An integrated pipeline for protein classification using specific PSSMs and existing protein annotations

Kyung Dae Ko[1] and Hongfang Liu[2]
[1]Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University
[2]Department of Health Sciences Research, Mayo Clinic College of Medicine

Abstract - *Protein classification has been performed by many protein databases to infer annotations of unknown proteins and therefore enhance the performance of protein annotation. In this study, we implemented an integrated pipeline for protein classification using specific PSSMs and proteins with the same entity name. After clustering sequences on the basis of their evolutionary distances, a target group is selected using Jarccard distance. Finally, each group is represented using specific PSSMs generated from sequences in the target group. Using 76 p53-relative and 155 non-relative sequences to validate the performance of our pipeline, we measured 100% accuracy of protein classification by our pipeline. In addition, we identified 35 homologous proteins of p53 among 86,718 sequences through high-throughput analysis of human proteome.*

Keywords: PSSM, protein classification, text mining, Jarccard distance, p53

1 Introduction

There is a high demand of automated protein annotation approaches and methods due to the latest advance in high throughput genomics and proteomics technology. However, automated protein annotation is a very challenging task in computational biology. In general, the first step in annotating a novel protein is to identify homologous proteins related to the protein. If its homologous proteins are well annotated, we can infer the characteristics of the protein from the homologous proteins' annotations.

One of the simplest methods to identify homologous proteins is to measure the similarity between novel and reference sequences [1, 2]. If their identity is high, they can be structural and/or functional homologous. However, for sequences that are distantly related, sequence-sequence comparison algorithms may lose the sensitivity in detecting the homologous relationship [3]. To increase the sensitivity in detecting remote homologues, instead of comparing two proteins directly through pair-wise sequence alignment, the new sequence can be compared with profiles, which contain common information from known protein sequences belonging to the same families. Indeed, after building multiple sequence alignments of related sequences in the same family, a PSSM (Position Specific Scoring Matrice) or HMM (Hidden Markov Model) model is then generated on the basis of the common information from the alignments. Using PSSM or HMM, sequence-profile comparison methods such as PSI-BLAST (Position specific iterative-BLAST) and SAM (Sequence Alignment and Modeling System) can increase the sensitivity in detecting the distant homologous sequences with low sequence identities [4, 5]. In addition, the sensitivity and specificity of PSSM or HMM tend to depend on sequences used for building multiple sequence alignments. Thus, specific PSSMs generated from functional related sequences can improve the sensitivity of protein classification.

In this study, we implemented an integrated pipeline for protein classification using specific PSSMs and considering proteins with the same name based on the observation that biologists tend to assign related genes or proteins similar names. Sequences are clustered on the basis of their evolutionary distances. After selecting a target group using Jarccard distance, specific PSSMs are generated from sequences in the target group. Finally, each group is represented using specific PSSMs.

In next section, we describe the background information of tools and resources used in the pipeline. We will then introduce our classification pipeline. A case study based on p53 (tumor suppressor protein) is provided in detail.

2 Method and Resources

2.1 Tools and Resources

The tools in this study contain PSSM and RPS-BLAST (Reverse Position specific iterative-BLAST). A PSSM profile is a position-specific scoring matrix with 21 columns and *M* rows where *M* is the length of probe. Each row matches a sequence position of the probe [6]. The first 20 columns in each row show the score for searching each of 20 amino acid residues at the specific position of the target sequence. A penalty for insertions or deletions (INDELs) at each position of the target sequence is encoded in the 21st column. When a target sequence is compared with PSSMs, the highest score or scores above a specified threshold are retained as outputs [6]. RPS-BLAST searches homologous sequences in the inverse way of PSI-BLAST [7]. Thus, it reverses the role of a sequence and PSSMs, comparing a query sequence against a library of position-specific scoring matrices (PSSMs).

The resources used in the study include UniProtKB, a comprehensive knowledgebase about protein sequences and functional information, BioThesaurus, a comprehensive collection of gene/protein names collected from over 30 molecular databases for UniProtKB records, and several gene/protein family classification and functional annotation knowledge bases including PANTHER, PIRSF, and Gene Ontology. The following summarizes them.

UniProtKB provides the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information [8, 9]. It consists of a manually annotated and reviewed component, Swiss-Prot, and an automatically annotated component, TrEMBL. Proteins with sequence similarities of 50% or 90% were grouped into UniREF50 and UniREF90 clusters [10].

BioThesaurus is a thesaurus aiming to provide a comprehensive collection of protein and gene names for protein records in the UniProtKB. Currently covering six million proteins, the latest version of BioThesaurus consists of over eight million names extracted from multiple molecular biological databases according to the database cross-references in UniProtKB and iProClass [11].

The PANTHER (Protein ANalysis THrough Evolutionary Relationships Classification System) is a resource that classifies genes by their functions, using published scientific experimental evidence and evolutionary relationships to predict function even in the absence of direct experimental evidence [12]. Proteins are classified by expert biologists into families and subfamilies of shared function, which are then categorized by GO terms.

The PIRSF (Protein superfamily classification system) is a protein classification system based on the domain information of the whole proteins. It provides comprehensive and non-overlapping clustering of UniProtKB sequences into a hierarchical order to reflect their evolutionary relationships [13].

Gene Ontology (GO) presents a structured vocabulary about biological roles of gene and proteins from different species [14]. GO defines three different parts including molecular function, biological process and cellular component. GO terms are organized in directed acyclic graphs (DAG) whose nodes have child-parent relationships [14].

PHYLIP (Phylogeny Inference Package) is a package of programs for inference of phylogenies from sequences. Data types of the package include molecular sequences, gene frequencies, restriction sites and fragments, distance matrices, and discrete characters. Methods in the package are to generate distance matrix and consensus trees, and calculate bootstrapping, parsimony, and likelihood [15].

2.2 Method

Figure 1 shows the pipeline that consists of three modules. The first module is to collect sequences from public databases based on names collected in BioThesaurus, then calculate evolutionary distances among sequences, and finally cluster proteins in groups on the basis of their evolutionary distances.

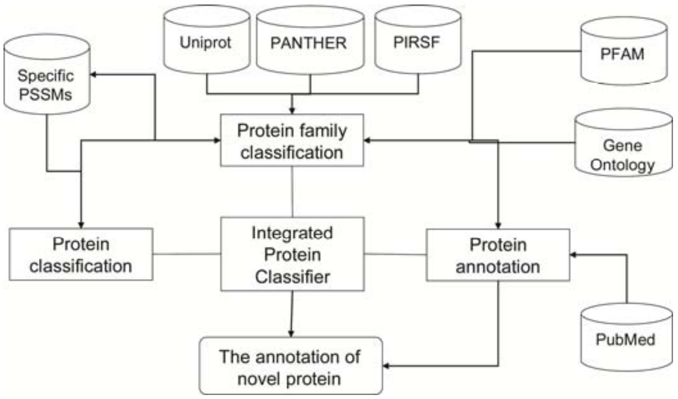

Figure 1: Diagram showing the workflow of the pipeline for protein classification

The second module is to characterize clustered groups by measuring the dissimilarity between the groups and reference protein families in PIRSF and PANTHER. After calculating the relative frequencies of domain architectures and Gene ontology terms which each protein family has, we use weighted Jaccard distance to measure their dissimilarity. Jaccard distance is generally used to measure dissimilarity between sample sets [16], and is calculated by subtracting the Jaccard coefficient from 1 in equation (1) and (2). Then, we give a relative frequency weight to Jaccard distance for reflecting the number of domain architectures or GO terms. Since the sum of the relative frequencies of domain architectures or GO terms is 1 in a protein family, we assume that the probabilities that protein family C_1 and C_2 have the same domain architecture or GO terms are $P(C_1)$ and $P(C_2)$. Then, the weight is defined in equation (3) assuming independency and mutual exclusiveness. We finally define weighted Jaccard distance in equation (4) :

$$J(C_1,C_2)=\frac{|C_1\cap C_2|}{|C_1\cup C_2|} \tag{1}$$

$$J_d(C_1,C_2)=1-J(C_1,C_2) \tag{2}$$

$$E_J=\frac{P(C_1)+P(C_2)-P(C_1)\times P(C_2)}{P(C_1)+P(C_2)} \tag{3}$$

$$J_M = E_J \times J_d \tag{4}$$

The third module is to identify the characteristics of a protein using specific PSSMs. After selecting a target group for the classification, we generate specific PSSMs using sequences in the group. PSSM generally describes the distribution of residues at each position in a conserved pattern such as motif or domain. Thus, if we generate specific PSSMs using sequences in a specific group, the specific PSSMs can allow us to identify proteins whose functional charactersitics are similar to the specific group in novel proteins or proteomes.

Based on this assumption, a pipeline first generates PSSMs from sequences which have similar domain architectures and functional GO terms. Second, each query sequence is searched against the specific PSSMs using RPS-

BLAST. If the alignments returned from the search do not satisfy our e-value threshold, they are filtered out. Then, given the alignments to specific PSSMs, a residue score is calculated. For every alignment returned from the RPS-BLAST search of each query against specific PSSMs, each amino acid of a query which is identically or positively (identical, but conserved) aligned is scored with BLOSUM62 score for the aligned pairs. These scores are summed for each amino acid of the query (i.e., residue score). The specific score for a query protein is calculated using equation (5).

$$\frac{1}{n}\sum_{i=1}^{n} p_i \text{ if } p_i > 0 \tag{5}$$

where n is the length of a protein sequence and p_i is a positional score of ith amino acid of the protein.

3 Results

To validate the performance of our pipeline, we first selected p53 tumor suppressing protein as a key word. Using BioThesaurus, we collected 205 sequences, which have the entity name as “p53”, and 3204 sequences, whose sequence similarities are over 50%, from UniProtKB, based on UniREF50. We calculated evolutionary distances among these sequences using phylip library and clustered them into 38 groups.

Table 1. The weighted Jaccard distances of domain architecture and functional GO term between group1 and protein families in PIRSF.

	PIRSF002089	PIRSF025230	PIRSF031080
Domain architecture	0.5555	0.6545	0.8889
Functional GO term	0.1935	1	1

Calculating the weighted Jaccard distances of these groups against protein families in PIRSF, the weighted Jaccard distance of the biggest group is very close to PIRSF002089 (tumor suppressor p53) in Table 1. Among 111 sequences in the group, we selected 35 reviewed sequences for the generation of specific PSSMs. We then chose 76 sequences as a positive dataset, and 26 RRM (Rna Recognition Motif)s and 127 non-nucleic binding proteins as a negative dataset.

Shown in Figure 2, the specific PSSMs successfully identified the conserved patterns related to p53 in a sequence of testing dataset. X-axis represents the position of amino acid, and Y-axis represents residue score. Since Figure 2 (b) shows only the distribution of residue scores, we filtered residue scores using smoothing filter for the identification of conserved regions. Shown in Figure 2 (c), the conserved regions match the domain regions identified by BLAST. This indicates that our pipeline can predict the conserved regions such as domains or motifs in a protein sequence using specific PSSMs.

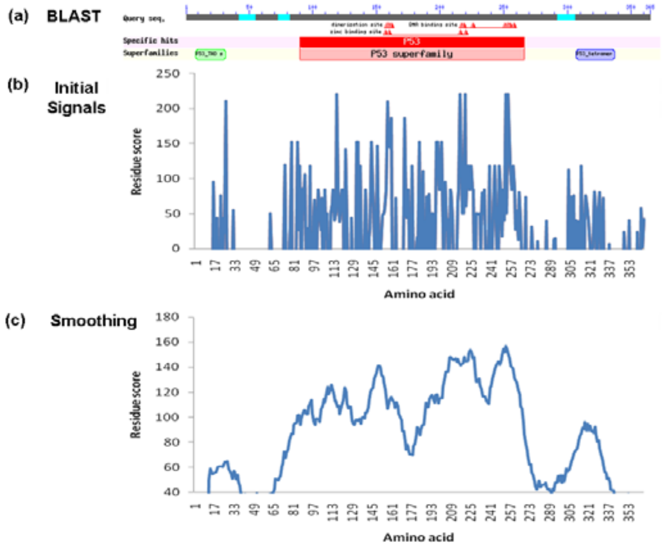

Figure 2: The identification of conserved regions in a protein sequence using specific PSSMs. (a) domain regions predicted by BLAST (b) the distribution of residue scores (c) the conserved regions predicted by the new pipeline using specific PSSMs.

To test the accuracy of the prediction, we selected 76 sequences as a positive dataset and 155 sequences (RRM: 37, non-nucleic binding protein: 127) as a negative dataset. Then, we calculated sensitivity, specificity, and accuracy using equation (5), (6), and (7). The pipeline did not identify any conserved region in proteins not related to p53 proteins, the sensitivity, specificity, and accuracy of the pipeline are 100%.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \times 100\% \tag{5}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}} \times 100\% \tag{6}$$

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \times 100\% \tag{7}$$

Table 2. The sensitivity, specificity, and accuracy of positive and negative datasets about p53 proteins.

TP	FP	TN	FN	Sensitivity (%)	Specificity (%)	Accuracy (%)
76	0	155	0	100	100	100

For further validation, we identified p53 related proteins in human proteome. In fact, after downloading 86,718 proteins from International Protein Index (IPI) site, we did high-throughput analysis of these proteins using specific PSSM for p53 proteins. Among 86,718 sequences, we identified 12 of p53, 13 of p63, and 10 of p73 proteins.

Even though we used specific PSSMs for p53 proteins, our pipeline identified tumor-related proteins including p63 and p73 proteins in the proteomic analysis. In 2009, Dr. Vladimir’s group proved that they are evolutionary close to each other and they have very similar structures [17]. Because of that, our pipeline captured all of p53, p63, and p73 proteins in human proteome. Therefore, these two experiments suggest that, generating functional specific PSSMs for sequences with similar functional characteristics is able to identify new proteins that have similar characteristics.

4 Conclusions

Many protein databases use homology-based approaches to build protein families and improve their protein annotations. While these protein families provide important resources for biologists to predict structures and functions of novel proteins, it is not clear how well those protein families capture the characteristics of proteins. Generally, we use sequence similarity, domains (or domain architectures), and GO terms to annotate proteins. Since protein families are used to infer protein annotations, proteins from the same family should tend to share similar GO terms and domain architectures. The names of biological entities related to these proteins can also be shared.

Based on the above, we add reliable information related to the characteristics of protein families into a pipeline for protein classification. As we use sequences which are collected on the basis of similar domain architectures and functional GO terms for specific PSSMs, these specific PSSMs allow RPS-BLAST to identify proteins which have similar characteristics in human proteome with high accuracy. Thus, this study suggests that additional information such as the entity name, evolutionary distance, domain architecture, and functional GO terms besides sequence similarity is helpful in improving protein classification. Finally, the integration of different methods in different fields into one pipeline can be cornerstone to implement a unified protein classifier.

5 Acknowledgement

This study was supported by National Science Foundation ABI:0845523 and National Institute of Health R01LM009959A1.

6 References

[1] Sander, C. and R. Schneider, Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins, 1991. **9**(1): p. 56-68.

[2] Hilbert, M., G. Bohm, and R. Jaenicke, Structural relationships of homologous proteins as a fundamental principle in homology modeling. Proteins: Struct. Funct. Genet., 1993. **17**: p. 138-151.

[3] Rychlewski, L., et al., Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci, 2000. **9**(2): p. 232-241.

[4] Altschul, S.F., et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 1997. **25**(17): p. 3389-3402.

[5] Karplus, K., SAM-T08, HMM-based protein structure prediction. Nucleic Acids Res, 2009. **37**(Web Server issue): p. W492-W497.

[6] Gribskov, M., A.D. McLachlan, and D. Eisenberg, Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci U S A, 1987. **84**(13): p. 4355-4358.

[7] Marchler-Bauer, A., et al., CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res, 2002. **30**(1): p. 281-283.

[8] Liu, H., et al., BioThesaurus: a web-based thesaurus of protein and gene names. Bioinformatics, 2006. **22**(1): p. 103-105.

[9] Wu, C.H., et al., The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res, 2006. **34**(Database issue): p. D187-D191.

[10] Suzek, B.E., et al., UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics, 2007. **23**(10): p. 1282-1288.

[11] Wu, C.H., et al., The iProClass integrated database for protein functional analysis. Comput Biol Chem, 2004. **28**(1): p. 87-96.

[12] Mi, H., et al., The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res, 2005. **33**(Database issue): p. D284-D288.

[13] Nikolskaya, A.N., et al., PIRSF family classification system for protein functional and evolutionary analysis. Evol Bioinform Online, 2006. **2**: p. 197-209.

[14] Couto, F.M., M.J. Silva, and P.M. Coutinho, Measuring semantic similarity between Gene Ontology terms. Data & Knowledge Engineering 2006. **16**: p. 15.

[15] Retief, J.D., Phylogenetic analysis using PHYLIP. Methods Mol Biol, 2000. **132**: p. 243-258.

[16] Zhou, T., et al., An approach for determining evolutionary distance in network-based phylogenetic analysis. ISBRA'08 Proceedings of the 4th international conference on Bioinformatics research and applications, 2008.

[17] Belyi, V.A. and A.J. Levine, One billion years of p53/p63/p73 evolution. Proc Natl Acad Sci U S A, 2009. 106(42): p. 17609-17610.