SUPPLEMENTAL MATERIAL

R.C. Edgar, *Quality measures for protein alignment benchmarks*.

BALIBASE set BBS11013

BBS sets are described as being trimmed to homologous regions, and should therefore have high agreement by superfamily annotations. An example of a BBS set with low scores by superfamily and DSSP agreement is BBS11013, shown in Fig. 3 (main text). SCOP assigns the structures to three different superfamilies (a.4.1, a.4.5 and a.4.12), as does CATH, which indicates a lack of convincing evidence for homology. Core block 2 has maximal DSS, aligning a surface loop in 1hst_A to a helix in 1idy. This region does not meet the stated criteria for core blocks, which are intended to correspond to conserved secondary structure and exclude loops.

BALIBASE sets BBS11025 and BBS20029

Set BBS11025 is maximally incorrect by the domain measures (CFLD=0, ECLS=100%) according to both SCOP and CATH. This set contains four sequences; all have known structure (1sap_, 1prtF, 1lt5D and 1tvxA). SCOP and CATH agree that 1prtF and 1lt5D belong to the same superfamily and lack convincing evidence of homology to 1tvxA and 1sap_. CATH assigns 1tvxA and 1sap_ to the same superfamily, but SCOP does not. These structures all contain, among other secondary structure elements, three anti-parallel beta strands followed by an alpha helix. Core blocks correspond approximately to the first and second beta strands and to a pair of turns of the helix, respectively. The first core block aligns a region where the secondary structure clearly differs (Fig. S1); for example, according to DSSP the first two positions are in a 3/10 helix in 1tvxA and 1sap_, a beta strand in 1prtF and a transition region without defined secondary structure in 1lt5D. The third core block aligns two turns in the alpha helix, but the correspondence appears arbitrary since the helices have different numbers of turns. Set BBS20029 has the same four structures; its remaining sequences are full-length proteins from Uniprot. However, despite having identical structures, the core blocks in BBS20029 are different from BBS11025. The first block in BBS20029 is extended by two columns despite having a more diverse set of sequences. The third block is extended to three instead of two turns of the helix, covering most of the helix in 1tvxA but not in the other structures. While a distant evolutionary relationship between the beta-alpha motifs in these structures cannot be ruled out, the sequences and structures are sufficiently dissimilar that it is not appropriate to assess residue accuracy on this set as the correct alignment is not well-defined by structure or knowable by sequence homology. This raises the question of whether the rankings of alignment methods could be biased by spurious agreements. Scores of the trimmed and full-length versions of BBS10025 are shown for some representative programs in Table S1. In the full-length set BB11025, all programs I have tested score CS=0 (including several not reported here), meaning that no columns were aligned in agreement with the reference, while in the trimmed version scores range from CS=0 to CS=68%. This suggests that CS scores greater than zero in the trimmed set are explained by the fact that the trimmed sequences have similar lengths (58 to 63) so that a global alignment has a significant probability of containing "correct" columns by chance.

FIGURE

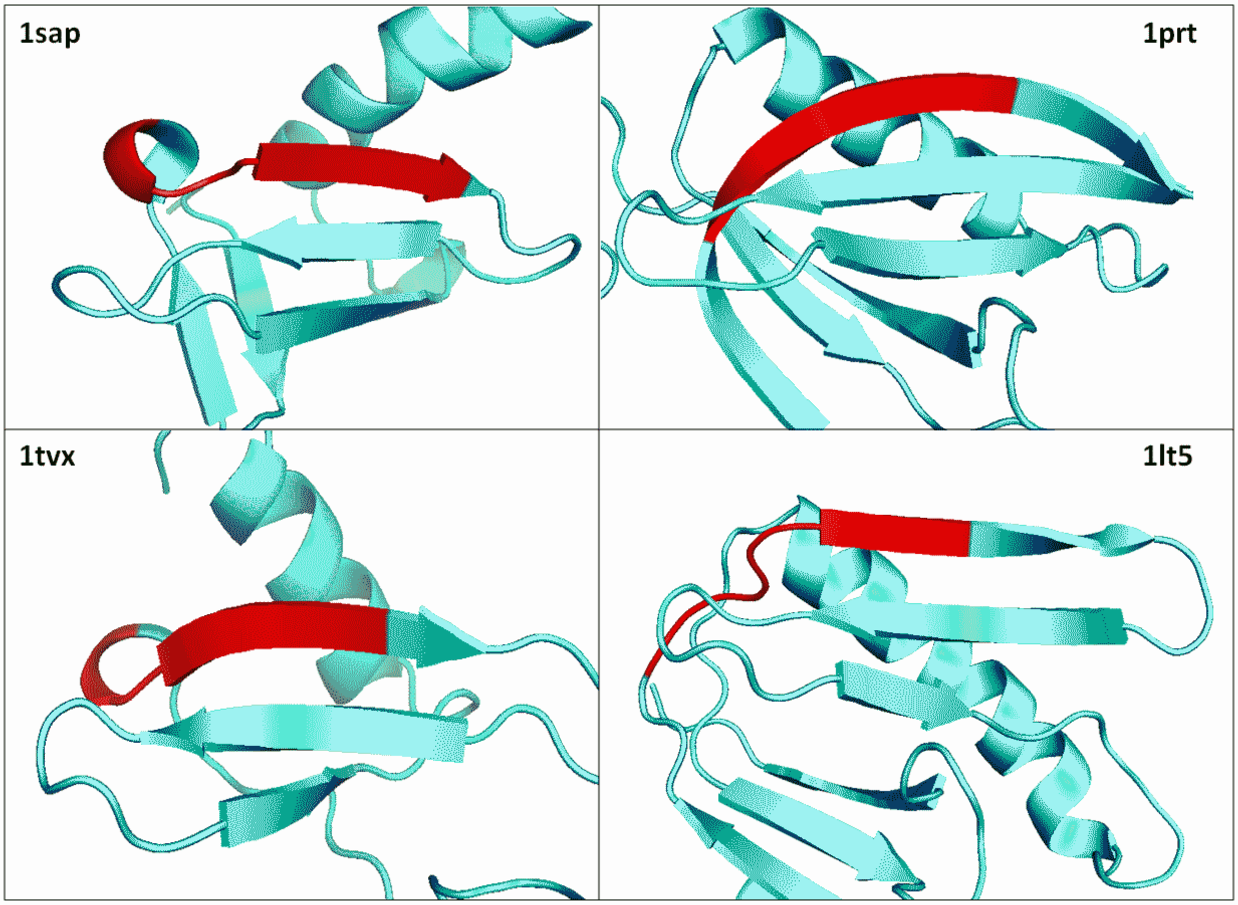

Figure S1. Core block in BALIBASE set BB20029

The first core block in BB20029 is highlighted in red in Pymol cartoons using secondary structure assignments in the PDB records. While the last few residues in the block are all in a beta strand with a possible distant relationship, the first few residues have different secondary structure and the annotation of this region as a reliably aligned core block with conserved secondary structure is contradicted.

TABLE

Method	SPS_{trim}	CS_{trim}	SPS_{full}	CS_{full}
MUSCLE	66%	53%	35%	0%
ALIGN-M	47%	37%	0%	0%
FSA	45%	26%	6%	0%
FSA -maxsn	74%	68%	10%	0%
MUSTANG	55%	11%	43%	11%

Table S1. Residue accuracy scores on BBS11025

SPS and CS scores for selected methods on set BB11025, which is arbitrarily aligned by BALIBASE. Subscript *trim* indicates scores for the trimmed set BBS11025, and *full* for the full-length set BB11025. Column scores of up to 68% are obtained on the trimmed set, while none of the sequence methods reproduce any columns of the full-length reference and the structural aligner MUSTANG (1) agrees on only two of the 19 core block columns (11%).

References

1. Konagurthu, A.S., Whisstock, J.C., Stuckey, P.J. and Lesk, A.M. (2006) *Proteins*, **64,** 559-574.