Partitional Clustering of Protein Sequences - An Inductive Logic Programming Approach*

Nuno A. Fonseca[1,2], Vitor S. Costa[2], Rui Camacho[3], Cristina Vieira[1], and Jorge Vieira[1]

[1] Instituto de Biologia Molecular e Celular (IBMC), Universidade do Porto
Rua do Campo Alegre 823, 4150-180 Porto, Portugal
[2] CRACS-INESC Porto LA, Universidade do Porto,
Rua do Campo Alegre 1021/1055, 4169-007 Porto, Portugal
[3] LIAAD-INESC Porto LA & FEUP, Universidade do Porto,
Rua Dr Roberto Frias s/n, 4200-465 Porto, Portugal

Abstract. We present a novel approach to cluster sets of protein sequences, based on Inductive Logic Programming (ILP). Preliminary results show that the method proposed produces understandable descriptions/explanations of the clusters. Furthermore, it can be used as a knowledge elicitation tool to explain clusters proposed by other clustering approaches, such as standard phylogenetic programs.

Keywords: Clustering, Inductive Logic Programming

1 Introduction

Inductive Logic Programming (ILP) is a machine learning method for discovering logical rules from examples and relevant domain knowledge. There are two major motivations for the use of ILP. First, ILP provides an excellent framework for learning in multi-relational domains. Relations are often used to encode complex structured objects, which may have various number of attributes and which may interact with each other. Second, the models learnt by general purpose ILP systems are in a high-level formalism often understandable and meaningful for the domain experts.

In this paper we describe how ILP can be applied to cluster protein sequences. We focus on two key points: features that can be used to describe protein sequences; and estimation of the distance between two sequences using multiple features. Moreover, we present preliminary results on two data sets.

2 Clustering Protein Sequences

Our approach relies on ILP to obtain a set of features of interest[4] associated to each sequence. Following a significant body of work in ILP[1], in our work a

* This work has been partially supported by the project ILP-Web-Service (PTDC-/EIA/70841/2006) and by Fundação para a Ciência e Tecnologia. Nuno A. Fonseca is funded by FCT grant SFRH/BPD/26737/2006.

[4] Relevant from the domain expert point of view

feature corresponds to a clause, and it holds for a sequence if the clause satisfies the sequence. We followed the approach described in [2] to map each sequence in a set of features. The partitional clustering algorithm is based on the well-known Lloyd's algorithm.

To devise a clustering algorithm it is necessary to define how to estimate a distance between sequences (objects), more precisely, between the sets of features characterising each sequence. We chose a distance widely used within the Bioinformatics community - the Tanitomo distance or coefficient [3] (also known as Jaccard index):

$$m(a,b) = \frac{|S_a \cap S_b|}{|S_a \cup S_b|} = \frac{|S_a \cap S_b|}{|S_a| + |S_b| - |S_a \cap S_b|}$$

where a and b are two sequences and S_a and S_b are, respectively, the set of features valid for each sequence.

To determine the clustering quality, while searching for a (local) best clustering, we implemented the following measure from [4] that aims at minimising the distance within the clusters wc and maximising the distance between clusters bc:

$$quality(C) = bc(C)/wc(C)$$

The features associated to each sequence are of two main types of knowledge: properties and relations. By properties we mean inherent characteristics of the protein sequences that can be computed from the sequence. This includes the isoelectric point, charge, molecular weight, average residue weight, number of residues, and k-mers (for $k > 5$ and number of occurrences greater than 10% of the set of sequences) contained in the sequence. The properties are computed using utilities available in EMBOSS [5] and for the k-mers we use wd [6]. The features involving relations encompass similarity between sequences in the data set (computed using Blast), and gene ontology (GO) annotations of similar sequences in NCBI. To obtain GO annotations for a sequence, the NCBI database is queried for similar sequences and then GO annotation information is gathered using the Blast2GO software [7].

In general, a cluster may have more than a single explanation, i.e., different features of the examples can justify the cluster. Arguably, the features over-represented may help, or even be sufficient to understand a cluster. We therefore want to look for features that are most likely to have a different distribution in the cluster. To this end we followed a widely used way to estimate distances between distributions, the Kullback-Leibler (KL) divergence:

$$D_{KL}(P \parallel Q) = P\frac{log(P)}{log(Q)} + (1-P)\frac{log(1-P)}{log(1-Q)}$$

where Q is the probability of a feature being found in the whole set of sequences and P is the probability that a feature being found in the cluster. Therefore, each cluster is represented by the feature with higher KL divergence.

3 Preliminary Experiments and Results

The goal of the experiments was two fold: i) determine to what extent the clusterings created are meaningful for a molecular biologist; ii) assess the differences, if any, between the clusters produced and the groups suggested by a phylogenetic approach. Two data sets of protein sequences were considered: the `serpin` data set with 66 serpin genes from human and insect; and the `human serpin` data set composed by the 35 human serpin genes from the `serpin` data set. The sequences in the data sets are very divergent. The average level of identity between each sequence in the `human serpin` data set is 31%, and is considerable less in the `serpin` data set.

In the `serpin` data set we would expect a clustering that partitions the data set into a cluster of human and a cluster of insect serpines. The clustering, when considering three groups, splits the data set into two homogeneous clusters of 7 and 6 sequences from insects and a third cluster containing the remaining sequences of insects and human serpines. The majority of interesting rules on each cluster include k-mers information. For instance, the rule *has word fkgqwk* is observed *exclusively* in *all* elements of the cluster containing 7 sequences.

For the `human serpin` data set, a clustering partitions the set of sequences into two clusters: the *cluster*1 contains the sequences SA1, SA3, SI2, SB4, SB12, SB8, SB2, SB13, SB10, SB6, SI1, SB9, and SA2; and the *cluster*2 with the remaining sequences. The two clusters are overlapped in a phylogenetic tree (see Figure 1). There is not a clear match between the clusters proposed and the groups in the tree. However, *cluster*1 has a good coverage of the group G2 in the phylogenetic tree. The *cluster*1 is characterised by all sequences in the group having an *isoeletric point below* 6.1313 - this characteristic is only observed in two sequences of group 2 (SB5 and SA6).

When we try to get an explanation for the well defined phylogenetic groups in the phylogenetic tree (G1, G2, and G3), the majority of the interesting rules involve the sequences having a k-mer. For instance, the rule *has word gfqhl* is observed exclusively in four sequences (SA9, SA6, SA4, and SA7) of group G3.

The results presented although preliminary are encouraging. We plan to proceed by performing some refinements in the current implementation and a more in depth empirical evaluation.

References

1. F. Zelezný and N. Lavrač. Propositionalization-based relational subgroup discovery with rsd. Machine Learning, 62(1-2):33–63, 2006.
2. Nuno A. Fonseca, Rui Camacho, Ricardo Rocha, and Vitor Santos Costa. Compile the hypothesis space: do it once, use it often. *Fundamenta Informaticae*, Special Issue on Multi-Relational Data Mining(89):45–67, 2008.
3. L. Ralaivola, S. J. Swamidass, H. Saigo, and P. Baldi. Graph kernels for chemical informatics. *Neural Netw.*, 18(8):1093–1110, 2005.
4. David J. Hand, Padhraic Smyth, and Heikki Mannila. *Principles of data mining.* MIT Press, Cambridge, MA, USA, 2001.

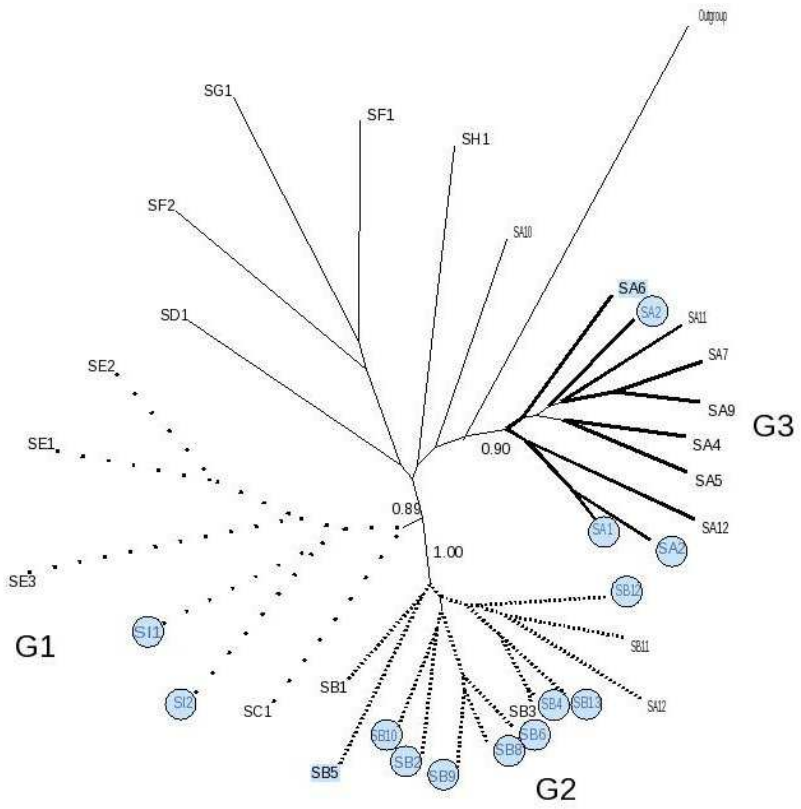

Fig. 1. Phylogenetic tree produced by MrBayes [8] for the `human serpin` data set. Each serpin is identified in the tree by its clade (A, B, ...) and membership (1,2, ...). The input alignment for MrBayes was produced by the Accurate mode of T-Coffee [9]. Circled names belong to cluster 1, non-circled ones belong to cluster 2.

5. P. Rice, I. Longden, and A. Bleasby. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, 6(16):276–277, 2000.
6. Pedro Pereira, Nuno A. Fonseca, and Fernando Silva. Fast Discovery of Statistically Interesting Words. Technical Report DCC-2007-01, DCC-FC & LIACC, Universidade do Porto, 2007.
7. Ana Conesa, Stefan Götz, Juan Miguel García-Gómez, Javier Terol, Manuel Talón, and Montserrat Robles. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676, 2005.
8. F. Ronquist and J. P. Huelsenbeck. Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574, August 2003.
9. C. Notredame, D. G. Higgins, and J. Heringa. T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1):205–217, September 2000.