The role of analogies in biological data: a study in the exploratory analysis of protein-protein interactions

January 18, 2007

Ricardo Silva
Gatsby Unit
University College London
London, UK WC1N 3AR
rbas@gatsby.ucl.ac.uk

Edoardo Airoldi
Carl Icahn Laboratory
Princeton University
Princeton, NJ 08544
eairoldi@princeton.edu

Katherine A. Heller
Gatsby Unit
University College London
London, UK WC1N 3AR
heller@gatsby.ucl.ac.uk

Abstract

We illustrate a new methodology for the exploratory analysis of protein-protein interactions motivated by the principles of *analogical reasoning*. In this work, we present an application which automatically discovers analogies between pairs of proteins that are known to interact. For example, given a pair of interacting proteins, P_1:P_2, which of two other interacting pairs, P_3:P_4 or P_5:P_6, interacts in a way that is most "similar" to P_1:P_2? In other words, is the interaction P_3:P_4 more analogous to P_1:P_2 than P_5:P_6 is? The goal of such exploratory analysis is to find new subclasses of interactions that might be relevant for further study: e.g., P_1:P_2 might belong to a class of interactions that is not yet fully formalized, and scientists exploring the interaction between P_1:P_2 might want to find other interactions which behave in an analogous way. We present a Bayesian formulation of this question and illustrate its potential application for exploring new taxonomies of protein-protein interactions.

The main principle of analogical similarity is: one should compare structured objects X and Y by how closely the relations within components of X correspond to relations within components of Y (French, 2002). This is in contrast to a direct comparison of features of the components of each respective object. In our domain, a structured object X_{ij} is a pair of proteins P_i:P_j, each protein being a component of X_{ij}, and the relation being a physical binding between the components.

Our application is built upon a measure of analogical similarity derived by Silva et al. (2007). In particular, it quantifies the similarity of two pairs of proteins (P_1:P_2, P_3:P_4) via a discriminative approach. It starts from the assumption that there is a (unknown) classification function $f_1(\cdot, \cdot)$ that classifies P_1:P_2 as interacting (as opposed to a class of proteins that do not interact), and a (unknown) classification function f_2 for P_3:P_4. We show

how this task can be reduced to a novel variation of the Bayesian sets method (Ghahramani and Heller, 2005) for relational data with a discriminative model.

The evaluation follows criteria commonly used in applications of information retrieval. Namely, we define subgroups of protein-protein interactions using the Munich Institute for Protein Sequencing (MIPS) database and its respective taxonomy for gene/protein role (Mewes and et. al, 2004). We then perform queries using a subset of protein-protein pairs from a particular subpopulation. Precision-recall curves are derived from the resulting rank of the remaining pairs by comparing how closely the MIPS taxonomy of the retrieved pairs match those in the query.

Topic: data mining
Preference: poster

References

R. French. The computational modeling of analogy-making. *Trends in Cognitive Sciences*, 6:200–205, 2002.

Z. Ghahramani and K. Heller. Bayesian sets. *18th NIPS*, 2005.

H. Mewes and C. Amid et. al. MIPS: analysis and annotation of proteins from whole genome. *Nucleic Acids Research*, 32, 2004.

R. Silva, K. Heller, and Z. Ghahramani. Analogical reasoning with relational Bayesian sets. *11th International Conference on Artificial Intelligence and Statistics, AISTATS*, 2007.