SIMPRO: simple protein homology detection method by using indirect signals

Inkyung Jung¹ and Dongsup Kim¹,²,*

¹Department of Bio and Brain Engineering and ²KAIST Institute for BioCentury, KAIST, Daejeon 305-701, South Korea

ABSTRACT

**Motivation:** Detecting homologous proteins is one of the fundamental problems in computational biology. Many tools to solve this problem have been developed, but development of a simple, effective and generally applicable method is still desirable.

**Results:** We propose a simple but effective information retrieval approach, named SIMPRO, to identify homology relationship between proteins. The key idea of our approach is that by accumulating and comparing indirect signals from conventional homology search methods, the search sensitivity can be increased. We tested the idea on the problem of detecting homology relationship between Pfam families, as well as detecting structural homologs based on SCOP and found that our method achieved significant improvement. Our results indicate that simple manipulation of conventional homology search outputs by SIMPRO algorithm can remarkably improve homology search accuracy.

**Contact:** kds@kaist.ac.kr

1 INTRODUCTION

The explosion of raw sequence data requires efficient computational methods for homology search. Detecting homologous proteins is valuable for predicting structure and function of unknown proteins. The functional relationship between proteins has been studied by probing evolutionary relationship and using specific structural or functional information. When homologous proteins have weak sequence similarity with a query sequence, detecting them with high sensitivity (SE) is quite challenging. Many effective approaches have been developed including profile–profile alignment methods combined with various machine learning algorithms, yet still there remains ample room for improvement.

In general, the most successful methods for homology detection typically rely on profile–sequence or profile–profile alignment methods. Some programs utilize structural information as well when available. The examples are hidden Markov model (HMM) (Karpplus et al., 1999), PSI-BLAST (Altschul et al., 1997), COMPASS (Sadreyev and Grishin, 2003, 2008; Sadreyev et al., 2007), COACH (Edgar and Sjoland, 2004) and HHsearch (Soding, 2005). The other examples that utilize structural information are PROSPECT (Kim et al., 2003; Xu and Xu, 2000), and ProfNet (Ohlson and Elofsson, 2005). PRC (Madera, 2008) is also a new scoring scheme which aligns profile HMM of protein families and well recognizes distance homology relationships.

Recently, introducing support vector machine (SVM) resulted in remarkable performance enhancement in remote homology searches. Examples are SVM-Pairwise (Liao and Noble, 2003), and profile–profile alignment with SVM (Han et al., 2005). Moreover, several kernel methods such as local alignment kernels (Saigo et al., 2004), profile-based direct kernels (Rangwala and Karypis, 2005) and cluster kernels (Weston et al., 2005) were proposed to develop more powerful remote homology detection methods. In addition, applying new feature extraction method such as non-negative matrix factorization (NMF) to profile–profile alignment features increased the performance of fold recognition significantly (Jung et al., 2008).

Despite high performance of profile-based SVM methods in remote homology detection, these methods require extensive training, and thereby simple and general algorithms that do not require extensive training have been pursued. In Rankprot (Weston et al., 2004) and distance-profile methods (Ku and Yona, 2005), they suggested a simple comparison process using protein pairwise similarities. Recently, a new method named SCOOP (Bateman and Finn, 2007) was reported. By considering common sequence matches between two Pfam HMM profile search results, SCOOP detected protein superfamily relationship more accurately than more elaborated methods such as HHsearch. In addition, methods based on simple reprocessing of PSI-BLAST results have been suggested; Combined E-values generated during different PSI-BLAST iteration steps eliminated false positives (Lee et al., 2008), and PSI-BLAST search for consensus sequences gave a significantly improved result (Przybylski and Rost, 2007, 2008).

In this work, we developed a new search algorithm utilizing outputs from conventional homology search methods such as PSI-BLAST and HHSearch. Our key idea is that by integrating indirect signals between proteins, we can detect homologous proteins more accurately even if those signals are too weak to infer correct homology relationships between proteins by conventional homology search methods. There are possibly many different ways to implement our idea, but it was found that a simple implementation described in this article greatly increased the performance of homology search.

SIMPRO and SCOOP are similar in that both measure a similarity between two proteins by using indirect match results. In SCOOP, they assumed that if search outputs of two family profiles shared many matches in common they could be related. In SIMPRO,
we assumed that if a certain family matches with a number of significantly matched families for a given query, the certain family is more likely to be related with the query family. The main difference between SIMPRO and SCOOP is that in SIMPRO homology search accuracy is increased by extracting hidden signals from significant hits for a given query, whereas in SCOOP by comparing search outputs including both significant and insignificant hits. Figure 1 is a schematic diagram of both methods. For SIMPRO, as shown in Figure 1A, the query Family A significantly matches with Families B, C, D and F. However, Family F does not significantly match with family B, C and D, while Family E significantly matches with Families C and D. In this case, we can infer that Family F is unrelated to the query Family A, while Family E is related (A). For the SCOOP they consider if two families share many common matched sequences, both families could be related (B).

2 METHODS AND ALGORITHMS

2.1 SIMPRO scoring function

A procedure of accumulating the indirect signals from conventional homology search algorithms implemented in SIMPRO is illustrated in Figure 2. Suppose that a protein database is composed of either N protein sequences or protein family profiles (for example, Pfam or SCOP databases), and we try to search for homologs (related proteins or protein families) of a query sequence \( p_k \) using a certain homology search algorithm (for example, PSI-BLAST, HHsearch or profile–profile alignment). Initial search using a chosen search algorithm would give similarity scores, typically E-values or probabilities, to all or potentially homologous proteins in the database (Step 1). However, since the algorithm is not perfect, lower ranks would be given to some proteins (for example, T6 in Fig. 2), despite they are homologous to \( p_k \).

The key idea of SIMPRO is that all proteins homologous to \( p_k \) should have high similarity scores not only with \( p_k \) but also with each other; therefore if those scores of a certain hit are consistently high, then it indicates that this hit is indeed a true homolog. Accumulation of all those indirect signals (Step 2) would give more proper rankings to the proteins in the database (Step 3).

More detailed implementation of SIMPRO is as follows. For a given query protein sequence \( p_k \), let \( \text{sim}(p_j; p_k) \) be the similarity score of the \( j \)-th protein \( p_j \) in the database. Note that \( p_j \) and \( p_k \) can be either protein sequence or protein family profile. Then, the indirect signal for \( p_j \) described in the above paragraph can be written in a simple mathematical form as

\[
\frac{1}{n} \sum_{j=1}^{n} \text{sim}(p_j; p_k)
\]

where the summation is over all \( p_j \)'s in the database whose \( \text{sim}(p_j; p_k) \) is greater than a certain threshold value, and \( n \) is the number of those proteins. Here we assumed that only those proteins with sufficient large similarity scores were considered as significantly matched hits, and therefore contained useful indirect signals. It should be noted that \( \text{sim}(p_j; p_k) \) is the similarity score of \( p_k \) when \( p_k \) is treated as a query. For the computational efficiency, all \( \text{sim}(p_j; p_k) \)'s for a given database can be pre-calculated and stored. Next, although there may be many alternative, potentially more accurate ways to define a new similarity score, we simply defined the combined similarity score of \( p_k \), \( S_{\text{comb}}(p_k; p_j) \), as a linear combination of the direct signal and the indirect signal,

\[
S_{\text{comb}}(p_k; p_j) = \text{sim}(p_j; p_k) + \frac{1}{n} \sum_{j=1}^{n} \text{sim}(p_j; p_k).
\]

To examine the effect of different weight between the direct and the indirect signal, we measured the performances after selecting various combinations of relative weights. The results indicated that for Pfam family classification the new scoring system greatly increased the accuracy of detecting correct relationship between Pfam families. To further increase the accuracy, we first...
defined the ‘signal profile’ of \( p_i \), \( p_j \) as an \( N \)-dimensional vector whose \( k \)-th element is \( S_{\text{comb}}(p_i; p_j) \) as follows:

\[
P_k = \begin{pmatrix}
S_{\text{comb}}(p_1; p_k) \\
\vdots \\
S_{\text{comb}}(p_N; p_k)
\end{pmatrix}
\]

where a majority of elements were simply zero. After we constructed signal profiles for \( p_i \), we measured the similarity between the signal profiles by using the following equation,

\[
S(p_i; p_j) = \sum_{k=1}^{N} S_{\text{comb}}(p_i; p_j) \times (p_k \cdot p_k)
\]

where \( P_k \) is the signal profile of \( p_k \) when \( p_k \) is treated as a query, and all \( P_k \)’s can be pre-calculated and stored. Note that because a majority of elements of signal profile vectors are zero and we only need the dot products of a pair of vectors, the typical size of the vector we need to create is fairly small.

The scoring system was designed to give a high score when the two proteins show similar tendency in similarity scores with respect to the other proteins in the database. This second step is conceptually similar to scoring schemes of SCOOP (Bateman and Finn, 2007) or distance profile (Ku and Yona, 2005) in that those methods measure how many common matches exist between the two proteins.

2.2 Tests and datasets

We tested SIMPRO on the problem of detecting homology relationship between Pfam families, and detecting structural homologs based on SCOP. For Pfam classification problem, we used Pfam version 22.0 that contained 10,255 families. Among them, 1808 families classified into 283 clans were selected. For this problem, the similarity score, \( \text{sim}(p_i; p_k) \), was defined as—\( -\log(E\text{-value/100}) \). Naturally, we assumed that \( \text{sim}(p_i; p_k) \geq 0 \) implying that all protein pairs with \( E\text{-value} \geq 100 \) were assumed to be non-related, and therefore their similarity scores were set to 0. We calculated \( E\)-values using HHsearch 1.2.0 with default parameters. To define the significantly matched hits, we tested several values (0.1, 0.5, 1, 5, 50 and 100) and then selected 50 (for Pfam classification) and 0.5 (for SCOP classification) as the threshold value for the significant hit. Our definition of the similarity scores and we found that for Pfam classification problem HHsearch with \( E\)-values performed better than with the probability scores, and vice versa for SCOP classification problem. Because of all these circumstances, we defined the similarity scores as probability score and \( \exp(-E\text{-value/100}) \) for HHsearch and PSI-BLAST, respectively. When we used HHsearch results, we set the threshold value for the highly ranked proteins as a probability score of 0.5. In addition, when the number of highly ranked proteins was less than 10, we selected top 10 proteins as highly ranked proteins. For PSI-BLAST, we set the threshold value as an \( E\)-value 0.5.

2.3 Performance assessment

Performances were measured by calculating the receiver operating characteristic (ROC) scores (Gribskov and Robinson, 1996). ROC score is defined as the areas under the ROC curves, the plot of true positives as a function of the number of false positives. Also, we measured the ROC\(_{50}\) score, which is the area under the ROC curve up to the first 50 false positives. For Pfam family classification, we defined a true match when two families are in the same clan, while when two families in different clans we considered as a false match. For the fold recognition problem, proteins in the same fold, but different superfamily are considered as homologous proteins. For remote homology detection problem, proteins in the same superfamily but different families are considered as homologous proteins. In both cases, proteins in the different folds are defined as non-homologous proteins.

2.4 Benchmarking of profile based SVM method for SCOP classification

For Pfam classification problem, we benchmarked SCOOP. We downloaded the results from the web site (http://www.sanger.ac.uk/Software/analysis/scoop). We benchmarked LA-kernel and SW-PSSM for fold recognition and remote homology search. LA-kernel is SVM based method using a string alignment kernel (downloaded from http://sunflower.kiwi.uct.ac.za/hits/hiroto/project/homology.html). The best performing method, SW-PSSM is a profile-based local alignment kernel method (downloaded from http://bioinfo.cs.unn.edu.au/supplements/profile-kernels). For the LA-kernel we used version 0.3.2 with \( \beta = 0.05 \) (recommended value). In case of SW-PSSM, we used the best-performing parameter set reported in the original paper (gap opening = 3.0, gap extension = 0.75, zero-shift = 1.5). To compare our method with those methods we modified SIMPRO output scores. Those methods give a score measuring that a certain protein may belong to a specific superfamily (or fold), while our method gives a score measuring that two proteins may belong to the same superfamily (or fold). Therefore, we used the mean value of output scores of the reference set in a specific superfamily (or fold).

3 RESULTS

3.1 Validation of SIMPRO for Pfam family matching

We investigated the ability of SIMPRO to classify Pfam families and compared its performance with that of a similar comparison method, SCOOP (Bateman and Finn, 2007). SIMPRO and SCOOP are similar in that both measure a similarity between two families by using indirect match results. Figure 3 shows the performance of SIMPRO compared with those of SCOOP and HHsearch results. In this figure, we drew using blastpff 1.2.1 to nr database with the options ‘j=5’ (the number of iterations) and ‘h=0.001’ (\( E\)-value threshold for including sequences) which were recommended parameters in the original PSI-BLAST paper (Schaffer et al., 2001). Because the datasets were designed to be deliberately difficult, there were not many significantly matched pairs that had small \( E\)-values. In addition, while PSI-BLAST reported the search outputs with \( E\)-values, HHsearch reported the search results with both \( E\)-values and probability scores and we found that for Pfam classification problem HHsearch with \( E\)-values performed better than with the probability scores, and vice versa for SCOP classification problem. Because of all these circumstances, we defined the similarity scores as probability score and \( \exp(-E\text{-value/100}) \) for HHsearch and PSI-BLAST, respectively. When we used HHsearch results, we set the threshold value for the highly ranked proteins as a probability score of 0.5. In addition, when the number of highly ranked proteins was less than 10, we selected top 10 proteins as highly ranked proteins. For PSI-BLAST, we set the threshold value as an \( E\)-value 0.5.
I. Jung and D. Kim

Fig. 3. ROC curves comparing SIMPRO to various methods for the Pfam classification. The graph shows the cumulative number of true matches that are found with increasing number of false matches. Here, HHsearch(Prob) and HHsearch(E-value) indicate the results of probability score and E-value of HHsearch output, respectively. SIMPRO is final scores of our method and SIMPRO (combined) is the results without calculating signal profile similarities.

various types of SIMPRO results. The line labeled by ‘SIMPRO’ is the results of our method using the scores \( S(p_k; p_i) \), and the line labeled by ‘SIMPRO (combined)’ is the results based on the scores without calculating the signal profile similarities, i.e. \( \hat{S}_{\text{comb}}(p_k; p_i) \). From these results, we notice that SIMPRO and SCOOP have overall better performance than HHsearch results. Especially, the fact that the performance of our new combined similarity scores, SIMPRO (combined), is better than HHsearch supports that using only indirect signals in our scheme can improve the ability to recognize homology relationships. Through those results, we can conclude that integrating indirect signals effectively reduce noises, which results in performance enhancement. Furthermore, SIMPRO has achieved better results than SCOOP. For example, at the 100 false matches, the number of true matches of SIMPRO is 3439, while that of SCOOP is 3040.

In addition, the number of true matches in SIMPRO increases faster than SCOOP. Those results indicate that SIMPRO is more specific, and has larger coverage than SCOOP. However, due to the each advantage of both methods, we can expect combining results of SIMPRO and SCOOP gives high SP and large coverage.

3.2 Evaluation of the SIMPRO for identification of SCOP relationship

To evaluate the general applicability of SIMPRO, we applied it to SCOP classification problem. Conventionally, there are two types of problems; one is fold recognition and another is remote homology detection. We constructed two protein sets for each problem. One is for a query set and another is a reference set. Based on the reference set, we retrieved indirect information and constructed signal profiles.

For the fold recognition, we only considered the situation where a query set and a reference set do not share the same superfamily. As a result, we created a reference set (1717) and a query set (686 proteins for 38 folds). For the remote homology detection we created a reference set (1573) and a query set (805 proteins and 75 superfamilies), where no proteins in the reference set shared the same SCOP families with any protein in the query set.

Figure 4 shows ROC curves of various methods. In fold recognition, at the 10,000 false matches the number of true matches of SIMPRO with HHsearch (or PSI-BLAST) is 20,982 (or 7766), while that of HHsearch (or PSI-BLAST) is 10,422 (or 2643). In remote homology detection, at the 10,000 false matches the number of true matches of SIMPRO with HHsearch (or PSI-BLAST) is 22,999 (or 12,643), while that of HHsearch (or PSI-BLAST) is 19,711 (or 8823). From Figure 4 it is clear that SIMPRO greatly increases the original search methods.

The results are also summarized in Table 1. For fold recognition the mean ROC\(_{50}\) score of SIMPRO after applying HHsearch (or PSI-BLAST) is 0.36 (or 0.12), while that of HHsearch (or PSI-BLAST) is 0.23 (or 0.07). Furthermore, 39% of a query set have >0.50 ROC\(_{50}\) scores when applying SIMPRO to HHsearch results, whereas that of HHsearch results is only 12%. Moreover, in terms of ROC scores SIMPRO improves HHsearch (or PSI-BLAST) results from 0.79 (or 0.60) to 0.86 (or 0.65). For remote homology detection, the mean ROC\(_{50}\) scores after applying SIMPRO to HHsearch and PSI-BLAST results are 0.71 and 0.38, respectively. On the contrary, those of HHsearch and PSI-BLAST results are 0.63 and 0.31, respectively. For 24% (or 13%) proteins in a query set, the ROC\(_{50}\) scores are increased by more than 0.2 after applying SIMPRO to HHsearch (or PSI-BLAST) results. The statistical tests (paired \(t\)-test) on the results
Table 1. The mean ROC50 and ROC scores at the fold and superfamily level

<table>
<thead>
<tr>
<th>Method</th>
<th>Fold recognition</th>
<th>Remote homolog detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIMPRO(HHsearch)</td>
<td>0.36</td>
<td>0.71</td>
</tr>
<tr>
<td>HHsearch</td>
<td>0.23</td>
<td>0.63</td>
</tr>
<tr>
<td>SIMPRO(PSIBLAST)</td>
<td>0.12</td>
<td>0.38</td>
</tr>
<tr>
<td>PSI-BLAST</td>
<td>0.07</td>
<td>0.31</td>
</tr>
</tbody>
</table>

shown in Table 1 indicate that both performance enhancements are statistically significant with P-values of near zero. From these results, it is easy to see that SIMPRO with HHsearch results is more effective than SIMPRO with PSI-BLAST results. This indicates that the effectiveness of SIMPRO algorithm depends on the quality of the results of original search method. Therefore, it is expected that performance degradation can occur in some cases where original search results are not good enough to provide any meaningful signals.

Next, we investigated which types of proteins benefitted the most from SIMPRO method. First, we investigated performance variation among different SCOP classes. In terms of ROC50 scores, both class a and b achieved the most improved results after applying SIMPRO to HHsearch results; SIMPRO increased the performance by >20% from original results. In particular, for fold recognition problem, the mean ROC50 score of HHsearch was 0.27 but after applying SIMPRO the performance was improved to 0.53 in class b. On the contrary, classes d and f achieved the least improvement. Reason for the performance variation is not clear; it could depend on many factors such as the accuracy of original search algorithm, the size of each SCOP class, the size of reference proteins, etc. However, close examination revealed that there was a correlation between the performance improvement and the size of reference proteins which were included in the same fold (or superfamily) with the query protein. For example, the Pearson correlation was 0.4 when SIMPRO was applied to HHsearch results for remote homolog detection problem. It appears that if there are more homologs in the reference set, integrating indirect signals can be more effective, which results in higher performance improvement.

To better understand how SIMPRO algorithm recognizes remote homologs, we consider one of specific examples, the homology relationship between the insulin receptor substrate 1, IRS-1 from Human (d1qga2), and the UNC-89 from Nematode (d1fhoa_). Both proteins are included in the PH domain-like superfamilies (b.55.1). With HHsearch the UNC-89 was ranked as 986th among 1573 proteins for the query of IRS-1. However, SIMPRO increased the rank of UNC-89 to 28th with score of 0.61 [positive predictive value (PPV)] > 0.95, see next paragraph for the discussion on PPV]. This is because there are enough number of intermediate sequences, providing high-quality indirect signals. For example, among the top 10 highly ranked proteins for IRS-1, seven proteins (Tapp1, Dynamin, PEPP1, etc.) gave high rank to UNC-89. Thereby, even though UNC-89 was recognized as a distant protein by HHsearch, by accumulating indirect signals through intermediate sequences

Table 2. SP, SE and PPV for various problems as a function of SIMPRO cutoff score

<table>
<thead>
<tr>
<th>Score</th>
<th>Fold recognition</th>
<th>Remote homology detection</th>
<th>Pfam classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SP</td>
<td>SE</td>
<td>PPV</td>
</tr>
<tr>
<td>1</td>
<td>1.00</td>
<td>0.04</td>
<td>0.79</td>
</tr>
<tr>
<td>0.8</td>
<td>1.00</td>
<td>0.07</td>
<td>0.82</td>
</tr>
<tr>
<td>0.6</td>
<td>1.00</td>
<td>0.13</td>
<td>0.83</td>
</tr>
<tr>
<td><strong>0.4</strong></td>
<td><strong>1.00</strong></td>
<td><strong>0.21</strong></td>
<td><strong>0.80</strong></td>
</tr>
<tr>
<td>0.2</td>
<td>0.99</td>
<td>0.29</td>
<td>0.68</td>
</tr>
<tr>
<td>0.1</td>
<td>0.99</td>
<td>0.36</td>
<td>0.59</td>
</tr>
<tr>
<td>0.05</td>
<td>0.97</td>
<td>0.45</td>
<td>0.49</td>
</tr>
<tr>
<td>0.01</td>
<td>0.91</td>
<td>0.65</td>
<td>0.27</td>
</tr>
<tr>
<td>0.001</td>
<td>0.69</td>
<td>0.85</td>
<td>0.13</td>
</tr>
</tbody>
</table>

SP = TN/(TP+TN), SE = TP/(TP+FN) and PPV = TP/(TP+FP), where TN, TP, FP and FN represents true negatives, true positives, false positives and false negatives, respectively.

SIMPRO recognized UNC-89 as a homologous protein for IRS-1. This example shows exactly how SIMPRO works to detect remote homologous proteins which are not recognized by conventional methods.

Finally, we investigated how consistent and reliable SIMPRO scores are depending on different types of databases and problems. We calculated SP, SE and PPVs for fold recognition, remote homology detection and Pfam classification (Table 2). At the cutoff score of 0.4 (bold line), 80%, 89% and 99% of matches are true for fold recognition, remote homology detection and Pfam classification, respectively. As expected, the database consists of more proteins with closer relationships, PPV at a certain cutoff score increases.

In conclusion, above findings supports that SIMPRO is generally applicable to various types of homology search problems, implying that if search outputs are reasonably good, any kind of search algorithm can be improved by employing SIMPRO.

3.3 Performance comparison with SIMPRO and kernel-based SVM method

We benchmarked several kernel methods based on profile–profile alignments, since it was shown that kernel-based SVM methods were the most outperforming algorithms in fold recognition and remote homology detection. The results of SIMPRO are summarized in Table 3 along with those of LA-kernel (Saigo et al., 2004) and SW-PSSM (Rangwala and Karypis, 2005), which are the best performing kernel-based methods. For fold recognition (or remote homology detection) we measured the mean ROC50 and ROC scores for 38 folds (75 superfamilies). The results indicate that the performance of new method is comparable with profile-based SVM methods even though LA-kernel and SW-PSSM are much elaborated methods that require extensive training with SVM. The performance of SIMPRO using PSI-BLAST results is better than that of LA-kernel, but less than that of SW-PSSM. At the fold level, the mean ROC50 score of SIMPRO by using PSI-BLAST results is 0.35, while that of LA-kernel and SW-PSSM is 0.34 and 0.56, respectively. However, when we used HHsearch results, the performance was increased to 0.61 for the mean ROC50 score and 0.93 for the mean ROC score, which
Table 3. Performance comparison among the SIMPRO, LA-kernel and SW-PSSM

<table>
<thead>
<tr>
<th>Method</th>
<th>Fold recognition</th>
<th>Remote homolog detection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean ROC _50</td>
<td>Mean ROC</td>
</tr>
<tr>
<td>SIMPRO (HHsearch)</td>
<td>0.61</td>
<td>0.88</td>
</tr>
<tr>
<td>SIMPRO (PSIBLAST)</td>
<td>0.35</td>
<td>0.64</td>
</tr>
<tr>
<td>LA-kernel</td>
<td>0.34</td>
<td>0.50</td>
</tr>
<tr>
<td>SW-PSSM</td>
<td>0.56</td>
<td>0.87</td>
</tr>
</tbody>
</table>

were slightly better than SW-PSSM, although it was not statistically significant (paired t-test). We could observe similar performance improvement for remote homology detection as well. Consequently, our results indicate that by combining SIMPRO and other fast search algorithm like HHsearch without any complicated algorithms, distant homologous proteins can be effectively detected.

4 DISCUSSION AND CONCLUSION

The present method is remarkable in the sense that it is simple and computationally costless (806s on an Intel Xeon 8-processor machine with 8 Gb of RAM for entire Pfam classification), and yet shows the outstanding performance improvement for the various kinds of homology detection problems such as Pfam family classification, fold recognition and remote homology detection. We showed that when we applied SIMPRO to the search outputs from PSI-BLAST and HHsearch, the search results were greatly improved. There is no obvious reason not to believe that SIMPRO can be applied to any type of search algorithm and problem. Our tests on three kinds of problems using two types of search outputs all suggest that SIMPRO method is generally applicable, and combining SIMPRO with any type of reasonably good method will achieve higher search SE for homology search.

The performance improvement is attributed to the effective integration of indirect signals from original homology search outputs. Performance degradation in a conventional homology search is mainly originated from improper scoring system, which will achieve higher search SE for homology search. Performance degradation in a conventional homology search is mainly originated from improper scoring system, which gives higher ranking to unrelated proteins, producing many false positives. However, by considering indirect relationships between the proteins in the database, we can reduce the number of false positives, and detect the true negatives from original search results with greater accuracy.

The basic underlying principle of SIMPRO, SCOOP and some other methods is similar; all those methods exploit the similarity relationships between proteins and compare the outputs of conventional homology search methods with achieve higher accuracy. However, SIMPRO is unique in the aspect that it first generates the noise-reduced scores, i.e. $S_{\text{comb}}(p_i,p_j)$, by taking into account the similarity relationships between the proteins in the database, and then further improves the scores by comparing the search outputs, i.e. by calculating the dot products of two signal profile vectors. Moreover, we showed that the present method outperformed those methods.

To conclude, applying SIMPRO produces good results without sophisticated profile–profile alignment or machine learning techniques. We can generalize the present method to determine homology relationship for any given two proteins. Furthermore, we expect that applying the new method to other kinds of biological analysis of prediction problems would achieve high-performance improvement.

ACKNOWLEDGEMENTS

We thank Dr Alex Bateman for advice of benchmarking SCOOP. I.J. thanks Y.L. for many discussions.

Funding: CHUNG Moon Soul Center for BioInformation and BioElectronics (CMSC); Korea Science and Engineering Foundation (grant number: 2008-05442).

REFERENCES


