# Weak Homology Detection by Profile-Profile Comparison

**Masashi Fujita**
fujita@kuicr.kyoto-u.ac.jp

**Tatsuya Akutsu**
takutsu@kuicr.kyoto-u.ac.jp

**Susumu Goto**
goto@kuicr.kyoto-u.ac.jp

**Minoru Kanehisa**
kanehisa@kuicr.kyoto-u.ac.jp

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto
611-0011, Japan

**Keywords:** structure prediction, distant homology, profile, secondary structure

## 1 Introduction

Experimentally determined protein tertiary structures are currently increasing in an enormous rate. The growth of the structure database makes structure prediction methods based on known tertiary structures more efficient. Detecting weak homology is crucial to extend applicable area of these methods. For this purpose, a profile or PSSM (position-specific score matrix) derived from multiple alignments of protein families has been proved to be quite effective. One of the most successful examples of profile-based methods is PSI-BLAST. But PSI-BLAST simply compares profiles with amino acid sequences. A more sensitive search must be achieved by comparing one profile with another profile, and several studies reported that sensitivity was improved when profile pairs were aligned [4, 5]. We combined this profile-profile comparison method and predicted secondary structure information, and succeeded in significantly improving distant homology recognition performance in comparison with PSI-BLAST.

## 2 Method and Results

### 2.1 Data Sets

In order to define distantly related protein pairs, we used the structural classification of SCOP 1.59. If two proteins were classified into the same superfamily but belonged to different families, we defined these proteins are "distant homologues."

A profile database of structure-determined proteins was constructed as follows. We extracted 4370 sequences from SCOP so that no pairs of proteins shared more than 40% identity. Each of these sequences was searched against nr-aa databae in the GenomeNet using PSI-BLAST, iterating less than 10 times. We also assigned secondary structures to each proteins using STRIDE [2].

Query profiles were constructed by a similar procedure. Superfamilies that consist of more than two families were extracted, and representative proteins from each of these superfamilies were selected as a test set. Each sequence of the test set was translated into a profile by running PSI-BLAST, and secondary structures were predicted with PSIPRED [3].

### 2.2 PSSM Similarity Score

To define similarity of two profiles, PSSM was translated into 20 dimensional vectors of observed amino acid frequencies, and log-odds of the inner product of two vectors were defined as a profile similarity score. If the secondary structure of the template protein and the predicted secondary structure of the query protein were matched, a positive constant was added to the profile similarity score.

### 2.3 Database Search and Performance Assessment

Each profile in the test set was searched against the profile database using global dynamic programming, and the "coverage and error" [1] test was carried out (Fig. 1.) To define empirical reliability, several scoring methods were compared each other, in global alignment the "score gap" method performed best. Based on this result, empirical reliability of structure prediction was defined for both our method and PSI-BLAST. With 1% error rate, our method could correctly identify 151 distant homologues, although PSI-BLAST could detect only 61. In practice, query sequences are often multidomain proteins. So we also tested "glocal" alignment (global with respect to template, local with respect to query) with 34 multidomain query profiles. By a normalized score with the template length, our method performed better than PSI-BLAST again (Fig. 2.)

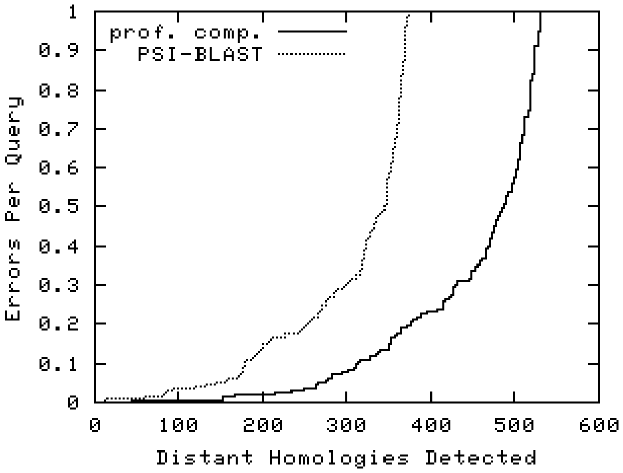

Figure 1: Coverage and error result for global alignment with single domain queries.

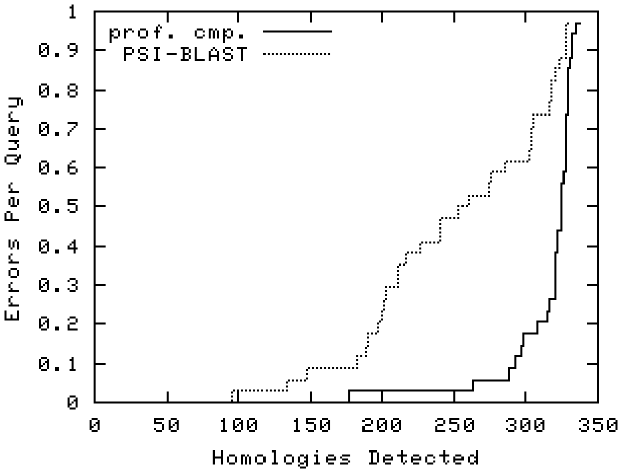

Figure 2: Coverage and error result for glocal alignment with multidomain queries.

## References

[1] Brenner, S.E., Chothia, C., and Hubbard, T.J., Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships, *Proc. Natl. Acad. Sci. USA*, 95:6073–6078, 1998.

[2] Frishman, D. and Argos, P., Knowledge-based protein secondary structure assignment, *Proteins*, 23(4):566–79, 1995.

[3] Jones, D.T., Protein secondary structure prediction based on position-specific scoring matrices, *J. Mol. Biol.*, 292:195–202, 1999.

[4] Rychlewski, L., Jaroszewski, L., Li, W., and Godzik, A., Comparison of sequence profiles. Strategies for structural predictions using sequence information, *Protein Sci.*, 9(2):232–241, 2000.

[5] Yona, G. and Levitt, M., Within the twilight zone: A sensitive profile-profile comparison tool based on information theory, *J. Mol. Biol.*, 315(5):1257–1275, 2002.