Comparison with other methods

Various methods that compute consensus sequences use various algorithms and sets of parameters in their procedures. Therefore, the direct comparison of remote homology recognition results from methods using different types of consensus sequences will not necessarily answer the question as to what general procedure is fundamentally better. Nevertheless such a comparison may be useful to the potential user of consensus sequences. Here we tested the ability to recognize SCOP superfamily level relationships using different types of consensus sequences. We have downloaded the latest versions of COBBLER (version 14.2) and ProDom (version 2005.1). The obtained sequence sets did not contain all of the SCOP sequences that were needed for the comparison. Therefore, for each SCOP sequence from our original data set of 2476 unique sequences, we have tried to find the closest corresponding COBBLER and ProDom consensus sequence using BLAST (e-value less than $10^{-6}$). For 1101 SCOP sequences we were able to find the related COBBLER consensus sequences. The corresponding ProDom consensus sequences were found for 2158 of the SCOP sequences.

In the case of COBBLER, the aligned fragments of consensus sequences were used as representatives of corresponding SCOP domains ($\text{consensus}^{\text{COBBLER}}$). The same fragments of actual raw protein sequences were used to derive our consensus sequences ($\text{consensus}^{\text{PSI-BLAST}}$) using the procedure described in this manuscript (except that we used only sequences found in Blocks for the construction of the PSI-BLAST PSSM). We found that our consensus sequences outperformed COBBLER sequences in recognition of remote homologues belonging to the same SCOP superfamily (Fig. A1).

A similar procedure was applied to the comparison with ProDom. The ProDom consensus sequences coincide with protein domains. Therefore, we represented the SCOP sequences from our test set with entire, most similar (according to BLAST) ProDom consensus sequences ($\text{consensus}^{\text{ProDom}}$). From among the raw sequences belonging to a multiple alignment of a ProDom domain we chose a raw sequence fragment that was most similar to the ProDom consensus sequence. We used that fragment for building our consensus sequence ($\text{consensus}^{\text{PSI-BLAST}}$) using the procedure described in this manuscript (except that we used only sequences found in ProDom for the construction of the PSI-BLAST PSSM). We found that our consensus sequences outperformed ProDom sequences in recognition of remote homologues belonging to the same SCOP superfamily (Fig. A2).
We found that 1192 SCOP sequences (from our original set of 2476) had a significant BLAST alignment score among COBBLER consensus sequences available with the latest Blocks distribution. The aligned consensus sequence fragments were used as representatives of the 1192 SCOP sequences. For each of those representatives we found a corresponding raw sequence fragment and used it to generate a new consensus sequence as described in this manuscript. We ended up with three sets of 1192 representatives composed of COBBLER, PSI-BLAST-based consensus and actual sequence fragments. We queried those three databases with 1192 actual SCOP sequences using BLAST. We sorted the alignments based on BLAST e-values. True positive pairs were composed of a protein and a representative fragments that belonged to the same SCOP superfamily, while false positives were pairs from different SCOP folds (pairs belonging to the same SCOP family were omitted as trivially related). The PSI-BLAST-based consensus sequences clearly outperformed COBBLER and raw sequences.
Fig. A2: Remote homology recognition with ProDom and PSI-BLAST-based consensus sequences. We found that 2158 SCOP sequences (from our original set of 2476) had a significant BLAST alignment score among consensus sequences of ProDom domains. For each of those 2158 representatives we found the corresponding most similar (smallest BLAST e-value) raw sequence fragment among the fragments used for derivation of ProDom domain consensus sequences. We used those fragments to generate new consensus sequences as described in this manuscript. Thus we ended up with three sets of 2158 representatives composed of ProDom, PSI-BLAST-based consensus and actual sequence fragments. We queried those three databases with 2158 actual SCOP sequences using BLAST. We sorted the alignments based on BLAST e-values. True positive pairs were composed of a protein and a representative fragment that belonged to the same SCOP superfamily, while false positives were pairs from different SCOP folds (pairs belonging to the same SCOP family were omitted as trivially related). The PSI-BLAST-based consensus sequences clearly outperformed ProDom and raw sequences.
Low error rate performance

Plotting numbers of true vs. false hits using a linear scale for both axes makes it easy to see how the rate of accumulation of true relations varies. However, the details of performance in the very low error rate region are sometimes obscure. The corresponding plot with the logarithmically scaled false positives axis reveals low error performance clearly (Fig. A3).

Fig. A3: Consensus sequence-based searches unraveled many new relations at any error rate. We sorted our consensus sequence-based PSI-BLAST (blue circles), the original sequence-based PSI-BLAST (green rectangles), and pairwise BLAST (grey triangles) alignments by e-values. The scores were generated by aligning 2476 query proteins against each other. True positive pairs were proteins from the same SCOP superfamily, while false positives were proteins from different SCOP folds. By construction, we excluded all pairs that were trivially related (belonging to the same SCOP family) which explained why the curves for the pairwise BLAST were so low. The logarithmic scale of the horizontal axis clearly shows that consensus alignments improved performance at any error rate.
