FROM PROTEIN SEQUENCE TO PROTEIN SPECIFICITY: COMPLETELY AUTOMATED DISCOVERY AND MAPPING OF SPECIFICITY DETERMINING RESIDUES

Kolesov G.*, Mirny L.A.

Harvard-MIT Division of Health Sciences and Technology, MIT, Cambridge, MA 02139, USA
* Corresponding author: e-mail: g.kolesov@mail.ru

Keywords: protein structure, specificity determining residues, mutual information, evolution, phylogenetics, multiple sequence alignment

Summary

Motivation: Most proteins bind and recognize specific targets such as other proteins, specific sites on DNA or RNA and small molecules. Specific recognition is essential for all cellular functions and is frequently affected in diseases. A detailed molecular understanding of protein specificity is essential for structure-based drug design, understanding of effects of mutations and rational manipulation of protein specificity.

Results: A robust, completely automated bioinformatics framework for the recognition of specificity determining residues has been developed and tested.

Availability: At the time of this writing the program is not available for public use. Computations for individual proteins can be performed on request to the authors. We plan to make our method available as a World Wide Web server in the future.

Introduction

In the recent years a variety of approaches have been proposed aiming to exploit evolutionary relationships within a given protein family and pinpoint amino-acid residues that determine the specificity of the protein. These approaches often rely on additional information and/or additional expert analysis. For instance, reliable predictions of specificity determining residues (SDRs) often require the computation of phylogenetic trees, availability of high quality multiple sequence alignments (MSAs) and visual inspection of protein 3D-structure models (Lichtarge, Sowa, 2002; Mirny, Gelfand, 2002; Kalinina et al., 2004).

Here we report a completely automated bioinformatics approach, which for a given protein sequence
1. finds members of the same family in the sequence database, 2. partitions the resulting set to orthologous groups, 3. computes the MSA, 4. locates putative SDRs (PSDRs), 5. estimates statistical significance of the PSDRs, 6. maps the significant positions on the protein’s 3D-structure model (if available) and 7. generates a web report page.

Methods

Assembling and partitioning sequence set. When entire protein sequences are being used, the construction of reliable MSAs is hindered or simply infeasible due to the multi-domain structure of proteins and domain shuffling events in the course of evolution. A better approach is to consider individual domains rather than entire protein.

In the first step of the algorithm a PFAM (Bateman et al., 2000) domain search is conducted for the query sequence. Domains contributing to the protein are then analyzed separately. Alternatively, a PFAM hidden Markov model (HMM) domain profile can be fed directly into the program to initiate the analysis. We shall refer to a domain sequence as simply a ‘sequence’ further in this text.
Associated to a sequence PFAM HMM profile is then run against a protein sequence database to harvest homologous domains originating from different organisms. To find orthologous domains all-against-all BLAST (Altschul et al., 1997) alignment is then conducted for the resulting sequence set. Bi-directional best BLAST hits are identified.

Sequences are clustered into orthologous groups by using two approaches depending on the size of the set. If the set is small (<150 sequences) the single linkage clustering of best-to-best BLAST hits is applied.

In the cases when the set is large enough we select sequences originating in the same organism (usually the organism containing original sequence or an arbitrary one) as the ‘parents’ of the clusters. The ‘parent’s’ best-to-best BLAST matches are included in the cluster. Overlaps are resolved by omitting the second and following occurrences of the same sequence in the clusters. Clusters containing a single sequence are removed. The underlying assumption of this method is that the organism contains only a single instance of each orthologous domain. While this is considered to be good estimate for prokaryotes with their compact genomes, in eukaryotic genomes the situation can be different and more sophisticated techniques such as INPARANOID (Remm et al., 2000) should be applied.

**MSA.** Multiple sequence alignment is conducted by either CLUSTAL W (Thompson et al., 1994) or the *hmmalign* program from the HMMER package (Eddy S.R., 1998) using default parameters.

We did not find any significant difference between MSAs obtained from these two programs.

**Locating PSDRs with mutual information.** To identify residues that can discriminate between paralogous proteins (different specificity), merging orthologs (same specificity) together, we used the mutual information as a measure of association with the specificity:

\[
I_i = \sum_{x=1,20} \sum_{y=1,Y} P_i(x,y) \log \frac{P_i(x,y)}{P_i(x)P_i(y)},
\]

where \(i, x\) and \(y\) denote the position in the alignment, the amino acid type and the orthologous group number (the same for all proteins of the same specificity group) respectively. \(Y\) is the total number of groups in an alignment. \(P_i(x,y)\) is the joint probability of \(x\) and \(y\), i.e. the probability of finding amino acid type \(x\) at position \(i\) and in group \(y\). \(P_i(x)\) is the marginal probability of finding amino acid type \(x\) at position \(i\) regardless of groups, and \(P_i(y)\) is simply the fraction of proteins belonging to group \(y\). Importantly, \(I_i\) measures the correlation between \(x\) and \(y\), and \(I_i = 0\) if and only if \(x\) and \(y\) are statistically independent (Cover, Thomas, 1991).

Mutual information has several important properties: 1) it is non-negative; 2) it equals zero if and only if \(x\) and \(y\) are statistically independent; and 3) a large value of \(I_i\) indicates a strong association between \(x\) and \(y\). Unfortunately, a small sample size and a biased amino acid composition of each column in the MSA influence \(I_i\) a lot. For example, positions with less conserved residues tend to have higher mutual information. Hence, we cannot rely on the value of \(I_i\) as an indicator of specificity association, instead we estimate the statistical significance of \(I_i\).

**Estimating statistical significance of PSDRs.** To evaluate the statistical significance of an \(I_i\), we need “control” MSAs to estimate the \(P(I)|\) and the \(p\)-value (the probability of observing this or higher \(I_i\) in the control). The control-MSA should carry most of the properties of the real MSA. We base our choice on the following reasoning. There are two major mechanisms of conservation for an amino acid position.

1. The first mechanism is independent of amino acid position in a protein. Examples of the first mechanism are chance and phylogeny (Wollenberg, Atchley, 2000). Position-independent signal can be take into account if we construct a “control” MSA using the same sample size and the same phylogenetic tree.
2. The second mechanism is position-specific, and conserve residues that have important structural
or functional role. For example, highly hydrophobic positions in MSA reflect selection of a residue in the hydrophobic core that stabilizes protein structure. A glutamate residue conserved across all proteins in the family can participate in the active site or in non-specific binding. In other words, position-dependent selection (other than sought diversifying selection of SDRs) is manifested in MSA columns with a biased amino acid composition. To compute statistical significance and rule out high $I_i$ due to other evolutionary signals, we construct a “control”-MSA that has the same amino acid composition at every position $i$.

We developed two different methods to compute $P(I_i)$, which, however, produced very similar results. The first one takes into account only (2) and then compensates for the position-independent signal. The second one simulates protein evolution under selection and explicitly generates MSA that satisfies (1) and (2). Both methods are described in detail in (Mirny, Gelfand, 2002).

**Mapping to a 3D-structure model.** Mapping to a Protein Databank (PDB) protein 3D-structure model is achieved through the ability of the hmmalign program to align a sequence to the MSA (using the MSA’s HMM-profile) while keeping the MSA intact. When such an alignment is available the translation of MSA coordinates into PDB coordinates can be performed easily.

A Perl script that automates these steps is available on conditions of GNU General Public License (GPL) at {http://web.mit.edu/~grigory/www/perl/sdr/pdbmsamap.pl}.

**Web display.** All results, including intermediate ones, are included into the web page report. MSAs are provided as both a FASTA-formatted file and a colored alignment with positions of statistically significant PSDRs highlighted.

PDB structures are displayed using excellent Rasmol/Chime-like Jmol software ({http://jmol.sourceforge.net}). Rasmol script highlighting PSDRs and selecting proper protein models (for NMR-derived structures) and chains is automatically supplied to a Jmol Java applet.

**Results and Discussion: a case study**

We have arbitrarily selected the GerE family of bacterial transcription regulators to illustrate our method. GerE-domains belong to the family of LuxR domains, which, in turn, belong to the larger family of classical ‘helix-turn-helix’ DNA-binding domains.

We chose the following genomes for our analysis: *Agrobacterium tumefaciens, Aquifex aeolicus, Bradyrhizobium japonicum, Brucella melitensis, Brucella suis, Campylobacter jejuni, Caulobacter crescentus, Escherichia coli, Mesorhizobium loti, Neisseria meningitidis, Rhodopseudomonas palustris, Rickettsia conorii, Rickettsia prowazekii, Sinorhizobium meliloti, Thermoplasma acidophilum, Thermotoga maritima* and *Wolbachia*. The choice of these genomes was not arbitrary but rather dictated by a related study we had been conducting at the time.

We used the GerE PFAM HMM profile (accession number PF00196.8) as an initial input to our program. As the result of the analysis 101 domains were found and clustered into 23 orthologous groups. Eight PSDRs were identified as having significantly high mutual information with $p$-value < 1.0E-4.

Screenshots of selected pages of the web report are presented in Fig.

As seen on Fig. b seven out of eight PSDRs are indeed directly contacting the DNA. Six of these residues fit tightly into the major groove of the DNA, one contacts the DNA backbone and one is located higher on the surface of the protein. The seven residues contacting a DNA are likely to be responsible for specific DNA site recognition while the one located higher on the surface of the protein can be responsible for contact with other proteins involved in the gene regulation network or can be a false positive.
References