Using the SCOP database as a gold standard to benchmark BLAST and PSI-BLAST

Oliver Delf-Rowlandson - Jing Qin - Daniëlle de Jong - Izak de Kom

ABSTRACT

Any bioinformatic tool must be adjusted and benchmarked to assess the accuracy of its output. We have conducted a project with the aim of benchmarking the BLAST and PSI-BLAST search tools using the SCOP database as a gold standard. For the purposes of this project several BLAST and PSI-BLAST searches were conducted using varying parameters; differing e-value and data thresholds. The outputs of said searches were then compared to the SCOP database for evaluation. Using this method it was discerned that BLAST is a search tool more appropriate for specific searches, whereas it is seen that PSI-BLAST is more sensitive to distant homology. As a byproduct of this project an evaluation of SCOP in its role as a gold standard also became apparent, the database is seen to be very accurate in its role as a benchmark but contains fewer protein identities compared to the GO and Pfam alternatives. SCOP is limited by known 3D structural definitions and is as such a less expansive, though highly accurate, database.

INTRODUCTION

The use of bioinformatic tools has greatly improved and hastened the classification of protein structure, function, and relationship. In order to determine the function and heritage of unclassified proteins these proteins must be subjected to sequence and structural comparison. In such a manner divergent evolution of proteins under different selection pressures, among other points of interest, can also be observed.

Commonly employed tools for protein amino acid sequence comparison are the Basic Local Alignment Search Tool (BLAST) and Position-Specific Iterated BLAST PSI-BLAST. These alignment tools serve as indicators of homologous sequences by detection of similarities between the sequences. Homologous sequences share a common ancestry (Schäffer *et* al., 2001) [Q2.1].

Contrary to other alignment tools, BLAST searches for more significant words in the query protein sequence. It starts searching for more common words in the query sequence and then for the most significant words in order to fasten the alignment (Altschul *et* al., 1990 and Mount, 2007). In addition, BLAST algorithms are able to mark low-complexity sequences and repetitive regions in the query and database sequences (Mount, 2007).

Position-Specific Iterated BLAST (PSI-BLAST) is a more sensitive tool than common BLAST. Before the alignment is performed in PSI-BLAST a position specific score matrix (PSSM) is constructed using the chance for occurrence of amino acids at certain positions. PSI-BLAST performs an alignment only with significant regions of the query sequence (Altschul *et* al., 1990) [Q2.2].

However, sequence comparison alone is far from a certain method of determining protein homology and as such function. Due to mutations the amino acid sequence often differs between species. On the other hand, protein structure is highly conserved, as mutant variations can lose integral structural components necessary for function.

A large number of databases concerning protein structure have been established. One such database is the Structural Classification of Proteins (SCOP) established in 1994 by Alexey G. Murzin. The SCOP database has been determined by high-level scientific knowledge and visual identification of major secondary structural arrangement (Csaba, Bizele & Zimmer, 2009). [Q3c.3] The proteins in this database are classified according to the following order of different levels: family, superfamily, fold, and a class with structural classes based on the composition of the units (alpha or beta strands). (Lo conte *et* al., 2000) [Q3c.1]. Each protein may have more than one classifications if they belong to various families, superfamilies, folds or classes [Q3c.2].

Benchmarking is needed in order to ascertain the most effective and accurate sequence comparison methods. Therefore, the aim of this study is to benchmark the BLAST and PSI-BLAST ability and

accuracy in determining protein homology, using the SCOP database as a gold standard. In addition, we benchmark PSI-BLAST under varying parameters. Therefore, the research question of this paper is as follows: when benchmarking BLAST and PSI-BLAST, which of these tools performs the best indication of homologous sequences, using SCOP as a gold standard database?

METHODS

The implemented method is illustrated in a flowchart in Figure 1. Firstly, the FASTA sequence data of 209 Uniprot ID's was downloaded from http://www.uniprot.org/uniprot/ (fetch_sequences.py). These 209 FASTA files were combined to build a local FASTA database (makeblastdb, Package BLAST, version 2.2.31). The makeblastdb generated a .phr .pin and .psq file, which contained the indexes of the compiled database [Q1].

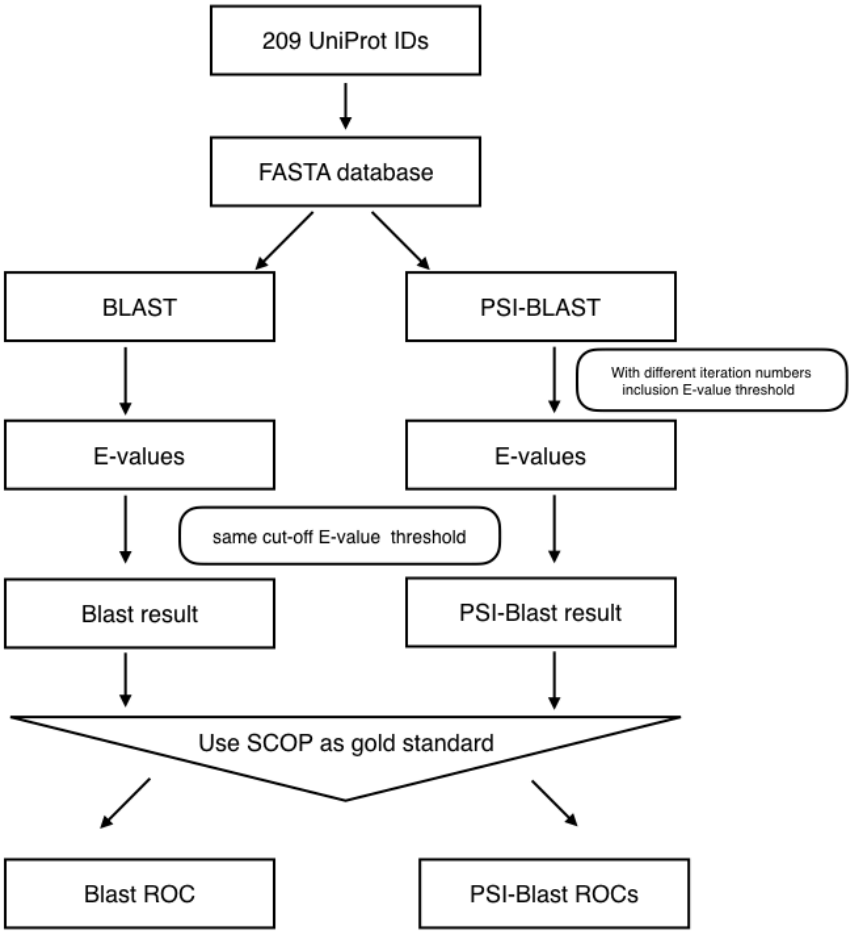

Figure 1. Flowchart illustrating the implemented method.

In order to find putative homologs, the BLAST and PSI-BLAST programs were used (blastp and PSI-BLAST respectively, Package BLAST, version 2.2.31). To assess the degree of similarity between two sequences, BLAST and PSI-BLAST calculate an alignment score. This is done using scoring matrices. While PSI-BLAST uses profiles (PSSM's), BLAST uses pre-defined scoring matrices (BLOSUM62). The calculated alignment score can be normalized to obtain the bit-score using

$$S' = \frac{\lambda * S - ln(K)}{ln(2)} \quad (1)$$

where λ is a scoring system scaling parameter, K is a search space scaling parameter and S is the alignment score. In the output of both BLAST and PSI-BLAST, the expected value (e-value) is reported for every sequence comparison. This value indicates how many homologies one could expect by chance given the size of the used database. A lower e-value is associated with a better alignment. The e-value is calculated by

$$E = m*n*2^{-S'} \quad (2)$$

where m and n are the lengths of the sequences and S' is the bit-score [Q2.3] (www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html, Zvelebil & Baum, 2008). The BLAST and PSI-BLAST programs can be run using the -evalue parameter. This parameter specifies the maximum e-value in the output. The PSI-BLAST program was run using two additional arguments: The -inclusion_ethresh parameter, which is the inclusion threshold in the profile construction [Q2.5] and the -num_iterations parameter, which specifies the number of search iterations executed PSI-BLAST. PSI-BLAST will stop iterating until either the user specified number of iterations is reached, or no new sequences are found after an iteration (convergence) [Q2.4]. The execution of both programs was automated in run_local_blast.py. The output of this script consisted of each protein pair and the associated e-value. If the e-value was lower than 100 the protein pair was flagged as a putative homolog ("Similar"). If the e-value was higher than this threshold the protein pair was flagged as a putative non-homolog ("Different").

The BLAST and PSI-BLAST program results were benchmarked using the SCOP database (1.75 release, June 2009) as a gold standard. Since the SCOP database classifies domains, a Uniprot ID of a protein can have multiple SCOP database entries [Q3c.2]. Consequently, when comparing the SCOP data of each protein pair for the gold standard all of the domains should be included in the comparison. The protein pairs were compared in two ways: comparing all domains by all SCOP data elements ("family", "superfamily", "fold" and "class") and comparing all domains only by the "fold" SCOP data element. In both comparisons, a similarity score was computed. This score consisted of two partial scores; the first partial score was calculated by counting the number of SCOP domains of protein 1 that can be found in protein 2, and then dividing by the number of domains of protein 1. The second partial score was calculated by doing a reverse comparison (e.g. counting the number of SCOP domains of protein 2 that can be found in protein 1, and then dividing by the number of domains of protein 2). Finally, the average of the two partial scores was returned as the final similarity score. A score higher than 0.5 indicated a homologous ("Similar") protein pair, and a score lower than 0.5 indicated a non-homologous ("Different") protein pair [Q3c.4].

Table 1. Classifications of (PSI-)BLAST and gold standard results.

	"Similar" by (PSI-)BLAST	**"Different" by (PSI-)BLAST**
"Similar" by gold standard	True positive	False negative
"Different" by gold standard	False positive	True negative

The true positives, false positives, true negatives and false negatives were determined as described in Table 1 [Q3c.4]. This information was used to compare the BLAST and PSI-BLAST outputs based on the resulting receiver operating characteristic (ROC) curves (create_roc_plot.py), and the corresponding area under the curve (AUC).

The (PSI-)BLAST and benchmarking runs were automated (experiment.py) for a more efficient experimentation phase and the option to run the PSI-BLAST program multiple times while varying the -num_iterations and -inclusion_ethresh parameters.

RESULTS & DISCUSSION

Similar AUC scores for BLAST and PSI-BLAST.

In order to include only related proteins in the search PSI-BLAST is more sensitive than BLAST in picking up distant evolutionary relationships (Bergman NH et al. 2007). It can be seen from the e-value distributions in Fig.2A and Fig.2B that PSI-BLAST found more hits than BLAST, especially in higher e-value ranges, which indicates that PSI-BLAST is more sensitive, and BLAST is more specific [Q2.7].

In the ROC curve, a random method would result in a diagonal line from the left bottom to the top right corner. The perfect method would give a point in the upper left corner or coordinate (0,1) of the ROC space. The AUC of a random method will be 0.5, and the perfect method will have an AUC score of 1. [Q4.1] The protein pair P0A6Y8 and P19120 is different in SCOP, while having the lowest e-value 0.0 for BLAST. Based on how the ROC curve is created, false positives at a low e-value can be found at the left side of the ROC curve, near the (0,0) point. Given this e-value, we think it is likely that these proteins have very similar sequence, but different structures. They may come from a common ancestor, but not share same function anymore. [Q4.3] Better results were expected when using PSI-BLAST rather than BLAST, as shown in Figure 3A. However, when the e-value was changed from 1 to 10 in order to get a higher AUC score for PSI-BLAST, the AUC value for BLAST became higher than the AUC for PSI-BLAST (Fig. 3B) [Q4.2]. This is thought to be due to the use of SCOP as a gold standard yielding a small pool of TP, both BLAST and PSI-BLAST gave more hits compared to SCOP's TP pool, and, since PSI-BLAST picked up more distant homologs this resulted in a lower TP/FP ratio.

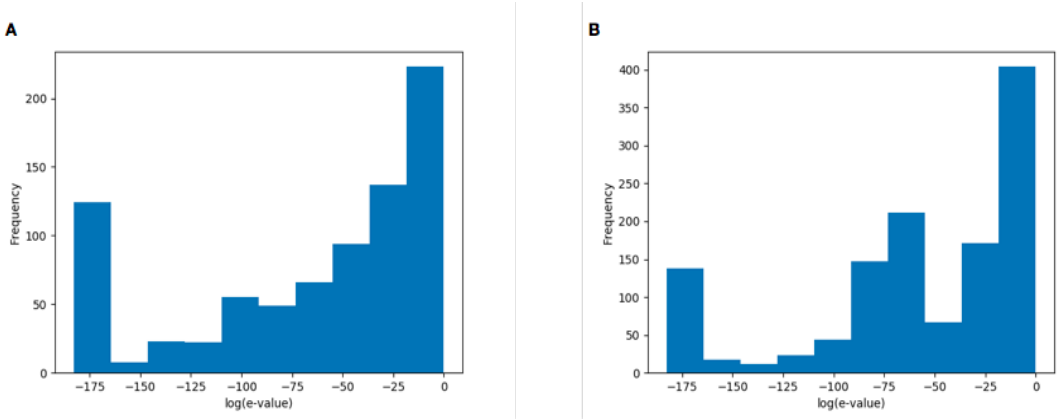

Figure 2. E-value distributions of the BLAST and PSI-BLAST runs. **A:** BLAST, -evalue parameter = 1. **B:** PSI-BLAST, -evalue parameter = 1, -num_iter parameter = 3, -inclusion_ethresh parameter = 1.01.

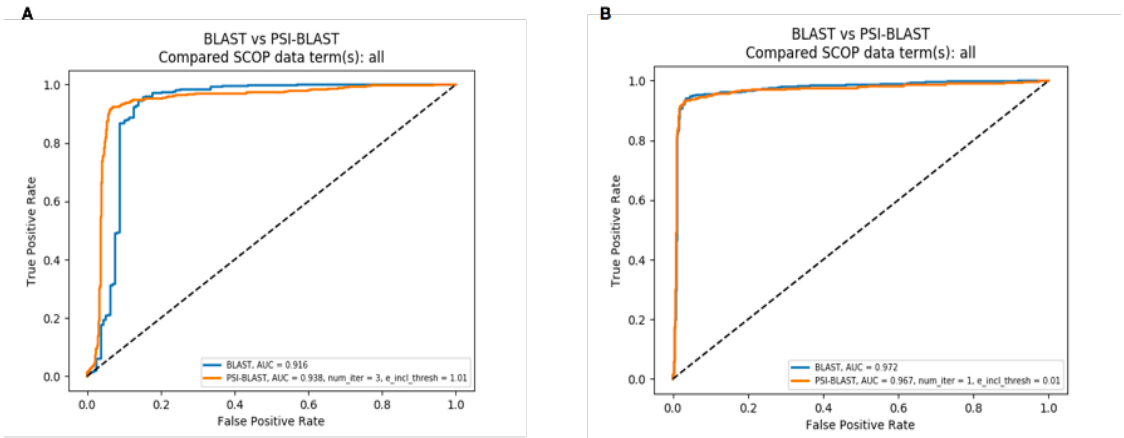

Figure 3. ROC curves of the BLAST and PSI-BLAST benchmark results. **A:** -evalue parameter = 1. **B:** -evalue parameter = 10. The PSI-BLAST parameters were chosen based on highest resulting AUC.

How the parameters change the AUC score

Table 2A and Table 2B show how different parameters have influenced the final AUC score. Besides the previously discussed e-value setting a large disparity was observed due to the way the similarity score is calculated using SCOP as a gold standard. The strict standard which only considers two proteins homologous when they are similar in all classification categories gives a higher AUC score than the less strict standard which only compares the "fold" database elements. [Q2.6] When the number of iterations is 1, PSI-BLAST runs very similarly to BLAST, so the e-value inclusion threshold is not applicable in this circumstance. When the number of iterations is greater than 1 a different e-value inclusion threshold will result in different AUC scores. While the e-value inclusion threshold increases from 0.01 to 0.51, performance of PSI-BLAST also increases. This is due to PSI-BLAST gaining a broader search pattern after each round of BLAST which in turn gives more TP hits. After the 0.51 mark the increase of the e-value inclusion threshold leads to a lower AUC caused by an increased number of TP results.

Table 2. Different parameters result in different AUC scores.

A

AUC score of PSI-BLAST compared SCOP data term(s): All

e-value	e-value inclusion threshold	Number of iterations		
		1	3	5
1	0.01	0.9363	0.9163	0.9174
	0.51		0.9367	0.9310
	1.01		0.9384	0.9303
10	0.01	0.9672	0.9633	0.9630
	0.51		0.9630	0.9580
	1.01		0.9615	0.9547

B

AUC score of PSI-BLAST compared SCOP data term(s): fold

e-value	e-value inclusion threshold	Number of iterations		
		1	3	5
1	0.01	0.9151	0.8825	0.8841
	0.51		0.9038	0.9025
	1.01		0.8995	0.8960
10	0.01	0.9158	0.9040	0.9030
	0.51		0.9036	0.9001
	1.01		0.9008	0.8883

Table 3. Comparison between the GO, Pfam and SCOP databases.

	Size of samples	Categorised	Based on	Last updated
GO	1163655 gene products	cellular components, molecular function biological processes	experimental evidence	continuously
Pfam	16712 entries	families:16,712 clans: 604	seed alignment	March 2017
SCOP	110800 Domains 38221 PDB Entries	folds:1195 superfamilies:1962 families:3902	structural	2014

Choose the gold standard database wisely

GO, Pfam and SCOP are the three most commonly used gold standard databases in benchmarking. As shown in Table 3, GO is most related to experimental evidence, but often not guaranteed for homology, and it is used less than SCOP and Pfam in practice. Pfam can provide good sensitivity based on seed alignment but is more expensive in terms of computational power (Bateman et al. 2004). SCOP only includes proteins with a defined 3D structure, but is not up to date. Conclusively, SCOP is a solid database for use in a PSI-BLAST search or as a gold standard. However, it may give negative feedback depending on the proteins in question, which may not have a defined 3D structure yet (the basis of SCOP classification). In this case the Pfam database is a useful secondary choice. [Q4.4]

For SCOP, a true homologue is found by alignment of the structures and sequences, and by homologous features based upon visual identification. Homologous proteins appear in the same families, superfamilies, folds and classes. For GO, homology detection depends on terms created by

GO, which relate to two proteins showing the same function under experimental conditions. For Pfam, a true homologue is defined based upon shared HMMs, homologs will show up in the same clans and families. As the proteins are classified in different ways depending upon the database each benchmark will also result in a different ROC curve. [Q4.5]

If no local database was present, one could use a database such as NCBI (https://www.ncbi.nlm.nih.gov/) which contains built in BLAST and PSI-BLAST features for searching its database. This can give a basic outline of a protein's relatives and any homologous that might be present. The starting parameters may be altered and those that significantly affect output are; matrix, gap costs, word length, and e-value threshold. The matrix is the method in which the statistical probabilities are calculated. Commonly used is BLOSUM62. The gap cost is the number of gaps introduced, and increasing this number decreases the number of gaps introduced. The word length dictates the length of sequence for comparison, so the standard word size of three compares three residues at a time before moving on. Increasing word size will speed up a search but will lower sensitivity and the reverse is also true. [Q2.8]

CONCLUSIONS

By benchmarking the BLAST and PSI-BLAST search tools using the SCOP database as a gold standard, it was discerned that BLAST is a search tool more appropriate for specific searches, whereas PSI-BLAST is more sensitive to distant homology. During the evaluation of the SCOP database in its role as a gold standard it also became apparent that the database is seen to be very accurate but contains fewer protein identities compared to the GO and Pfam databases. Furthermore, it will be interesting to find out how to apply this method in larger database searches, and find a standard strategy to choose the best benchmarking method and parameters.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. Journal of Molecular Biology 215, 403–410. doi:10.1016/S0022-2836(05)80360-2

Aniba, M.R., Poch, O., Thompson, J.D., 2010. Issues in bioinformatics benchmarking: the case study of multiple sequence alignment. Nucleic Acids Research 38, 7353–7363. doi:10.1093/nar/gkq625

Bateman, A., 2004. The Pfam protein families database. Nucleic Acids Research 32, 138D–141. doi:10.1093/nar/gkh121

Brenner, S. E., Hubbard, T. J., Chothia, C., Murzin A. G., 2004. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Research, Volume 32, Issue suppl_1, 1 January 2004, Pages D226–D229,https://doi.org/10.1093/nar/gkh039

Bergman, N.H. (Ed.), 2007. Comparative Genomics: Volumes 1 and 2. Humana Press, Totowa (NJ).

Csaba, G., Birzele, F., Zimmer, R., 2009. Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis. BMC Structural Biology 9, 23. doi:10.1186/1472-6807-9-23

Finn, Rob; Mistry, Jaina (8 March 2017). "Pfam 31.0 is released". Xfam Blog. Retrieved 13 March 2017.

Lo Conte, L., Ailey, B., Hubbard, T.J., Brenner, S.E., Murzin, A.G., Chothia, C., 2000. SCOP: a structural classification of proteins database. Nucleic Acids Res. 28, 257–259.

Mount, D.W., 2007. Using the Basic Local Alignment Search Tool (BLAST). Cold Spring Harbor Protocols 2007, pdb.top17. doi:10.1101/pdb.top17

Pearson, R. W., 2014. An Introduction to Sequence Similarity ("Homology") Searching. doi: 10.1002/0471250953.bi0301s42

Schäffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V., Altschul, S.F., 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res. 29, 2994–3005.

SCOP: Structural Classification of Proteins. [ONLINE] Available at: http://scop.mrc-lmb.cam.ac.uk/scop/. [Accessed September 2017].

The Statistics of Sequence Similarity Scores. [ONLINE] Available at: https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html. [Accessed September 2017].

Zvelebil, M.J., Baum, J.O., 2008. Understanding bioinformatics. Garland Science, New York.