Poster B-30

Prediction from complex bioinformatic databases, with application to protein--protein interactions

Authors:
Berzuini Carlo (*MRC BSU*)
Tom M.W. Nye (*MRC BSU*)
Sarah A. Teichmann (*MRC LMB*)

Short Abstract: We consider predictions based on a synthesis of information from a heterogeneous collection of databases. Our method involves a weighted combination of predictive distributions generated by each individual database, where each weight measures the independent predictive contribution of the corresponding database. We apply the methodology to predict domain--domain contacts within a given multimeric protein complex, by synthesizing information from a database of atomic-level, three--dimensional, multimer structures and from rather crude, genome--wide, experiments indicating which pairs of proteins have potential to multimerise.

Long Abstract:
Biological research is increasingly database driven, and increasingly dependent on data generated with the aid of high-throughput experimental platforms at several levels of organization of the genome/ proteome/ metabolome. It is therefore important to rate each individual database in terms of scientific value, and one way of doing this is by assessing the quality of the predictions that flow from it. Our approach to such an assessment is based on Dawid's prequential method to model validation (Dawid:1984). This method assumes we have a series $y^t=(y_1, ..., y_t)$ of observed realizations of a quantity Y of predictive interest, and a system for generating a probabilistic prediction P_i for each Y_i on the basis of earlier observations $y^{i-1}=(y_1, ..., y_{i-1})$. We measure the inadequacy of each P_i, in the light of the materialized outcome, $Y_i=y_i$, in terms of a penalty, $S(y_i, P_{i})$. Having applied the system for $i=1$ to t, and resulting in predictions $(P_{1}, \ldots, P_{t})$, its performance can be measured by the total penalty $S_{1:t} = \sum_{j=1}^t S(y_j,P_{j})$.

\newpar Motivated by a typical analysis situation in bioinformatic studies, we consider the situation where each prediction P_{i} is a synthesis of information from a heterogeneous collection ${\cal D}$ of databases, and we propose that this synthesis take the form of a weighted combination of predictive distributions flowing from each individual database. We show that these weights can be robustly estimated via sequential optimization of the chosen penalty function. We are mainly concerned with the choice of a suitable penalty function, that we wish to tailor to the specific nature of the prediction of interest. As an aid in the choice, we examine general properties of penalty functions for a categorical predictand.

\newpar We apply the methodology to predict domain--domain contacts in protein complexes. At a coarse level of structural description, a protein can be regarded as a contiguous set of one or more modular components called {\em domains}, which take the

form of compact globules in three-dimensional space. Recent statistical methods allow domains to be extracted from a protein's amino acid sequence and classified into homologous families (Gough etal 2001)(Gough,2002). Proteins frequently aggregate into multimers, the 3D structure of a multimer being maintained by physical contacts between the domains of its constituent proteins. It is believed that some domain families are more promiscuous than others in forming contacts, so that an important intermediate target is to identify those family combinations (which we shall call signatures) which play an active role in the formation of multimers.

Experimentally determined, atomic-level, 3D structures are currently available in the Protein Quaternary Structure (PQS) database (Henrick:Thornton,1998), but unfortunately only for a minority of multimers, due to well known technical difficulties in crystallographic experiments. By contrast, protein--protein interaction (PPI) databases(Ito:Chiba:etal:2001)(Uetz:Giot:etal:2000) contain genome wide (albeit crude and qualitative) experimental data indicating which pairs of proteins have the potential to multimerize. The idea is then to synthesize information from the PQS and the PPI databases to predict domain--domain contacts in the many still uncharacterized multimers. Such predictions may contribute to understanding multimer structure and function on a genome-wide scale, well beyond the horizons of experimental crystallography, and ultimately to understanding networks of proteins (Tong:etal:2002), as well as to predicting the effects of specific single-nucleotide polymorphic mutations (Wang and Moult,2001) or of alternative splicing events (Resch etal,2004). It would also contribute to the development of drugs to inhibit pathological protein--protein interactions (Loregian etal, 2002)(Gadek and Ockey,2002)(Zutshi etal,1998}, and to designing novel protein interactions from appropriate domain scaffolds (Dueber etal,2004).

We provide insight into the the problem by assessing the contribution of PPI data to predicting domain--domain contacts in a multimer from sequence information about its constituent proteins. In this context, a major source of difficulty has been the complexity of the data and of the underlying stoichochemistry. Important steps of the analysis are the choice of a suitable penalty function, and the scoring of relevant information from the available databases. Concerning the latter, one method we propose involves extracting scores from the PPI database based on p-values obtained by shuffling domains across proteins. Wealso discuss and apply a particular form of prequential validation which reduces sensitivity to the particular order in which the data of the reference database are analyzed.