known as omics technologies.^{1,15} The data generated by these technologies are generally suitable for being mined to extract new knowledge from them. Genomics technologies can read the whole genome from a biological sample (i.e., identify the DNA composition of the genome of an individual) or identify common genome variations across individuals of a species (e.g., single-nucleotide polymorphisms, SNPs). Transcriptomics technologies measure the abundance of RNA sequences, and proteomics technologies measure the abundance of proteins, to mention but a few of them.

The common trend across omics technologies is that they generate high-dimensional data. To give a couple of examples, a typical human microarray generates instances with

“BIOINFORMATICS HAS BEEN A DRIVER FOR INNOVATION IN DATA MINING AND KNOWLEDGE DISCOVERY.”

~50,000 attributes. An SNP or methylation chip generates instances with about half a million attributes. These technologies are (somehow) limited in the dimensionality of the samples they generate because they measure a predefined set of elements (be it RNA sequences, proteins, SNPs, etc.) and ignore anything else. Next-generation sequencing technologies, rather than looking for a predefined set of DNA/RNA sequences, read every sequence existing in the biological sample. Hence, the data they generate have a far larger dimensionality (potentially, as large as the size of the genome being sequenced) and size: The 1000 Genomes Project¹⁶ is an international consortium with the aim of thoroughly characterizing the human genome variation. That is, the variations in our DNA that differentiate individuals across the world. In its pilot stage, it performed next-generation sequencing of 742 individuals, producing 4.8 terabytes of *compressed* data just to store their genome sequences, and a total of 12.4 terabytes of compressed data when including the postprocessing of the raw sequence data. The project is currently in its main stage, sequencing the genome of new individuals with the target goal of reaching 2500 genomes. Its whole public data repository currently stores 579 terabytes of compressed data.

The data stored in biological databases are extremely diverse. DNA/RNA sequences use a four-letter alphabet of nucleotides, while protein sequences use the 20-letter alphabet of the natural amino acids. Most omics data are quantitative in nature (measuring the abundance of certain molecules). Entries in the Protein Data Bank describe the structure of proteins. That is, each entry stores the 3D coordinates of each atom that forms a protein. Moreover, bioimage analysis technologies (e.g., confocal microscopy and computed tomography) can capture 2D/3D images of a tissue, many times at/close to cellular resolution.

There is a rich and heterogeneous annotation (i.e., metadata) associated with most biological entities. This adds another layer of complexity to biological data. For instance, the Gene Ontology project currently stores annotation for 36 million protein sequences using a hierarchy of 41,000+ terms. The protein families (PFam) database catalogs 14,000+ protein families and their members, that is, groups of proteins that share a certain property (e.g., function). There are also databases of gene–protein interactions (with varying degrees of certainty) or about scientific publications related to biological data (PubMed), to mention just a few examples. In addition, the majority of these databases have cross-links between them. For instance, an entry in the Uniprot protein sequences database may link to the entry of the GenBank DNA database containing the sequence from which it is derived, a link to the

entry in Protein Data Bank that contains the 3D structure of that protein, and a link to the publication from PubMed that presented the experimental verification of this protein's existence.

Our aim in this section is to illustrate the fact that biological databases contain large volumes of data, broad varieties of sources, varying degrees of veracity of information, and high velocity of data generation (and data update). Hence, the major V's of big data¹⁷ are present. The case studies in the next two sections will show that this type of data is not only large and diverse, but also represents real challenges for the fields of data mining and knowledge discovery.

Case Study: Protein Structure Prediction

PSP refers to the problem of finding an algorithm that would be able to determine the 3D shape of a protein (the 3D coordinates of all the atoms in a protein) from its amino acid sequence. When a protein is constructed, its amino acid chain folds to create complex 3D shapes, and the function of a protein is (primarily) a consequence of its final folding state. With recent advances in large-scale DNA sequencing methods, the number of known protein sequences is reaching 80 million. At the same time, the experimental methods to determine a protein's 3D structure remain expensive and time-consuming, resulting in increasing gaps between known sequences and structures. Currently, the structure has been determined for less than 0.2% of proteins with known sequence.

Given that proteins, in their natural environment, mostly fold on their own as they are being constructed,¹² the general consensus of the PSP community is that the sequence of a protein should contain enough information to predict its structure. This problem is the longest standing open challenge in computational biology, because knowing the structure of a protein is a crucial step for understanding its function. In fact, the Nobel Prize for chemistry in 2013 was awarded (among others) to Michael Lewitt

for his pioneering work in this field. PSP is not a single estimation problem but rather a family of related problems, which in many cases can be treated as classification or regression tasks. In the next subsections, we describe several of them.

Evolving energy functions for PSP methods

Problem description. Energy functions are at the core of all major 3D PSP methods. The problem of generating complete 3D models of a protein's structure is usually treated as an

“...BIOLOGICAL DATABASES
CONTAIN LARGE VOLUMES
OF DATA, BROAD VARIETIES
OF SOURCES, VARYING
DEGREES OF VERACITY
OF INFORMATION,
AND HIGH VELOCITY
OF DATA GENERATION...”

(extremely complex and costly) optimization problem. Top methods (e.g., Rosetta) can dedicate tens of thousands of central processing unit (CPU) days (through collaborative computing environments such as Rosetta@home and more recently FoldIt¹⁸) to generate a model of the structure of just a single protein. The most advanced 3D structure modeling methods attempt to mimic the protein folding process from scratch using physical principles. They perform a stochastic search over the vast space of possible solutions by iteratively applying torsions and rotations to the amino acid chain and produce a set of candidate models. The best models from that set are then selected based on their thermodynamic stability, which is approximated using *energy functions*. The key elements to a successful prediction are (1) an accurate energy function that is able to relate a low energy stable state of a model, to the correctly folded native protein state, and (2) the ability to select native-like models from the set of candidates. The classic approach to the design of energy functions¹⁹ assumes that the total energy is a linear combination of energy terms and focuses on the optimization of the weights assigned to each term.

Data mining formulation and challenge. Given (1) a set of known, correctly folded protein structures, (2) a set of candidate 3D structure models generated for each protein, and (3) a candidate set of energy terms, we try to find an energy function that would rank the models by the similarity to the native state. This could be seen as a regression problem, where an energy function is constructed as a combination of energy terms. The main challenge lies in the size of the training set used to generate the energy functions, which can contain millions of models and close to 100 energy terms. Additional difficulty arises from the fact that models vary greatly in quality. As extremely good/bad models are much less frequent than those “in between,” the training set is unbalanced.

Our approach. In contrast to the classic approach, we assumed that an energy function could be any combination of energy terms, not necessarily a linear one. We solved this symbolic regression problem using a genetic programming algorithm, which generates tree-like functions with arithmetic and trigonometric operations of (potentially) arbitrary complexity.⁵ Our objective function minimized the distance between the ranks of models assigned by the evolved energy function and the reference ranks based on the similarity to the native state. As distance function between ranks we used the Spearman distance coupled with a sigmoid function to apply larger penalties to errors in low (better) positions of the rank. Moreover, we constrained the depth of the genetic programming trees to control overfitting. The experiments

were executed in the high performance computing (HPC) environment and took around 100 CPU days. We compared our method to the optimized linear combination of terms used at that time by the top prediction methods (using the simplex optimization method).

Table 1 reports the results for both methods, trained on two sets of models, (1) all models and (2) a small subset of models (d-100), selected by sampling uniformly the whole range of similarities to the real structure in order to reduce noise and increase balance in the data. In addition to the objective function values (similarity of ranks), we report the correlation between the energy function and the similarity of the models to the real structure. In the end, our method was able to outperform the linear method by over 10% on

both measures. Given that the symbolic expressions used in the energy functions were human-understandable, we were able to analyze the best evolved functions, which would not be possible with a black-box method such as a neural network. We found that the algorithm was able to independently discover the most/least useful energy terms. Given a larger training set and refined noise reduction strategy, it could provide more insight into the relevance of individual energy terms, which in

consequence could lead to improvements in their design.

Predicting structural aspects of a protein

Problem description. A very broad range of structural aspects of a protein associated with individual positions of a protein chain can be quantified. The most well-known structural aspect is its *secondary structure*. The secondary structure of a protein is a set of recurrent structural patterns that the amino acids of a protein undertake due to interactions with their immediate neighbors in the chain. Each amino acid is characterized as belonging to a finite set of secondary

“THE EVOLVED ENERGY FUNCTIONS WERE HUMAN-UNDERSTANDABLE, AND WE WERE ABLE TO ANALYZE THEM, WHICH WOULD NOT BE POSSIBLE WITH A BLACK BOX METHOD.”

TABLE 1. COMPARISON BETWEEN THE PERFORMANCE OF THE ENERGY FUNCTIONS GENERATED BY OUR METHOD (GP) AND A LINEAR COMBINATION OF ENERGY TERMS (NELDER-MEAD DOWNHILL SIMPLEX METHOD)⁵

Method	Spearman-sigmoid		Correlation	
	d-100	All	d-100	All
Simplex	0.734	0.638	0.65	0.166
GP	0.835	0.714	0.74	0.2

All, complete training set; correlation, Pearson's correlation coefficient between the scores given by the energy function and the similarity of a model to the real protein structure; d-100, subset of the training set that sampled uniformly the whole range of similarities to the real structure; Spearman-sigmoid, objective function in the GP method. Reproduced from Widera et al.'s (5) Table 6 with permission from Springer Science+Business Media.

structure states. Other typical structural aspects are the *solvent accessibility* of an amino acid, the area of an amino acid's surface that is accessible from the outside of a protein; the *residue depth*, the distance between a given amino acid and the surface of the protein; and the *contact number*, the number of other amino acids that are within a given distance of an amino acid in the protein structure. These structural aspects can be estimated (as classification/regression tasks) from a protein's amino acid sequence, and these estimations become stepping stones toward solving the overall PSP problem by applying a divide-and-conquer strategy.

Data mining formulation and challenge. Training sets for the prediction of structural aspects of amino acids can easily reach sizes of almost a million instances (individual amino acids) taken from a few thousand different proteins for which we experimentally know their structure. Moreover, the state-of-the-art representations for these problems contain hundreds of attributes. We treat these datasets as regression problems (when a structural aspect is continuous, e.g., the solvent accessibility) or as classification problems (when the

structural aspect is discrete [e.g., the secondary structure] or it has been discretized).

Our approach. The definition of the typical structural aspects of a protein (described above) is motivated by the chemical properties of a protein (which, after all, is a molecule formed by chaining together amino acids). Mathematics offers many alternative ways of characterizing a protein's structure from an abstract perspective. Borrowing from the fields of geometry and topology, we have defined novel structural aspects of protein amino acids that are able to capture information that is complementary to the standard ones. Figure 1 represents two classes of these structural aspects. Panel A shows our recursive convex hull measure,² which borrows the formal geometry concept of convex hull²⁰ and applies it to identify the layers of amino acids in a protein structure, hence providing a new way of quantifying whether an amino acid is at/near the core of a protein or at/near its surface. Panel B represents four structural aspects that belong to a family of topology measures called proximity graphs.³ The proximity graphs provide a formal way of identifying who are the *structural neighbors* (i.e., neighbors in the 3D

FIG. 1. Two measures to represent structural aspects of protein amino acids based on geometry (A) and topology (B). (A) Recursive convex hull measure, represented in 2D in an idealized representation and in 3D for real protein 1MUW. Each color identifies a layer of amino acids formally defined as a convex hull of points. The aim of this measure is to quantify the degree of burial of a given amino acid within the 3D structure of a protein where it belongs. Reproduced from Stout et al.² by permission of Oxford University Press. (B) Proximity graphs family of topological structural aspects, represented for real protein 153L. Each amino acid in a protein is represented as a vertex in the graph. Edges connect the amino acids considered to be *structural neighbors* (in the 3D space) by each of the four topology measures. DT, Delanuy tessellation; GG, Gabriel graph; MST, minimum spanning tree; RNG, relative neighborhood graph. Reproduced from Stout et al.³ with permission from Springer Science and Business Media.

space, instead of neighbors in the protein chain) of an amino acid within the protein structure. These measures are complementary to the typical procedure of defining an *arbitrary* distance threshold, and declaring all amino acids within this distance in the protein's structure as neighbors, as described above for the *contact number* structural aspect.

In our experiments we generated classification models to estimate these structural aspects from the protein sequence. The levels of accuracy obtained vary across structural aspects and depending on the sophistication of the data representation, ranging from 62% to 80% accuracy for the 2-class formulation of the structural aspects.²⁻⁴ In addition, we demonstrated that the estimation of these structural aspects could be fed back to improve the prediction of other structural aspects. We report in Table 2 the use of the estimations of a variety of structural aspects of protein residues (two variants of recursive convex hull, solvent accessibility, residue depth, and exposure) as extra input attributes to estimate another structural aspect, contact number.² The basic contact number classification model had an accuracy of 77.2%. When the estimated recursive convex hull was added to its problem representation, the accuracy increased to 78.5%. When both recursive convex hull and solvent accessibility were used, the accuracy of predicting the contact number increased to 79.7%, that is, an

accuracy level of 2.5% over the basic contact number classification model. These results indicate that our novel measures of structural aspects contain useful information complementary to that of standard structural aspects, and that

the general strategy of divide and conquer used in PSP is effective. Moreover, given that the training sets generated to create the models for these structural aspects represent a broad range of (relatively large) classification and regression datasets with a variety of types and sizes of attributes and classes, they are ideal benchmarks to test the scalability of data mining methods. To this aim we have set a public repository of 140 different PSP datasets with

varying characteristics.²¹

“OUR NOVEL MEASURES OF STRUCTURAL ASPECTS CONTAIN USEFUL INFORMATION COMPLEMENTARY TO THAT OF STANDARD MEASURES.”

Protein contact map prediction

Problem description. Two amino acids of a protein are defined as being in contact if the distance between them in the 3D structure of the protein is less than a certain threshold. A contact map (CM) is a square binary matrix indicating whether every possible pair of amino acids in a protein is in contact or not. A CM is a simplified 2D profile of the 3D protein's structure.

Data mining formulation and challenge. The prediction of a protein's CM is an extremely challenging classification task for several reasons. First, training sets contain every possible pair of amino acids in a protein. As a result, the size of the training set grows quadratically in relation to the number of amino acids in a protein, and can easily reach tens of millions of instances (pairs of amino acids) by just using a few thousand proteins with known structure for training. Second, a CM is a very *sparse* matrix, and the number of real contacts generally is around 2% of the total number of possible amino acid pairs. Hence, this is a classification problem with an extremely high class imbalance. Finally, state-of-the-art knowledge representations for CM prediction are the fusion of complex and heterogeneous sources of information, including the prediction of several of the structural aspects of protein amino acids described in the section Predicting Structural Aspects of a Protein, and easily reach hundreds (if not thousands) of attributes.

Our solution. We have created our own CM prediction method,⁹ for which we generated a training set of 32 million instances and 631 attributes (taking 56GB of disk space). The 631-attribute problem representation we designed was the fusion of 9 different sources of information derived or estimated from the amino acid sequence of a protein. As data mining tools, we used our own highly scalable rule-based data mining method called BioHEL,^{22,23} generating an ensemble of

TABLE 2. USING THE ESTIMATION OF A VARIETY OF STRUCTURAL ASPECTS OF PROTEIN AMINO ACIDS AS INPUT ATTRIBUTES TO HELP ESTIMATE BETTER THE CONTACT NUMBER STRUCTURAL ASPECT²

Data representation	Accuracy
CN	77.2 ± 0.8
CN + RD	77.4 ± 0.8
CN + RCHr	77.6 ± 0.7
CN + Exp	77.7 ± 0.8
CN + Exp + RCHr	77.7 ± 0.7
CN + RCH	78.5 ± 0.9
CN + RCH + RCHr	78.8 ± 0.7
CN + Exp + RCH	78.9 ± 0.8
CN + SA	78.9 ± 0.8
CN + Exp + RCH + RCHr	78.9 ± 0.7
CN + Exp + SA	79.1 ± 0.8
CN + Exp + SA + RCHr	79.1 ± 0.8
CN + SA + RCHr	79.1 ± 0.8
CN + SA + RCH	79.7 ± 0.8
CN + Exp + SA + RCH	79.8 ± 0.8
CN + SA + RCH + RCHr	79.8 ± 0.8
CN + Exp + SA + RCH + RCHr	79.8 ± 0.7

The BioHEL rule-based machine learning method was used for all experiments in this table.²¹ CN, contact number; Exp, exposure; RCH, recursive convex hull; RCHr, reverse recursive convex hull; RD, residue depth; SA, solvent accessibility. Reproduced from Stout et al.² by permission of Oxford University Press.

rule sets trained from several samples of the training set where the class imbalance has been corrected. The whole training process took around 25,000 CPU hours.

Our CM predictor participated in the 9th edition of the bi-annual Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiment, the main competition of the PSP field. From all the proteins used in the CASP competition, the *free modeling* subset (proteins for which no other protein with a similar structure is known so far) was used to evaluate CM methods. The evaluation system used in CASP is quite ad-hoc: for each protein that is used in the competition, participants are asked to submit only the predicted contacts (the minority class with about 2% of the instances in the dataset). For each predicted contact, a confidence estimate (between 0 and 1) has to be submitted. Next, the predicted contacts of each protein are sorted by confidence, and only the top N predictions are used to compute the score metrics, hence varying a confidence threshold to accept predicted contacts. This N is proportional to the length of the protein (number of amino acids it contains). Three values are used: $N=L/10$, where L is the length of the protein; $N=L/5$; and $N=5$ (top 5 predictions per protein). Two metrics are computed for each protein, precision (true positives/[true positives+false positives]) as well as XD, a metric specific to CASP that assesses how better than random the predictions are. We summarize in Table 3 the performance of the 12 methods that participated in the CM prediction category of CASP9.

We report, for each metric and confidence threshold, the average score and average rank (for each protein, methods

were ranked from 1 to 12, and then these ranks were averaged) across the *free modeling* proteins. Our predictor obtained an average precision ranging from 21.1% to 25.7% across different confidence thresholds. This figure may look very low, but it just reflects the difficulty of the problem: our method was ranked as the best *ab initio* (i.e., a method explicitly designed to predict contacts, rather than deriving contact predictions from models of the 3D structure) CM predictor^{9,24} and third best method overall (although there is a considerable gap with the first two methods). Our rule-based ensemble method outperformed several other participants in the competition using a variety of machine learning methods such as neural networks, support vector machines, and random forests. Furthermore, the reason that we opted for a rule-based machine learning method was to improve knowledge discovery. After the whole training process, we analyzed the ensemble of rule sets generated by BioHEL, and because we had a *white-box* classification model that we could understand, we were able to numerically estimate the relevance of all the sources of information⁹ used in the CM prediction knowledge representation, and identify the interactions across attributes.

Figure 2 visually represents the set of attributes used in the CM representation of our CASP9 predictor and the new CASP11 predictor that is currently (Spring 2014) taking part in the experiment. Individual attributes are represented as bubbles. Colors identify sources of information, and the size of each bubble represents the relevance of the attribute within the classification model. This visualization technique allows us to represent datasets of high dimensionality and

TABLE 3. PERFORMANCE OF THE CONTACT MAP PREDICTION METHODS THAT PARTICIPATED IN THE CASP9 COMPETITION USING TWO PERFORMANCE METRICS (PRECISION AND XD) AND THREE CONFIDENCE THRESHOLDS (L/5, L/10, AND TOP5)⁹

Team ID	Method	Precision L/5		Precision L/10		Precision Top5		XD L/5		XD L/10		XD Top5	
		Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank
RR490	3D PSP	33.8±24.5	4.30 (2)	38.6±28.0	4.78 (2)	46.7±31.7	4.59 (2)	15.1±8.0	3.96 (2)	17.5±8.5	3.98 (2)	19.9±9.4	3.96 (2)
RR391	3D PSP	32.8±25.0	3.74 (1)	37.2±28.3	4.30 (1)	45.9±33.0	3.91 (1)	15.3±8.4	3.28 (1)	17.3±9.4	3.28 (1)	19.6±10.5	3.35 (1)
RR051	Rule-based	21.1±13.3	5.87 (3)	24.1±16.4	5.98 (3)	25.7±23.2	6.11 (4)	10.6±5.3	5.33 (3)	11.7±7.1	5.63 (3)	11.8±9.0	5.96 (3)
RR002	SVM	21.0±16.0	6.15 (4)	24.3±21.7	6.48 (6)	23.0±25.4	7.43 (10)	10.2±7.4	6.07 (4)	11.6±8.4	6.07 (4)	12.0±9.1	6.5 (7)
RR138	Random forest	20.6±13.9	6.24 (5)	23.3±21.1	6.39 (5)	27.7±24.2	6.15 (5)	10.5±5.3	6.33 (5)	11.5±6.8	6.30 (6)	12.4±7.6	6.32 (6)
RR103	Neural network	20.1±16.2	6.35 (6)	26.9±25.7	6.04 (4)	31.4±34.0	5.78 (3)	9.1±5.6	6.96 (7)	11.1±8.2	6.11 (5)	11.9±10.7	6.15 (4)
RR375	Random forest	19.3±14.5	7.09 (8)	21.9±17.5	6.54 (7)	24.6±21.7	6.22 (6)	10.1±6.0	6.46 (6)	9.9±7.3	6.41 (7)	10.5±8.9	6.22 (5)
RR244	Neural network	18.8±15.0	6.41 (7)	21.5±18.3	6.70 (8)	22.3±23.1	7.11 (7)	9.0±5.8	7.39 (9)	10.0±6.7	7.26 (9)	10.2±8.4	7.43 (10)
RR422	Not specified	16.4±13.7	7.61 (9)	18.3±17.8	7.11 (9)	18.6±19.9	7.09 (8)	7.9±5.2	7.30 (8)	8.6±7.3	7.2 (8)	8.4±7.9	7.41 (9)
RR119	Neural network	15.6±17.3	8.13 (11)	19.0±22.2	7.26 (10)	20.0±24.8	7.26 (9)	7.5±6.9	8.13 (10)	8.6±8.2	7.8 (10)	9.6±8.5	7.09 (8)
RR080	Neural network	14.8±17.4	8.30 (12)	16.7±22.8	8.30 (12)	15.4±25.0	8.26 (12)	7.3±7.2	8.39 (11)	7.7±8.7	8.57 (11)	7.1±9.6	8.5 (11)
RR214	Neural network	14.0±13.2	7.80 (10)	12.1±13.4	8.11 (11)	11.4±16.4	8.09 (11)	6.0±5.7	8.41 (12)	5.0±6.1	9.39 (12)	4.6±7.2	9.1 (12)

Participating teams are identified by their team ID. The row corresponding to our team is marked in bold. For each metric and confidence threshold, we report the average score (higher=better) and average rank of score (lower=better) across the proteins used for test. A brief description of the machine learning method used by each team is provided.

FIG. 2. Structure of our contact map predictors for the CASP9 (left) and CASP11 (right) experiments. Each bubble represents one attribute in the representation of contact map classification problem, and the size of the bubble indicates its importance in our classification method. Color identifies the source of information. The difference between the CASP9 and CASP11 predictors is the addition of two attributes (top right in light green) that produce a noticeable change in the shape of the overall predictor.

heterogeneous structure. The difference between the CASP9 and CASP11 predictors is the addition of two new attributes (*correlation_dca* and *correlation_mutual*, represented in light green in the top-right part of Fig. 2) that implement the latest step change in CM prediction.²⁵ Even though only two attributes were added, there is a visually noticeable change in the overall structure of the CASP11 predictor, as the two new attributes also affect the rest (the size of some of the major bubbles has changed). Moreover, the impact is not just visual: the accuracy of our method increased by $\sim 6\%$ based on our own experiments (pending the official CASP11 assessment).

Furthermore, we have recently used this dataset as a generic big data benchmark. During the spring/summer of 2014, we organized a big data competition²⁶ within the First Evolutionary Computation for Big Data and Big Learning Workshop. The competition worked in a similar way to the Kaggle competitions: A training set and a test set without labels were released, and we set up a web-based infrastructure where the participants could submit the classification outcome for the test instances. A 2-month window to submit solutions was established. Participants could see their progress in the competition's leaderboard. The product of the true-positive rate and true-negative rate was used as performance score, in order to force methods to focus on the (very minority) positive examples.

Seven teams participated in the competition, with a total of 364 submissions through the 2 months of competition. The succession of submissions through the competition reflected the progressive improvement in performance of the different teams. Among the participants, a variety of classification algorithms were used: random forests, ensembles of rule sets, a deep learning architecture, linear genetic programming, and support vector machines. Moreover, among the computational infrastructure used by the teams, there were Hadoop clusters, classic HPC clusters using both batch and parallel solutions, GPUs, and so on. We would like to remark, however, that this competition is not directly comparable to the CASP competition described above for two reasons: (1) The main challenge in CASP is to *engineer* a problem representation containing useful information for predicting contacts. In here all teams were provided with the same dataset, and so the representation was fixed and teams had to focus only on model training. (2) The performance metrics used to assess participants in each competition are very different from each other.

Case Study: Analysis of Omics Data

Problem description

Omics data are produced by biotechnologies that can measure, at once, a very large number of biological elements (e.g.,

DNA, RNA, proteins, and lipids) from a biological sample, and hence are able to generate data records that are characterized by a very large number of attributes.

Data mining formulation and challenge

When samples belong to categories (e.g., cases vs. controls), the problem can be treated straightaway as a classification problem. When samples are generated from the same organisms at different time points, the role of the data mining is to identify clusters of temporal patterns in the data. When there is no annotation in the samples, the goal of the mining process is to identify clusters of distinct phenotypes. Generally, the knowledge discovery aim of analyzing omics data generally falls into two categories: identifying biomarkers (individual or small groups of attributes that are important to characterize the samples) and inferring networks of interactions from the data. However, due to the high cost of generating each record, most omics datasets have far lower number of instances than number of attributes. As a result, feature selection is an important aspect in essentially all omics data analysis in the literature.²⁷

Our solution

We have used rule-based machine learning methods to both identify biomarkers and infer networks of interactions.^{6–8,10,11} From an omics dataset, we first generate an ensemble containing a large number of rule sets (e.g., using the BioHEL

method described above for the CM prediction challenge). BioHEL is enhanced for this type of analysis with a sparse knowledge representation containing a rule-wise embedded feature selection:²² each rule being generated by our method will only use a very small fraction of attributes. The relevant attributes for that particular rule are discovered automatically during the learning process (hence being an embedded feature selection), and each different rule may use a different subset of attributes (hence being rule-wise).

In our experiments we have used this representation to learn directly from microarray data containing tens of thousands of attributes without needing to perform a global feature selection before the training process, hence avoiding the potential information loss from prefiltering the data before learning. Next,

the attributes and pairs of attributes most frequently used within rules of the ensemble are identified. Variables appear in rules because they have high classification power *in combination* with the rest of attributes in the same rule. Thus, this method provides a multivariate approach to ranking the attributes of an omics dataset. Moreover, by identifying pairs of attributes often appearing together in rules, interactions between attributes can be inferred. Using this principle, we can build interaction networks. We call this network inference method co-prediction:⁶ we use this name in contrast with the most typical method of inferring networks from biological data, called co-expression.²⁸

“CO-EXPRESSION NETWORKS APPLY WHAT IS CALLED AS GUILT-BY-ASSOCIATION PRINCIPLE, WHILE OUR CO-PREDICTION METHOD CONNECTS ATTRIBUTES THAT ACT TOGETHER WITHIN OUR RULE SETS TO CLASSIFY SAMPLES.”

FIG. 3. (A) Simplified representation of the co-prediction principle for network inference from rule-based machine learning. (B) Co-prediction network generated from a plant seed transcriptomics dataset.

TABLE 4. ACCURACY OF BioHEL, SUPPORT VECTOR MACHINES, NAÏVE BAYES, AND C4.5 IN A PLANT SEED TRANSCRIPTOMICS DATASET⁶

Method	Accuracy
BioHEL	92.8% ± 1.6%
SVM	82.4% ± 0.4%
Naïve Bayes	88.0% ± 0.4%
C4.5	79.8% ± 3.6%

Accuracy estimated using 10 × 10-fold cross-validation.

In co-expression networks, two attributes are connected if their profile of values across the instances of the dataset is similar (using, e.g., Pearson's correlation to measure similarity). Hence, they apply what is called as guilt-by-association principle, while our co-prediction method connects attributes that act together within our rule sets to classify samples. Afterward, we apply statistical permutation tests²⁹ to filter out the attributes appearing just by random chance. Our co-prediction network inference methodology is illustrated in Figure 3, which also shows the network generated from a plant seed microarray dataset⁶ to functionally represent the process of seed germination. In this dataset we had seed samples for which we knew if they had germinated or not. We first built a rule-based classification model that was accurately estimating the germination outcome of the samples. We compared our method against standard machine learning methods.

Table 4 reports the results. Our method obtained close to 93% accuracy, way ahead of other machine learning methods (e.g., Naïve Bayes obtained an 88% accuracy, and support vector machines [SVMs] an 82%). Moreover, Table 5 reports the confusion matrix for our method on this dataset. Our predictor is slightly better for germination samples than for nongermination ones, which happen to be the minority class in the dataset (but without much class imbalance, it has 43% of instances). Afterward we used our knowledge extraction techniques to generate rankings of important variables, where many genes known from the literature to be important in the germination process were highly ranked. Finally, experi-

mental validation in the lab (by generating and planting seeds where the genes under investigation had been disabled) revealed that our rule-based biomarker discovery methodology was able to identify that four genes that until then had unknown function were indeed influential in the seed germination process. We have also successfully applied this methodology to generate accurate classification models and identify biomarkers from cancer transcriptomics datasets,⁷ piglet lipidomics data as a model of early development,⁸ and canine proteomics data^{10,11} as a model of human osteoarthritis. In all these cases, the biomarkers identified by our method were well supported by the literature. Other researchers have applied similar techniques (in terms of both rule-based machine learning and knowledge extraction) to synthetic and real human SNP data.^{29–31}

“TREATING THE KNOWLEDGE GENERATION ONLY AS AN OUTPUT OF THE MINING PROCESS JUST SCRATCHES THE SURFACE OF ITS POTENTIAL.”

Toward Knowledge Intensive Biodata Mining

Through these four case studies, we have illustrated the central role that data mining has in the analysis of biological data and the variety of challenging mining tasks that such analysis requires. We also demonstrated the use of white-box data mining techniques such as rule-based machine learning or genetic

programming, coupled with sophisticated information visualization techniques, to discover new knowledge from the data. However, treating the knowledge generation only as an output of the mining process just scratches the surface of its potential. There are many avenues that can make knowledge extraction a central player in the mining of biological data, and lead to the next generation of knowledge-intensive biodata mining methods.

For instance, given that rich and heterogeneous metadata about biological datasets exist in public databases, it is a missing opportunity not to fully integrate such metadata into the mining process. Examples exist that perform feature selection from text mining³² or from network data,³³ hence just as a preprocessing step. There are many opportunities for embedding the domain knowledge much deeper within the mining process: for instance, by creating template models from domain data before learning, by biasing the exploration process during learning, or by refining final models after learning. Strategies such as this are just starting to get explored.³⁰

Moreover, most biodata mining methods use general-purpose knowledge representations that do not perform any assumption about the data. Can we optimize the representations to improve knowledge discovery when faced with the particular case of mining biological data? In our work

TABLE 5. CONFUSION MATRIX FOR BioHEL IN A PLANT SEED TRANSCRIPTOMICS DATASET⁶

Belonging to	Predicted as	
	Germination	Nongermination
Germination	98.1%	1.9%
Nongermination	11.3%	88.7%

FIG. 4. Tree map representation of the structure of our CASP9 rule-based contact map predictor. Each box represents an attribute, and the size of the box indicates the attribute's relevance. Color of the box identifies the source of information. Attributes that participate in the rules being activated for a specific protein are connected with edges. Red edges: activations for a particular class of protein structure, an all-alpha protein (code 1ECA). Blue edges: activations for another particular class of protein structure, an all-beta protein (code 1CD8). An edge bundling technique is used to visualize the resulting network. All rules classify instances as belonging to the same class (contact), but using different strategies depending on the type of protein.

analyzing plant seed omics data,⁶ we introduced a very subtle change in the structure of the rules allowed in our knowledge representation: our rule representation follows a classic “*If predicate then classify instance as class C.*” Where *predicate* is a conjunction of tests, each of them associated with an attribute in the problem. In our standard representation, each test (for continuous attributes) would have a general shape “ $X < \text{Attribute} < Y$.” By removing the “ $< Y$ ” part of our representation, we forced the rule learning algorithm to focus on predicates such as “ $\text{Attribute1} > X$ **and** $\text{Attribute2} > Y$.” In general biological terms, when such a predicate appears, we can infer that we have identified the co-occurrence of up-regulated genes. In the particular case of our work in plant seed omics data, this change in knowledge representation was able to identify biomarkers that were well supported by both the literature and, especially, experimental validation (as mentioned above). There is a lot of scope to create bio-knowledge representations that are designed (initially designed by hand, but ideally created through a data-driven process) to capture specific patterns of biological data, for instance, by using grammars to express patterns.³⁴

Finally, can we feed all the discovered knowledge back to the mining algorithm to help improve the learning process? For

instance, there are examples of rule-based machine learning methods applied to SNP data, where the attribute relevance is estimated from the rules and then this information is used to bias the exploration process of learning new rules.³¹ This type of knowledge feedback into the learning process can be refined much further if, rather than focusing on individual attributes, we focus on their interactions.³⁵ There are many possible strategies to feed interaction knowledge back into the mining process. In Figure 4 we illustrate one possible strategy aimed at integrating human experts in the refinement process. Figure 4 represents (using a tree map visualization) the structure of our CASP9 CM predictor, described in the section Protein Contact Map Prediction. Larger boxes represent more relevant attributes. On top of the tree map, we show the co-prediction network of this dataset, but we only display the subset of the network that is activated to classify certain particular instances: red edges represent the activations for an all-alpha protein, while the blue edges represent the activation pattern for an all-beta protein.

The technique we use to display the network (force-directed edge bundling)³⁶ will collapse the parts more frequently used (in essence the backbone of the network/classifier) but will display in more detail the less frequently used parts. The

colors of the edges allow us to see that for both types of instances there is a common part of the classification model being activated (where blue and red overlap) but also several areas where only blue or red appears. This shows how different types of data (classes of protein structures) activate different parts of our predictor. Hence, by joining together knowledge extraction and sophisticated information visualization, (1) we are able to visually explain how our classification models work for very complex and heterogeneous data, where there may be multiple explanations for the same concept (in this particular case, a contact between amino acids), and, beyond this, (2) we are able to identify how to guide the data mining process and adjust our models to learn better on specific types of data.

Conclusions

This article illustrates that hard biological data analytic problems are an excellent test bed for data mining methods due to their sheer size, diversity, and rich metadata. Through the case studies we have shown how these problems motivated many innovations in data mining and knowledge discovery performed at our research group, and helped us perform concrete contributions to these application domains, such as better methods for generating energy functions for PSP, the definition of novel and useful aspects of a protein's structure, CM predictors scoring very high in competitions, and identifying new biomarkers for several application domains. Finally, we have made a case for tightly integrating knowledge discovery into the core of the mining process. We have described three types of integration strategies by (1) using domain knowledge throughout the whole process of learning, (2) designing knowledge representations that capture specific patterns of biological interactions, and (3) feeding back the extracted knowledge into the learning process. In combination, this strategy allows us to (1) improve knowledge extraction, (2) improve the efficiency of the mining methods, and (3) produce better classification and regression models.

Acknowledgments

We would like to acknowledge the support of the UK Engineering and Sciences Research Council (EPSRC) under grants EP/H016597/1, EP/I031642/1, and EP/J004111/1, and the support of the EU FP7 program under grant 305815.

Author Disclosure Statement

No conflicting financial interests exist.

References

1. Schneider MV, Orchard S. Omics technologies, data and bioinformatics principles. *Methods Mol Biol* 2011; 719:3–30.
2. Stout M, Bacardit J, Hirst JD, Krasnogor N. Prediction of recursive convex hull class assignments for protein residues. *Bioinformatics* 2008; 24:916–923.
3. Stout M, Bacardit J, Hirst JD, et al. Prediction of topological contacts in proteins using learning classifier systems. *Soft Comput* 2009; 13:245–258.
4. Bacardit J, Stout M, Hirst JD, et al. Automated alphabet reduction for protein datasets. *BMC Bioinform* 2009; 10:6.
5. Widera P, Garibaldi JM, Krasnogor N. GP challenge: evolving energy function for protein structure prediction. *Genet Program Evol Mach* 2010; 11:61–88.
6. Bassel GW, Glaab E, Marquez J, et al. Functional network construction in *Arabidopsis* using rule-based machine learning on large-scale data sets. *Plant Cell* 2011; 23:3101–3116.
7. Glaab E, Bacardit J, Garibaldi JM, Krasnogor N. Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data. *PloS One* 2012; 7:e39932.
8. Fainberg HP, Bodley K, Bacardit J, et al. Reduced neonatal mortality in Meishan piglets: a role for hepatic fatty acids? *PloS One* 2012; 7:e49101.
9. Bacardit J, Widera P, Márquez-Chamorro A, et al. Contact map prediction using a large-scale ensemble of rule sets and the fusion of multiple predicted structural features. *Bioinformatics* 2012; 28:2441–2448.
10. Swan AL, Hillier KL, Smith JR, et al. Analysis of mass spectrometry data from the secretome of an explant model of articular cartilage exposed to pro-inflammatory and anti-inflammatory stimuli using machine learning. *BMC Musculoskel Dis* 2013; 14:349.
11. Swan AL, Mobasher A, Allaway D, et al. Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. *Omics* 2013; 17:595–610.
12. Alberts B, et al. *Molecular Biology of the Cell*, 4th edition. New York: Garland Science, 2002.
13. GenBank. Growth of GenBank and WGS. Available online at www.ncbi.nlm.nih.gov/genbank/statistics (Last accessed on 8/13/2014).
14. UniProt. Current Release Statistics. Available online at www.ebi.ac.uk/uniprot/TrEMBLstats (Last accessed on 8/13/2014).
15. Omics. Alphabetically ordered list of omes and omics. Available online at http://omics.org/index.php/Alphabetically_ordered_list_of_omes_and_omics (Last accessed on 8/13/2014).
16. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* 2010; 467:1061–1073.
17. IBM. Infographics & Animations. Available online at www.ibmbigdatahub.com/infographic/four-vs-big-data (Last accessed on 8/13/2014).
18. Cooper S, Khatib F, Treuille A, et al. Predicting protein structures with a multiplayer online game. *Nature* 2010; 466:756–760.

19. Wu S, Skolnick J, Zhang Y. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol* 2007; 5:1:17.
20. Wolfram MathWorld. Convex Hull. Available online at <http://mathworld.wolfram.com/ConvexHull.html> (Last accessed on 8/13/2014).
21. The ICOS PSP benchmarks repository. Available online at: http://ico2s.org/datasets/psp_benchmark.html (Last accessed on 4/9/2014).
22. Bacardit J, Burke EK, Krasnogor N. Improving the scalability of rule-based evolutionary learning. *Memetic Comput* 2009; 1:55–67.
23. Bacardit J, Llorà X. Large-scale data mining using genetics-based machine learning. *WIREs Data Mining Knowl Discov* 2013; 3:37–61.
24. Monastyrskyy B, Fidelis K, Tramontano A, Kryshtafovych A. Evaluation of residue–residue contact predictions in CASP9. *Proteins* 2011; 79:119–125.
25. Morcos F, Pagnani A, Lunt B, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *PNAS* 2011; 108:E1293–E1301.
26. The ECBDL Big Data Competition. Available online at: <http://cruncher.ncl.ac.uk/bdcomp/index.pl> (Last accessed on 4/9/2014).
27. Haury A-C, Gestraud P, Vert J-P. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS ONE* 2011; 6:e28210.
28. Childs KL, et al. Gene coexpression network analysis as a source of functional annotation for rice genes. *PLoS ONE* 2011; 6:e22196.
29. Urbanowicz RJ, Granizo-Mackenzie A, Moore JH. An analysis pipeline with statistical and visualization-guided knowledge discovery for Michigan style learning classifier systems. *IEEE Comput Intell M* 2012; 7:35–45.
30. Urbanowicz R, Granizo-Mackenzie D, Moore JH. Using expert knowledge to guide covering and mutation in a Michigan style learning classifier system to detect epistasis and heterogeneity. In: Coello Coello CA, Cutello V, Deb K, Forrest S, Nicosia G, Pavone M, (eds.), *Parallel Problem Solving from Nature—PPSN XII*. Berlin: Springer, 2012, pp. 266–275.
31. Urbanowicz R, Granizo-Mackenzie D, Moore JH. Instance-linked attribute tracking and feedback for michigan-style supervised learning classifier systems. *Proceedings of the Fourteenth International Conference on Genetic and Evolutionary Computation Conference*. New York: ACM, 2012, pp. 927–934.
32. Yang X, Ye CY, Bisaria A, et al. Identification of candidate genes in *Arabidopsis* and *Populus* cell wall biosynthesis using text-mining, co-expression network analysis and comparative genomics. *Plant Sci* 2011; 181: 675–687.
33. Rapaport F, Zinovyev A, Dutreix M, et al. Classification of microarray data using gene networks. *BMC Bioinform* 2007; 8:35.
34. Pappa GL, Freitas A. *Automating the Design of Data Mining Algorithms*. Berlin: Springer, 2009.
35. Moore JH, et al. Exploring interestingness in a computational evolution system for the genome-wide genetic analysis of Alzheimer's disease. In: Riolo RL, Moore JH, Kotanchek ME, (eds.), *Genetic Programming Theory and Practice XI*. Berlin: Springer, 2014, pp. 31–45.
36. Holten D, Van Wijk JJ. Force-directed edge bundling for graph visualization. *EuroVis'09 Proceedings of the 11th Eurographics/IEEE—VGTC conference on Visualization*. Aire-la-Ville, Switzerland: Eurographics Association, 2009, pp. 983–998.

Address correspondence to:

Jaume Bacardit
 Interdisciplinary Computing and Complex BioSystems
 Research Group
 School of Computing Science
 Newcastle University
 Claremont Tower
 Claremont Road
 Newcastle-upon-Tyne
 Tyne and Wear NE1 7RU
 United Kingdom

E-mail: jaume.bacardit@newcastle.ac.uk

This work is licensed under a Creative Commons Attribution 3.0 United States License. You are free to copy, distribute, transmit and adapt this work, but you must attribute this work as “Big Data. Copyright 2013 Mary Ann Liebert, Inc. <http://liebertpub.com/big>, used under a Creative Commons Attribution License: <http://creativecommons.org/licenses/by/3.0/us/>”