Performance Comparison between Profiles from Hidden Markov Models and from Pseudocount Method

Makihiko Sato\textsuperscript{1}
\texttt{maki@maebashi-it.ac.jp}

Katsuhisa Horimoto\textsuperscript{2}
\texttt{khorimot@ims.u-tokyo.ac.jp}

\textsuperscript{1} Department of Systems and Information Engineering Graduate School, Maebashi Institute of Technology, 460-1 Kamisatori-Cho, Maebashi-City, 371-0816, Japan

\textsuperscript{2} Laboratory of Biostatistics, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

Keywords: profile-profile comparison, hidden Markov model, remote homolog detection

1 Introduction

Remote homolog detection is one of the main issues in bioinformatics, and the profile-profile comparison is one of the most effective approaches. Recently, we have proposed a novel profile, match-node profile, constructed via profile HMM [6]. To evaluate the performance of match-node profile, we construct the profile by two popular HMM packages, HMMER [2] and SAM [4], and compared the performance of match-node profile for the remote homolog detection with that of a representative profile-profile method, COMPASS [5], in which the profile is constructed by a pseudocount method.

2 Methods and Results

2.1 Match-Node Profile

The profile of a representative HMM is composed of three types of nodes; match node, delete node, and insert node [2, 4]. The three nodes in the HMM profiles represent the probability distribution of amino acid residues, deletions, and insertions in each site of multiple alignment, respectively. We construct the profile, the match-node profile, from the series of probability distribution in the match-nodes, which can describe essential characteristics of the multiple alignment [6].

2.2 Data Set

The PDB40D [9] was adopted as a data set to perform a benchmark test. The sequences in PDB40D (after purification of shorter or rare chains, 4289 protein chain from one complete domain were remained) corresponded with those in SCOP [10] to evaluate the performance of our approach for fold recognition and alignment accuracy: 3327 chains in family level, 537 superfamily level, 193 fold level, and 232 class level.

2.3 Profile Construction and Scoring

For construction of match-node profile, we adopt HMMER [2] and SAM [4]. In the case of HMMER, each sequence in PDB40D was searched against KIND [3] by PSI-BLAST [1], the HMM profiles were generated from the multiple alignments thus obtained, and the match-node profiles were extracted from them. The multiple alignments were used for the profiles of COMPASS [5]. In the SAM, the built-in methodology, SAM-T2K with w0.5 script, was applied for construct the HMM profiles against NR database from NCBI [11], according to the SAM-T2K default, and the match-node profiles were constructed from them. As for the scoring of similarity between profiles, we adopted the log average scoring based on Bayesian theory, which is one of the most suitable for the profile-profile comparison [8]. Thus, we performed three methods in this study: two log average methods for the match-node profile from HMMER and SAM, and COMPASS as a representative profile-profile method.

2.4 Performance Comparison
Figure 1 shows the fold recognition performance in the three SCOP levels. Each number indicates percentage of correctly predicted targets.

Figure 2 shows the alignment accuracy; the profile alignment and the structural alignment by CE [7] are compared, and the mean length is estimated as the alignment accuracy from the number of residue pairs that are correctly aligned in both alignments. In this case, the match-node profile methods show the superiority to COMPASS in the accurate alignment length.

3 Discussion

Overall, two match-node profile methods and COMPASS are equal in the performance of remote homolog detection. Note that the match-node profile methods have an advantage in the quality of alignment. The detailed results and the merits and pitfalls of methods will be further discussed.

References