BMC Systems Biology

BioMed Central

Poster presentation

Open Access

A Bayesian framework for integrating genomic data to aid function prediction

Chris Bridson* and Richard J Morris

Address: Computational and Systems Biology, John Innes Centre, Norwich, Norfolk, NR4 7UH, UK

Email: Chris Bridson* - Chris.Bridson@bbsrc.ac.uk

* Corresponding author

from BioSysBio 2007: Systems Biology, Bioinformatics and Synthetic Biology
Manchester, UK. 11–13 January 2007

Published: 8 May 2007

BMC Systems Biology 2007, **1**(Suppl 1):P62 doi:10.1186/1752-0509-1-S1-P62

This abstract is available from: http://www.biomedcentral.com/1752-0509/1?issue=S1

Introduction

The function of a protein can be associated with a number of factors, from its cellular location, expression profile, interaction partners, down to its molecular structure, intrinsic physico-chemical properties and sequence. Machine learning methods such as Neural Networks and Support Vector machines are active bioinformatics research areas which attempt to combine these various properties into a grand scheme for predicting protein function. Powerful as these methods may be, they can be computationally expensive, are prone to over-fitting unless used with great care, require larger training sets than currently available for many biological problems, and most importantly provide little insight into the relative importance and contributions of different data sources. In addition, the transition to a proper probabilistic model of such approaches is not trivial. Bayesian [1] methods offer a number of advantages and possible solutions to these issues and in this contribution we present our experience with Bayesian Networks on a few toy examples.

Method

A set of associations between different factors represented within BRENDA [2] (such as ligand, co-factor, substrates) and their related gene ontology (GO) identification numbers were established using a set of perl scripts that build up prior distributions from counting statistics. The prior distribution of, for example, co-factors given the GO term, can so be obtained from counting items within the database. A naive Bayes classifier was setup to analyse the joint probability distributions of these different factors within BRENDA. The posterior distribution was then computed following Bayes' law to update the prior knowledge with an empirical likelihood function of the observed data (in this case, the outcome of, for instance sequence alignments, localisation prediction, docking experiments, etc.). The maximum posterior probability determines the best functional hypothesis and the entropy of this distribution the uncertainty.

Results

Using a low-dimensional toy example, we illustrate how initial distributions (prior knowledge) can be updated with new pieces of evidence and how the full uncertainty and errors can be correctly propagated through each step to ensure that the results are not biased. We show how this can be applied to protein function prediction and highlight the importance of probabilistic reasoning in this area.

Conclusion

Initial results have demonstrated that it is possible to differentiate between certain biochemical GO terms using features taken from BRENDA (such as ligand and co-factor to GO function). Although the method we have used is relatively simple it can yield accurate results. We are currently incorporating more feature sets from sources other than BRENDA, as well as performing optimisation on the scripts, with respect to run time and likelihood formulation to improve the classification performance. Given the success of the method we envision these approaches being at the core of data integration and function prediction pipelines such as ProFunc [3].

References

1. Needham CJ, Bradford JR, Bulpitt AJ, Westhead DR: **Inference in Bayesian networks.** *Nat Biotechnol* 2006, **24:**51-53.
2. Schomburg I, Chang AJ, Hofmann O, Ebeling C, Ehrentreich F, Schomburg D: **BRENDA: a resource for enzyme data and metabolic information.** *Trends Biochem Sci* 2002, **27:**54-56.
3. Laskowski RA, Watson JD, Thornton JM: **ProFunc: a server for predicting protein function from 3D structure.** *Nucleic Acids Res* 2005, **33:**W89-W93.