Testing statistical significance scores of sequence comparison methods with structure similarity

Tim Hulsen[1], Jack A.M. Leunissen[2], Jacob de Vlieg[1,3], Peter M.A. Groenen[3]
[1]CMBI, Radboud Univ. Nijmegen [2]Wageningen Univ. & Research Centre [3]NV Organon, Oss

Summary

Due to improved implementations and rapidly increasing computing power the Smith-Waterman sequence comparison algorithm has gained popularity over time. However, the quality and sensitivity of a database search is not only determined by the algorithm but also by the statistical significance testing for an alignment. The e-value is the most commonly used statistical validation method for sequence database searching. The CluSTr database and the Protein World database (part of Biorange SP 3.2.2.) are created using an alternative statistical significance test: a Z-value based on Monte-Carlo statistics. From a theoretical point of view it has been demonstrated that the Z-value is statistically superior to the e-value. We were interested if this holds true for evolutionary related protein sequences.
All experiments are performed on the ASTRAL SCOP database. The Smith-Waterman sequence comparison algorithm with both e-value and Z-value statistics is evaluated, using ROC, CVE and AP measures. The BLAST and FASTA algorithms are used as reference. We find that two out of three Smith-Waterman implementations with e-value are better at predicting structural similarities between proteins than the Smith-Waterman implementation with Z-value. Especially SSEARCH has very high scores.

Materials & Methods

We used a non-redundant protein-domain sequence database derived from PDB as the target database. It is automatically generated using the ASTRAL system (Brenner et al., 2000). According to the structural classification of proteins (SCOP release 1.65), it includes 9498 sequences and 2326 superfamilies. True positives are those in the same superfamily as the query sequence. SCOP as an independent and accurate source for evaluating database search methods has been used by other researchers (Brenner et al., 1998; Park et al., 1998). ASTRAL SCOP sets with different maximal percentage identity thresholds (10%, 20%, 25%, 30%, 35%, 40%, 50%, 70%, 90% and 95%) were downloaded from http://astral.stanford.edu/scopseq-1.65.html. The tested methods are shown in table 1, together with the parameters used. We created Receiver Operating Characteristic (ROC), Coverage Vs. Error (CVE) and Average Precision (AP) values for all methods and all ASTRAL SCOP sets, using the top 100 hits for each query.

method	abbr.	version	matrix	gap open	gap ext.	rand.
Paracel SW e-value	pc e	-	BL62	3*IS	0.3*IS	0
Biofacet SW Z-value	bf z	2.9.6	BL62	12	1	100
NCBI BLAST e-value	bl e	2.2.9	BL62	12	1	0
FASTA e-value	fa e	3.4t24	BL62	12	1	0
SSEARCH e-value	ss e	3.4t24	BL62	12	1	0
ParAlign SW e-value	pa e	4.00	BL62	12	1	0

*) IS = average matrix identity score

Table 1 | **Sequence comparison methods and parameters**

Results

The ROC, CVE and AP scores give quite similar results. SSEARCH scores best of all methods, followed by ParAlign and FASTA. Biofacet Smith-Waterman with Z-value scores only better than one out of three Smith-Waterman implementations with e-value, i.e. Paracel. Smith-Waterman scores better than BLAST, as predicted by theory. FASTA however gives almost as good results as the average Smith-Waterman The Z-value seems to score better when more similar proteins are compared (PDB050-PDB095).

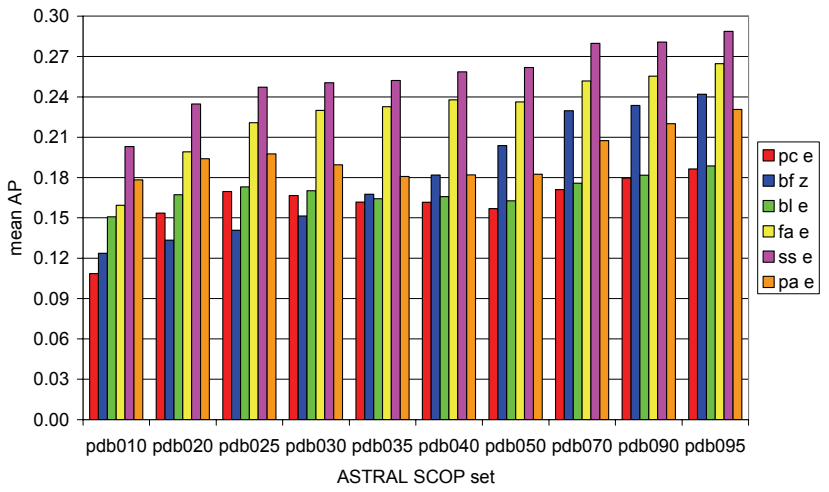
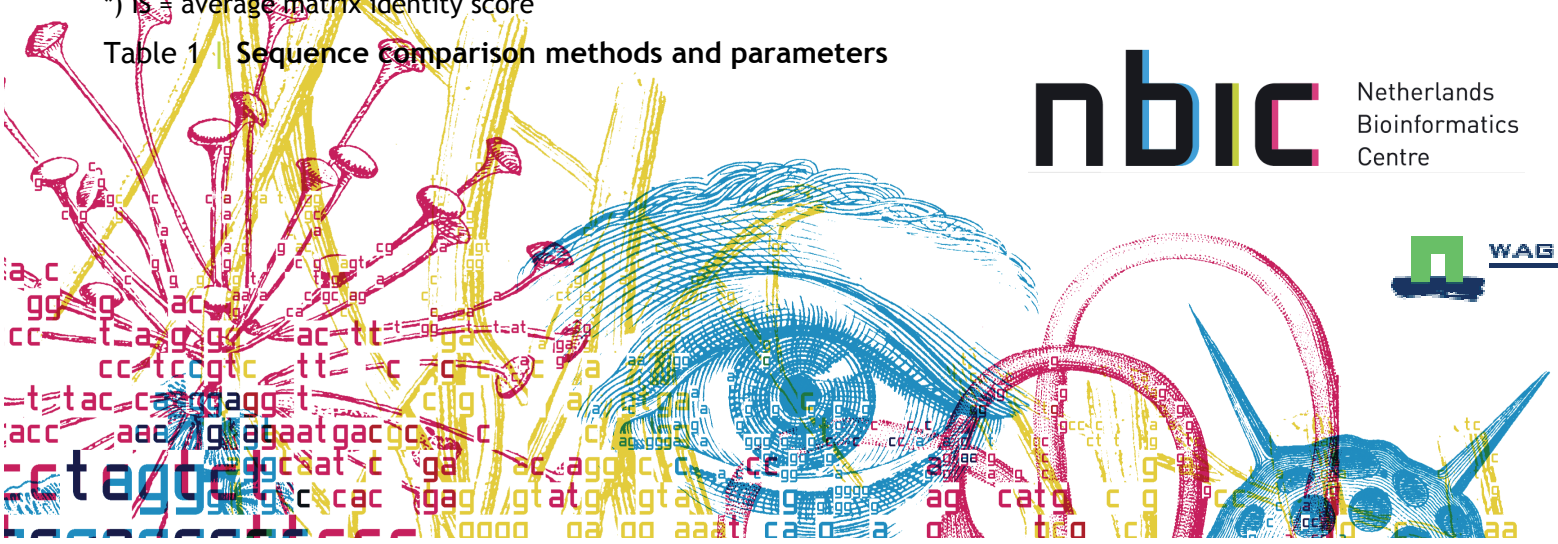


Figure 1 | **The average precision values for the tested methods**

Discussion

The theoretical advantage of the Z-value over the e-value appears to be disproven by our results. Our results show that the e-value calculation gives an advantage over the computationally intensive Z-value, at least when looking only at the results from the Smith-Waterman algorithm. Some caution should be taken however, drawing any definite conclusions. First, the Z-value was designed to make a distinction between significant hits and non-significant hits that have high SW scores. It might have an advantage over the e-value when applied only to the top hits, and might have less advantage for the hits with lower SW scores. Second, the Z-value can differ for each run, because of it different randomizations. Third, the PDB set is some-what biased: it only contains crystallized proteins, and no hypothetical proteins. For a complete analysis we need a more extended database, having a wide range of proteins classified by structure similarity. Regardless of all these theoretical assumptions, the computational disadvantage of the Z-value is smaller for larger databases. Z-values do not have to be recalculated when sequences are added to the data-base, in contrast to e-values, which are dependent on database size. For very large databases containing all-against-all comparisons, this is an important advantage of the Z-value. Despite these considerations, we recommend using SSEARCH with e-value statistics for pairwise sequence comparisons.



nbic Netherlands Bioinformatics Centre

CMBI

WAGENINGEN UNIVERSITEIT
WAGENINGEN UR

Organon