BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1–2

RANKPROP: a web server for protein remote homology detection

Iain Melvin [1], Jason Weston [1], Christina Leslie [2] and William Stafford Noble [3*]

[1]NEC Laboratories of America, Princeton, NJ, [2]Computational Biology Program, Sloan-Kettering Institute, Memorial Sloan-Kettering Cancer Center, New York, NY, [3]Department of Genome Sciences, Department of Computer Science and Engineering, University of Washington, Seattle, WA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: We present a large scale implementation of the RANKPROP protein homology ranking algorithm in the form of an openly accessible web server. We use the NRDB40 PSI-BLAST all-vs-all protein similarity network of 1.1 million proteins to construct the graph for the RANKPROP algorithm, whereas previously, results were only reported for a database of 108,000 proteins. We also describe two algorithmic improvements to the original algorithm, including propagation from multiple homologs of the query and better normalization of ranking scores, that lead to higher accuracy and to scores with a probabilistic intepretation.

Availability: The RANKPROP web server and source code is available at `http://rankprop.gs.washington.edu`.

Contact: iain@nec-labs.com

1 INTRODUCTION

RANKPROP [Weston et al., 2004] is a network-based inference algorithm for identifying subtle protein sequence similarities, corresponding to remote homology relationships or to structural similarities. The algorithm operates on a *protein similarity network*, a graph in which each node is a protein and each weighted edge connecting two proteins indicates their similarity. Such a network can be built using existing tools such as PSI-BLAST [Altschul et al., 1997].

The key idea of the RANKPROP algorithm is to extract global information from a protein similarity network by propagating outward from a user-specified query protein. Effectively, the algorithm sums over all possible paths from the query to each target protein. Thus, after propagation, the resulting activation scores for each node include global information about that protein's relationship to the query. Ranking proteins by these scores is analogous to performing a database search using a tool such as PSI-BLAST, except that the ranking induced by RANKPROP reflects the global topology of the protein similarity network.

In [Weston et al., 2004], PSI-BLAST is used to measure sequence similarity, and the unnormalized weight for the edge from node i to node j is $W_{ij} = exp(-S_j(i)/\sigma)$, where $S_j(i)$ is the PSI-BLAST E-value assigned to protein i given query j, and the parameter σ is a positive constant. Edges are only included in the network for E-values smaller than a fixed threshold. We obtain a stochastic connectivity matrix M for the protein similarity network by row-normalizing edge weights W_{ij} to obtain transition probabilities: $M_{ij} = W_{ij} / \sum_j W_{ij}$.

Given such a network and a query sequence q, the RANKPROP algorithm is simple to describe. First, all nodes are assigned initial *activation scores* that reflect each target protein's similarity to q. Like the edge weights, these scores are computed from PSI-BLAST E-values using the same equation. At each iteration of the algorithm, the activation score at a given node is replaced by the weighted sum of the scores from all of its incoming edges. The update rule includes a diffusion constant α that controls the rate of diffusion through the network. Formally, we define the initial activation scores as $P_i^0 = \exp(-S_q(i)/\sigma)$. Viewing P^t as the column vector of activation levels at iteration t, the algorithm is given by $P_i^{t+1} = \alpha M P_i^t + P_i^0$ if $P_i \neq q$ and $P_i^{t+1} = 1$ otherwise, where $\alpha \in (0, 1)$. One can show that this iterative procedure converges to a fixed point, which in practice happens in a small number of iterations. The output of the RANKPROP algorithm is a ranking of the nodes in the network according to their final activation values. Proteins that receive a high activation score are linked to the query via many strongly weighted paths and vice versa. A multidomain query protein will produce strong matches to any target protein that contains one or more of the query domains. A single domain query A may connect through a multidomain protein AB to infer a false relationship with B. However, previous work [Weston et al., 2004] has found that as long as the query sequence is connected to many other proteins, then the true homologs will be mutually reinforcing and receive a higher rank.

In this work, we extend the original RANKPROP algorithm in two ways: (1) improving accuracy by propagating simultaneously from proteins that are very closely related to the query, and (2) improving the interpretability of the scores produced by RANKPROP by empirically mapping them to probabilities. The mapped score can be interpreted as the probability that the target protein is a member of the same SCOP superfamily as the query. We also announce the availability of a free web server that allows individual queries against a protein similarity network derived from the NRDB40, comprising 1.1 million targets.

*to whom correspondence should be addressed

Table 1. Ranking Accuracy

method	Family ROC_1	Family ROC_{50}	S-Fam ROC_1	S-Fam ROC_{50}
PSI-BLAST	0.833*	0.851	0.609*	0.628
RankProp SWISSPROT	0.816*	0.906	0.592*	0.725
RankProp NRDB40	0.872	0.923	0.696	0.779*
RankProp+homologs NRDB40	0.884	0.928	0.710	0.775*

2 METHODS

The RANKPROP server uses the PSI-BLAST all-vs-all similarity matrix for NRDB40 provided by the PairsDB website [Heger et al., 2008]. NRDB40 is a subset of the non-redundant sequence database, filtered so that no pairs exhibit $> 40\%$ sequence identity. The NRDB40 collection contains 1.1M sequences.

We generalize the RANKPROP algorithm to accept a set Q of query proteins, rather than a single query protein. To use this extra information we perform propagation as usual, but we constrain the activation scores for all the query points such that they are ranked highly. In particular, we choose the set Q to be all the proteins that have a match with the initial query q with a PSI-BLAST E-value less than 0.001. We then constrain our algorithm to have $P_j = 1 - S_q(j), \ \forall j \in Q$. This modification is useful because, instead of propagating from a single query source node in the graph, we can propagate from several source nodes that all belong to the same family or superfamily that we are searching for.

The original RANKPROP algorithm outputs scalar values that are not directly interpretable. In the new version of the algorithm, we map each RANKPROP score to an estimate of the probability that the corresponding query and target proteins belong to the same structural superfamily. We employ the SCOP database [Murzin et al., 1995] to compute a histogram of empirical frequencies of the activation levels P_i for each protein i. More specifically, we choose bin centers v_k and compute the following quantities: n_k, the number of times P_i falls into bin v_k, and s_n, the number of times that the latter occurs and i is in the same superfamily as the query. We are interested in the value s_k/n_k, which can be interpreted as the probability for each activation value bin of the target being in the same superfamily as the query. We choose the bin centers $v = (0, 0.01, 0.02 \ldots, 0.2, 0.3, \ldots, 1)$, and we enforce monotonicity in the final output by setting $p_i/n_i = p_{i-1}/n_{i-1}$ if $p_i/n_i < p_{i-1}/n_{i-1}$.

3 RESULTS

Table 1 compares our large scale RANKPROP results with PSI-BLAST (using NRDB40 and the same blastpgp parameters as PairsDB: `-j 10 -e 1 -h 0.001 -b 10000 -v 10000`) and the previously published version of RANKPROP (using the SWISSPROT database, 108k proteins). RANKPROP NRDB40 is a straightforward scaling up of the previous RANKPROP algorithm to NRDB40. In addition, RANKPROP+homologs NRDB40 employs the extensions described in Methods. Accuracy is measured following the methodology given in [Weston et al., 2004]: SCOP version 1.59 is split into train and test portions, and hyper parameters are chosen by using the training set. Then, each test protein is treated as a query, and the quality of a method's protein ranking is measured by using the area under the receiver operating characteristic curve, up to the first (ROC_1) or 50th (ROC_{50}) false positive. We report results as average ROC_1 and ROC_{50} scores across all 3083 test proteins.

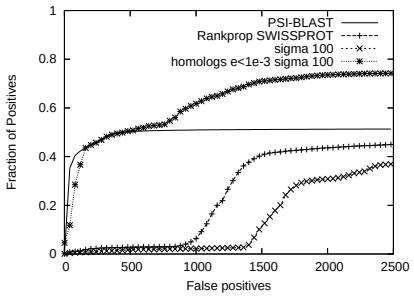

Fig. 1. Combined ROC curve across multiple queries. For each method, search results from 3083 queries were sorted into a single list. The figure plots, for varying thresholds in the ranked list, the fraction of all known homologs (SCOP superfamily members) that fall above the threshold, as a function of the number of non-superfamily members above the threshold.

Using a larger network yields improvements across all four performance metrics, and propagating from multiple queries improves the performance still further. A Wilcoxon signed rank test, corrected for multiple tests, shows that all differences in Table 1 are significant at 0.01, except for the three pairs of methods marked with asterisks.

We also evaluate the performance of RANKPROP using a combined ROC curve across all the queries in our test set, following the protocol of Altschul et al. [1997]. Figure 1 shows the combined ROC curves for RANKPROP NRDB40 (ranked by activation value), RANKPROP+homologs NRDB40 (ranked by probability) and PSI-BLAST (ranked by E-value). Compared to average per-query ROC scores, the combined ROC curve requires that scores are well calibrated from one query to the next. The figure shows that the mapping of RANKPROP scores to probabilities significantly improves the calibration, yielding better performance than PSI-BLAST for all but the first few false positives (across 3083 queries).

The RANKPROP web server first looks for an exact match of the query sequence against the sequences in NRDB40. If such a match is found, the server will retrieve the precomputed PSI-BLAST results from the database and then apply the RANKPROP algorithm. In this case the server takes around 90 seconds to process a query. If the sequence is not found in the database, then the server will run PSI-BLAST first, which on average takes an additional 15 minutes.

FUNDING

This work was funded by NIH award R01 GM074257.

REFERENCES

S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

A. Heger, E. Korpelainen, T. Hupponen, K. Mattila, V. Ollikainen, and L. Holm. Pairsdb atlas of protein sequence space. *Nucleic Acids Research*, 36(D276–D280), 2008.

A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 1995.

J. Weston, A. Elisseeff, D. Zhou, C. Leslie, and W. S. Noble. Protein ranking: from local to global structure in the protein similarity network. *Proceedings of the National Academy of Sciences*, 101(17):6559–63, 2004.