BIOINFORMATICS – THE MACHINE LEARNING APPROACH

A. Madevska-Bogdanova

Inst, Informatics, Fac. Natural Sc. and Mathematics

Arhimedova 5, 1000 Skopje, Macedonia

ana@ii.edu.mk

Abstract: Computational analysis of biological sequences – linear descriptions of protein, DNA and RNA molecules has completely changed its character since the late 1980s. The main driving force behind the changes has been the introduction of new, efficient experimental techniques, primarily DNA sequencing that has led to an exponential growth of data. As genome and other sequencing projects continue to advance, the interest progressively switches from the accumulation of data to its interpretation. There are some problems concerning the vast amount of data in the biological data-bases that has to be taken into account.

Keywords: bioinformatics, machine learning, SVM, neural networks

1 Introduction

Computational tools for classifying sequences, detecting weak similarities, separating protein coding regions from noncoding regions in DNA sequences, predicting molecular structure and function, finding genes in the DNA sequences, have become an essential component of the research process. This is essential to our understanding of life and evolution, as well as to the discovery of new drugs and therapies. Bioinformatics is emerging as a strategic discipline between biology and computer science, including medicine, biotechnology, pharmacy in many aspects.

Large databases of biological information create both challenging data-mining problems, each requiring specific approach. In this regard, conventional computer science algorithms have been useful, but increasingly unable to solve many of the most interesting sequence analysis problems. This is due to the complexity of biological systems developing in the process of the evolution and our lack of comprehensive theory of life's organization at the molecular level. Machine-learning approaches, like ANN, Support Vector Machines, on the other hand, are ideally suited for domains characterized by the presence of large amount of data containing noise, and the absence of general theories.

2 Problems with the biological data-bases

Although sequence data can be determined experimentally with high precision, they are generally not available to researchers without additional noise. The reasons lie in

the wrong interpretation of experiments and incorrect handling and storage in public databases. Given that biological sequences are stored electronically, that the public databases are maintained by a various group of people, and moreover that the data are annotated and submitted by an even more diverse group of biologists and bioinformaticians, it is understandable that in many cases the error rate from the subsequent handling the information may be much larger that the initial experimental error.

Another problem concerning the analysis of protein and DNA sequences is the redundancy of the data. Many entries in protein or genomic databases represent members of protein and gene families, or versions of homologous genes found in different organisms. Several groups of scientists may have submitted the same sequence, and entries can therefore be more or less closely related, if not identical, by using different names. The use of a redundant data set can be source of errors. First, if a data set of amino acid or nucleic acid sequences contains large families of closely related sequences, statistical analysis will be biased toward these families and will over represent features characteristic to them. Second, obvious correlations between different positions in the sequences may be an artifact of the specific sampling of the data. Finally, if the data set is being used for predicting a certain feature, and the sequences used for building a classification model – the training set are too closely related the sequences used in the test set, the apparent predictive performance may be overestimated, reflecting the method's ability to reproduce its own particular input rather than its generalization ability.

In order to avoid the redundancy problem, some algorithmic solutions have been proposed. Nevertheless, it may often be better to clean the data set first, and thereby give the under presented sequences equal opportunity.

3 Experimental results

The experiments are created with different simulation software packages: MatLab, ver. 6.5 for the Neural Networks experiments and SVM-Light, implementation of the SVM algorithm in C [3].

Recently SVM have been applied to biological issues, including gene expression data analysis or protein classification, particularly because of the high dimensionality of the data. SVMs can easily deal with a large number of features (thousands of genes) and a small number of training patterns (dozens of patients). As a result, the research in this area is increasing, especially in the machine learning community. It can be expected that SVM classifiers will become standard tool in the Bioinformatics, as the clustering algorithms and the dynamical programming are today.

In the following part, few problems from medicine and the molecular biology domain are considered.

3.1 Ovarian cancer problem

Due to hardware improvements (ultrasound equipment) over the last years, high quality measurements have become available to any well equipped gynecologist. Still, experience remains a key factor in the pre-operative discrimination between benign

and malignant (i.e. cancerous) ovarian tumors. The goal is therefore to develop a classifier based on recorded examples of correct classification. This classifier would assist in diagnosis.

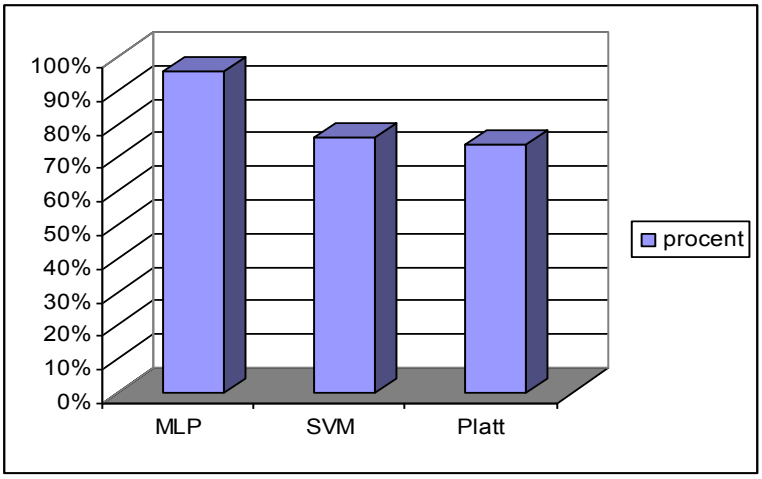

Fig. 1: Graphical representation of the recognition results with different classification models, given in percentages (MLP, SVM, Platt [5])

The molecular biology part begins with the problem of automatic recognition of the mouse chromosomes.

3.2 Mouse chromosome classification

Analysis of mouse chromosomes is important to many fields of genetics research, and there is a need for automatisation of the recognition process.

Mouse chromosomes are more difficult to classify than human chromosomes. The database is consisted of 3723 mouse chromosomes. The input vectors for the SVM classification are a set of 30 discrete values representing the banding profile and chromosome length. Mice have 21 different classes. The classification system is solved by constructing 21 decision hyperplanes. The classification is performed by the highest one-against-the-others values.

The best result: 88% was obtained for GRBF, with

- gamma value 0.1 (g = 0.1);
- train / test set : 2420 / 1303;
- difference between training error and margin C = 110;

The best published result over the same data base (2250 / 1473) is accomplished by using Radial Basis Function Neural Networks and the amount is 87.3 % .

Same training / test set (2250 / 1473), the SVM's result is 87.4 %.

The comparison of the published results on classifying mouse chromosomes using RBF, and the SVM method over the identical data shows slightly better results in favor of SVM. Moreover, they can be easily applied for broad range of problems in Molecular Biology.

Certain number of experiments is done by constructing the MLP classifier. The final results are:

partitions	MLP	RBF	SVM
2250 / 1473	29,7 %	87,3 %	87,4 %
2240 /1303	42,2 %	-	88%

Table 1: Percentage of correctly recognized test patterns for different partitions of the training / test set, with the different classification models

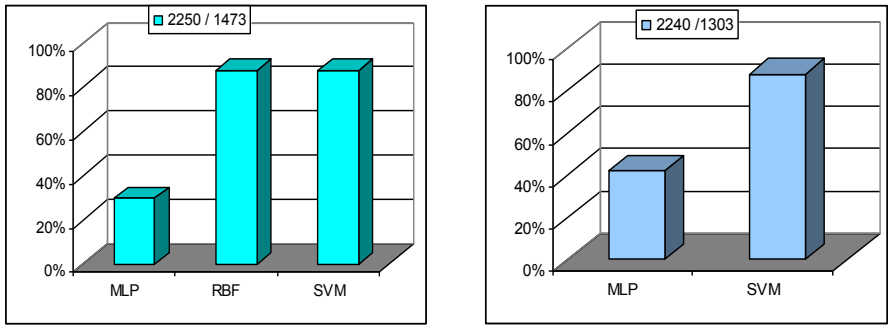

Fig. 2: Graphical representation of correctly recognized test patterns for different partitions of the training / test set, with the different classification models

Comparation of the SVM models results with the published RBF results and with the results from the MLP models, over the identical data set, shows the results in favor of the SVM models.

3.3 Regulatory sequences in the plant Arabidopsis thaliana

The proteins are created according the information provided by the genes. The main process that sets, controls or stops the creation of the proteins, is the binding of some specific proteins on the short parts of the DNA molecule. These parts are called *regulatory sequences* or *promoters*. The identification of this DNA-part is still unresolved problem, especially in the plants. The creation of an automatic systems is required that could recognize, or at least point to some DNA-parts that could be regulatory sequences.

Arabidopsis thaliana has shortest genome among plants. Its regulatory sequences (genetics switches) are short (5 – 10 base pairs). Each gene has a specific pattern of regulatory sequences. They are placed 40 – 1000 bp upstream of a gene. Repressor and activator proteins bind on both strands. By focusing on these regions, searching

through the large part of the genome is avoided and thus the rate of appearance of false positives is lowered. As a conclusion of the considered simulations, the following table and figures are given, where the best results of the different classification models are presented: MLP, MSVMO [4] and Platt [5]. The simulations are taken over the same training / test set, so the results are comparable.

test data	MLP	MSVMO	Platt
pos: 44	57 %	66 %	77 %
neg: 4732	99 %	99 %	98 %

Table 2: Percentage of correctly recognized test data, for the positive and negative examples separatelly, for the different classification models

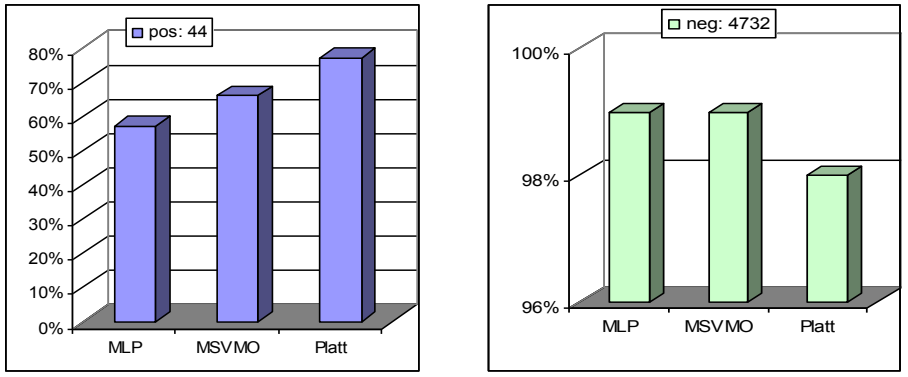

Fig. 3: Graphical representation of the percentage of correctly recognized test data for the positive and negative examples separatelly, for the different classification models

For this problem, best results are obtained with the Platt method. The reason is in the difference between the number of positive and negative examples in the training set. This information is used in obtaining better results compared to the SVM method. In order to have better results with the SVM, i.e. MSVMO model, one should introduce coefficients that would give bigger importance to the positive examples.

3.4 Mitochondrial sequences

Classification of the human mitochondrial sequences is part of the protein subcellular localization problem. Subcellular localization is a key functional characteristic of a potentional gene product such as proteins. Fully automatised classification system is needed, especially for the analysis of the higher organism's genomes. Experimental determination of a subcellular localization mainly is archived by: division of a cell, electronic microscopy and fluorescent microscopy. This methods are time consuming and subjective.

It is shown that function assignment to a protein is especially hard problem, when there is no clear homology with the proteins whose function is previously determined. The proteins should be classified to one of the subcellular components, in order to determine their functionality. The classification is based on the existence of the N-terminal targeting sequence (signal peptide, localization signal). The classification system should learn to recognize the targeting sequence and accordingly to accomplish the process of the protein classification.

The humans are a part of the eukaryotes. Mitochondrial subcellular localization is one of the four categories: nuclear, cytoplasmic, mitochondrial and extracellular.

The obtained simulation results with the chosen coding (coding of the mRNK, responsible for producing the mitochondria protein – positive examples) and the used SVM model, are much better than the published one [30], with coding based on the amino acid composition. SVM classifiers allow big dimensionality of the input vectors and also different length of the genes, presented by the input vectors.

Kernel	data	SVM	Plat	MSVMO (k=-3.42)
linear	pos	71.3%	71.2%	71.3%
	neg	54.8%	54.8%	54.8%
	total	65.5%	65%	65.5%
GRBF C=500 g = 0.005	data	SVM	Plat	MSVMO(k=-123.5)
	pos	73%	59.65%	73%
	neg	60%	65.1%	60%
	total	68%	61%	68%

Table 3: Percentage of correctly recognized testing mitochondrial sequences (train :297 / test: 359)

4 Conclusion

Understanding the biological data, turning it into biological knowledge will lead to revolutions in:

medicine - new or improved diagnosis and cures for many diseases including hereditary diseases and cancer;

pharmacology - easier drug development, more eficient drug production techniques; agriculture - pathogen-resistant crops;

chemistry - new biosynthetic pathways.

Machine learning techniques, as an untraditional approach, offer a great possibility for acheaving the before-mentioned goals.

5 References

1. Beale, R. and Jackson T., *Neural Computing: An Introduction*, IOP Publishing Ltd. 1990;
2. Bishop C., *Neural Networks for Pattern Recognition*, Oxford Press, 1998;
3. *SVM-light*: http://svmlight.joachims.org/
4. Madevska A., Nikolic D., Curfs L., Probabilistic SVM outputs for pattern recognition using Analytical, to appear in: *NEUROCOMPUTING, An International Journal, Elsevier*;
5. Platt J.C., Probabilistic Outputs for SVM and comparison to Regularized Likelihood Methods, in: *A. Smola, P. Bartlett, B. Scholkopf, D. Schuurmans (Eds.) Advances in Large Margin Classifiers*, pp. 61-74, MIT Press, 2000