Figures and figure supplements

A vocabulary of ancient peptides at the origin of folded proteins

Vikram Alva *et al*

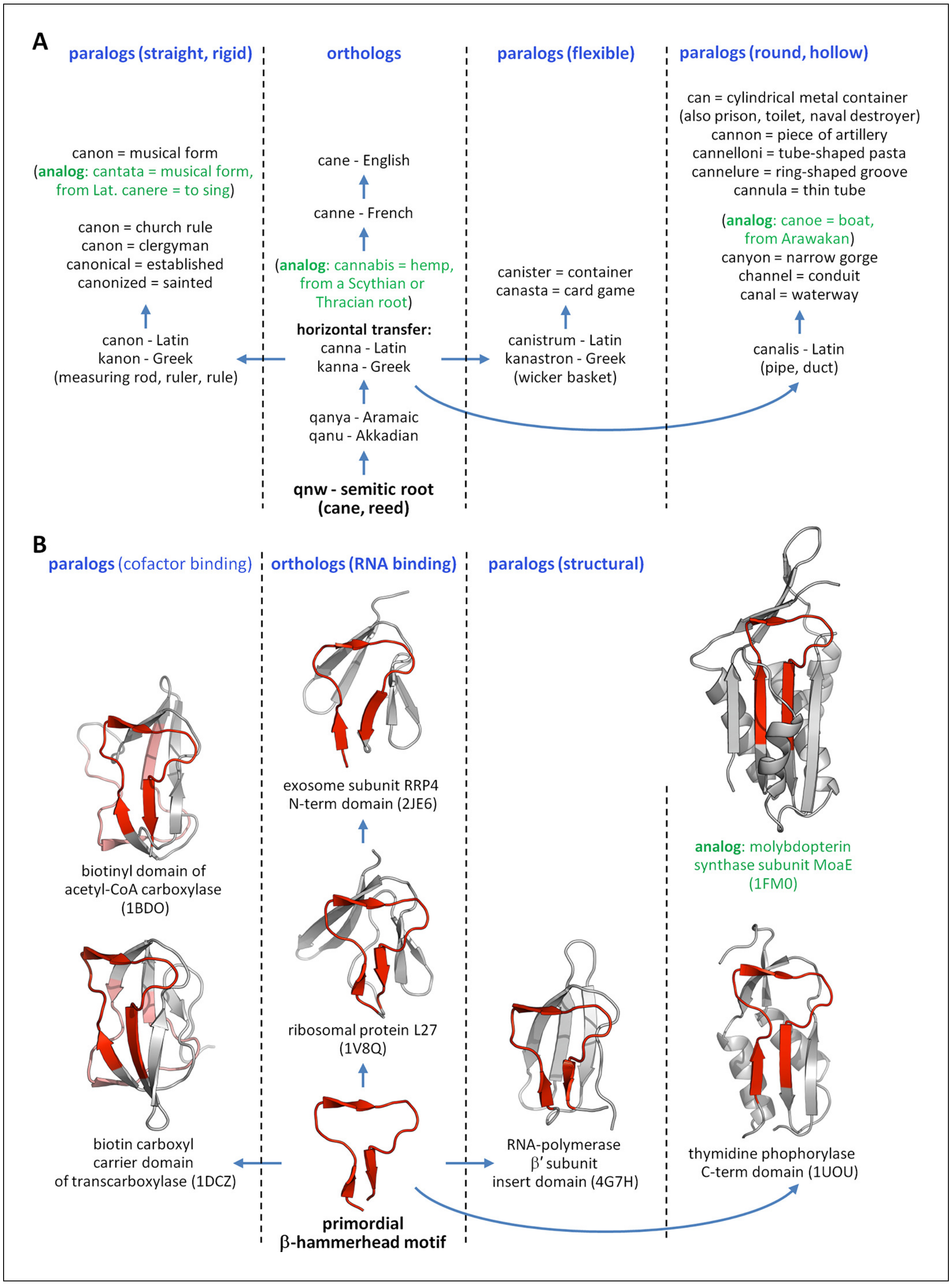


Figure 1. The evolution of words and proteins shows many parallels. (**A**) The Semitic root *qnw (*qanaw-)*, meaning reed, is the ancestor of hundreds of words in many different languages, following the same mechanisms as already known from biological evolution. Here we track the descendants of this root in the English language, arisen through the intermediary of Latin and Greek. In addition to the orthologous Greek word *kanna* (reed), paralogous cognates arose in antiquity based on certain attributes of reed, e.g., the levelling rule *kanon* (taking the straight and rigid attribute of reed), the wicker basket *kanastron* (flexible), and the Latin water duct *canalis* (round and hollow). A few examples of analogous words, which appear to be related to the descendants of *qnw* but have different evolutionary origins, are shown in green. (**B**) The primordial β-hammerhead motif (shown in red) is seen in four different folds, which cover a wide array of functions. Following our hypothesis of an origin in the RNA world, we propose that RNA binding is the orthologous function of this peptide, seen today in ribosomal protein L27 and

Figure 1 continued on next page

Figure 1 continued

exosome subunit RRP4. Paralogous functions arose around the time of the Last Universal Common Ancestor from its ability to form a biotin-binding domain by duplication, yielding the biotin-dependant enzymes of the barrel-sandwich hybrid fold, and to serve as a structural element in domains formed by accretion, yielding a domain of RNA-polymerase β′ subunit, as well as a range of enzymes with an α/β-hammerhead fold. By our analysis, enzymes classified in the α/β-hammerhead fold superfamily d.41.5, such as MoaE, are analogous to the other superfamilies in this fold, due to a lack of detectable sequence similarity, but nevertheless contain a supersecondary structure resembling the β-hammerhead.

DOI: http://dx.doi.org/10.7554/eLife.09410.003

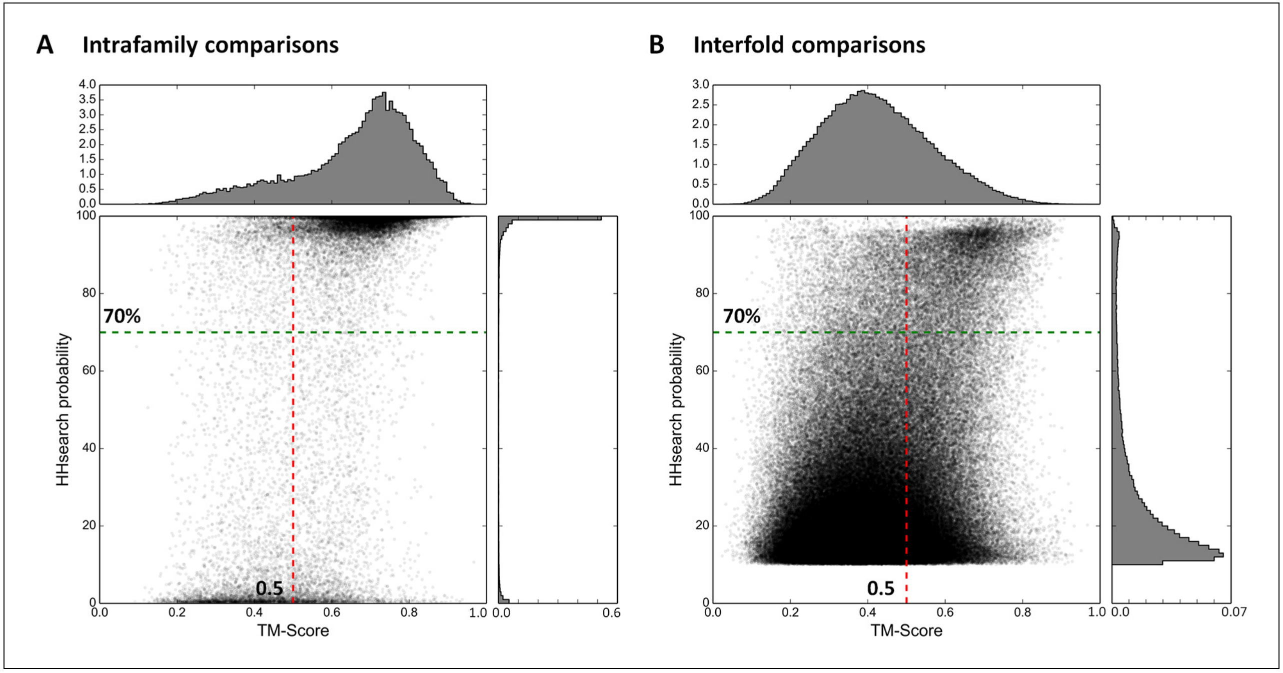


Figure 2. Estimation of cut-offs for HHsearch probability and TM-score. We compared all domains in the SCOPe30 set in sequence space using HHsearch and subsequently in structure space using TM-align (see 'Materials and methods'), and plotted the obtained scores. Separate plots for comparisons of domains within families (**A**) and between folds (**B**) were generated. Scores would have been expected at high HHsearch probabilities and TM-scores for intrafamily comparisons (presumed homologs, Panel A) and low HHsearch probabilities and TM-scores for interfold comparisons (presumed analogs, Panel B), but the score distributions were in fact bimodal, as also illustrated by the histograms top and right in each panel, which are plotted as probability density functions. In the comparison of domains of different fold, matches with an HHsearch probability of < 10% are not plotted.

DOI: http://dx.doi.org/10.7554/eLife.09410.004

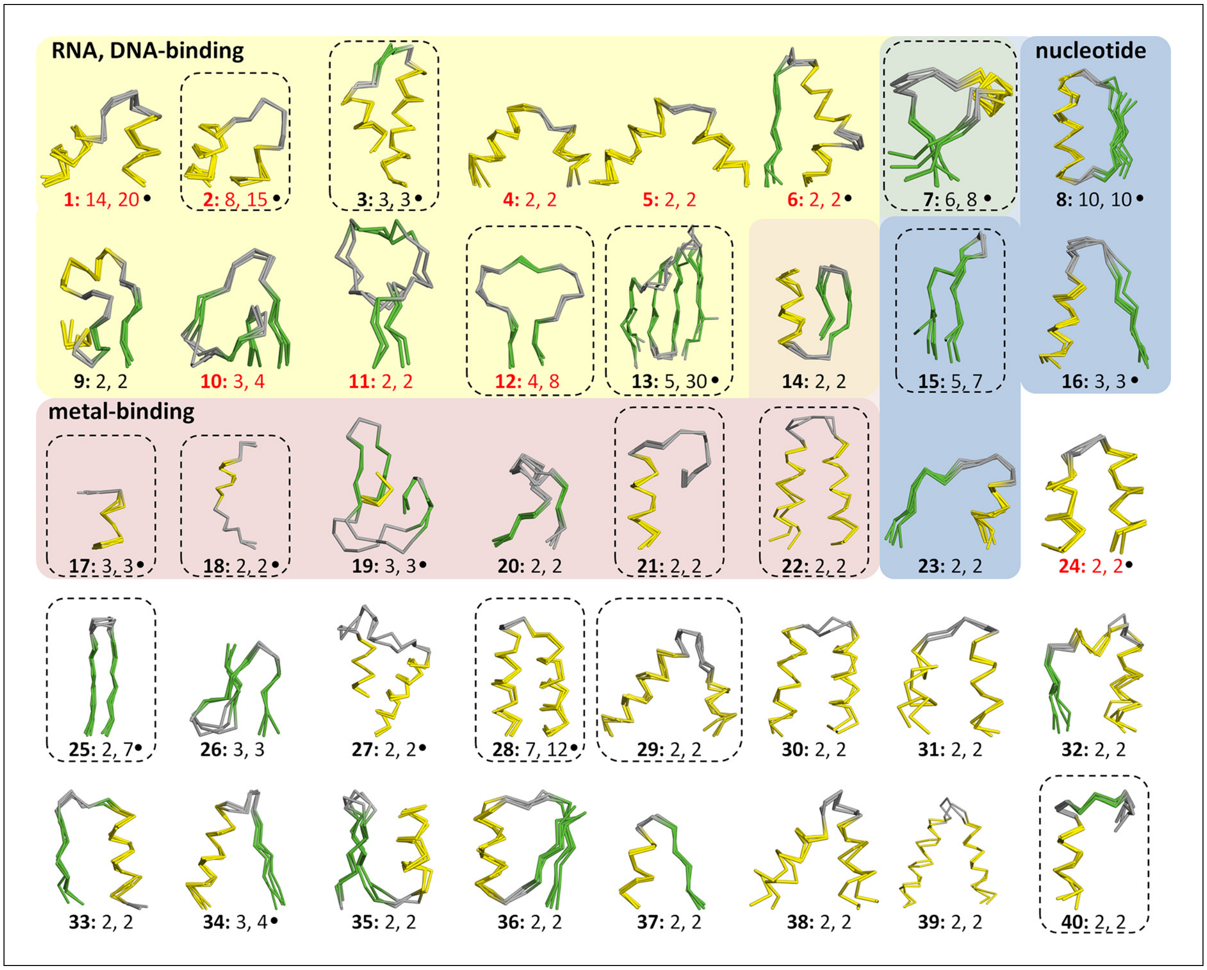


Figure 3. Vocabulary of primordial peptides that gave rise to folded proteins. The 40 peptides we detected are shown as ensembles in backbone representation; α-helices are coloured in yellow, β-strands in green, and loops in gray. Detailed information on each fragment is provided in ***Table 1*** and ***Figure 3—source data 1***. The fragments are numbered sequentially and their occurrence in different folds and superfamilies of SCOPe is given. Fragments reported individually before are indicated by a dot. Nucleic-acid binding, nucleotide-binding, and metal-binding motifs are highlighted in yellow, blue, and red, respectively. Fragments found in ribosomal proteins are indicated by red font colour. Fragments that form folds by repetition are boxed.

DOI: http://dx.doi.org/10.7554/eLife.09410.005

The following source data is available for figure 3:

Source data 1. Multiple sequence alignments and accession details for the 40 primordial fragments shown in ***Figure 3*** and the 5 B-set fragments shown in ***Figure 3—figure supplement 1***.

DOI: http://dx.doi.org/10.7554/eLife.09410.006

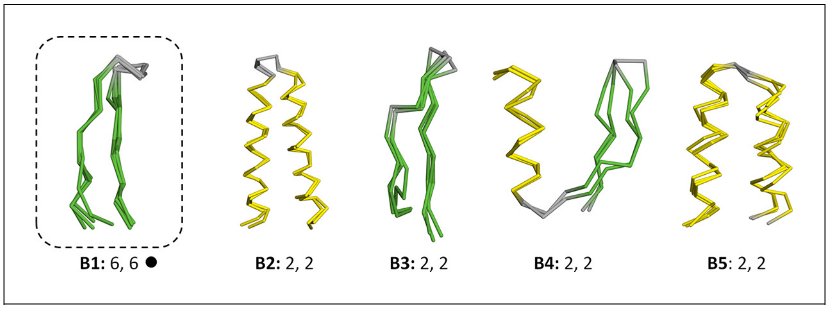
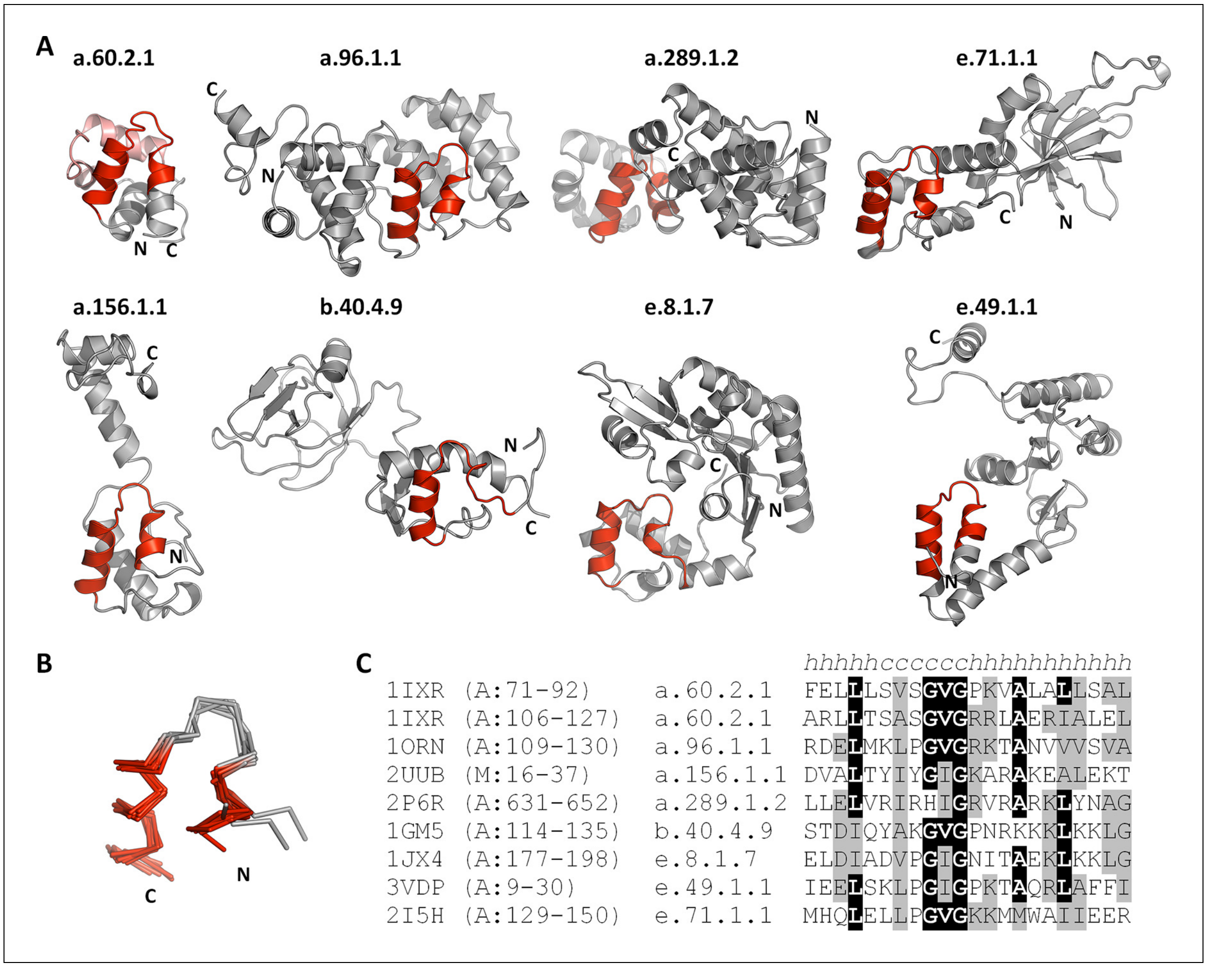


Figure 3—figure supplement 1. Vocabulary of primordial peptides (B-set). In addition to the 40 fragments described in the main text, we detected five further fragments after relaxing the sequence similarity requirement to an HHsearch probability of 60%. One of these fragments, the Asp box (B1), has been previously described and it forms folds by repetition (emphasized by dotted boxes). Detailed information is provided in ***Figure 3—source data 1***.

DOI: http://dx.doi.org/10.7554/eLife.09410.007



```
                                          hhhhhccccchhhhhhhhhhh
1IXR (A:71-92)     a.60.2.1   FELLLSVSGVGPKVALALLSAL
1IXR (A:106-127)   a.60.2.1   ARLLTSASGVGRRLAERIALEL
1ORN (A:109-130)   a.96.1.1   RDELMKLPGVGRKTANVVVSVA
2UUB (M:16-37)     a.156.1.1  DVALTYIYGIGKARAKEALEKT
2P6R (A:631-652)   a.289.1.2  LLELVRIRHIGRVRARKLYNAG
1GM5 (A:114-135)   b.40.4.9   STDIQYAKGVGPNRKKKLKKLG
1JX4 (A:177-198)   e.8.1.7    ELDIADVPGIGNITAEKLKKLG
3VDP (A:9-30)      e.49.1.1   IEELSKLPGIGPKTAQRLAFFI
2I5H (A:129-150)   e.71.1.1   MHQLELLPGVGKKMMWAIIEER
```

Figure 4. The nucleic-acid binding helix-hairpin-helix motif is found in 8 different folds comprising 15 superfamilies. (**A**) Representative domains from the eight SCOPe folds. The motif is coloured in red and the remainder of the structure is shown in gray. The SCOPe family a.60.2.1 contains two copies of this motif, whereas the remaining folds contain one copy each. (**B**) Structural superimposition of the helix-hairpin-helix motifs displayed in panel A. (**C**) Sequence alignment of the motifs shown in panel A. Residues conserved in at least half of the aligned sequences are highlighted in black and similar residues are highlighted in gray.

DOI: http://dx.doi.org/10.7554/eLife.09410.008

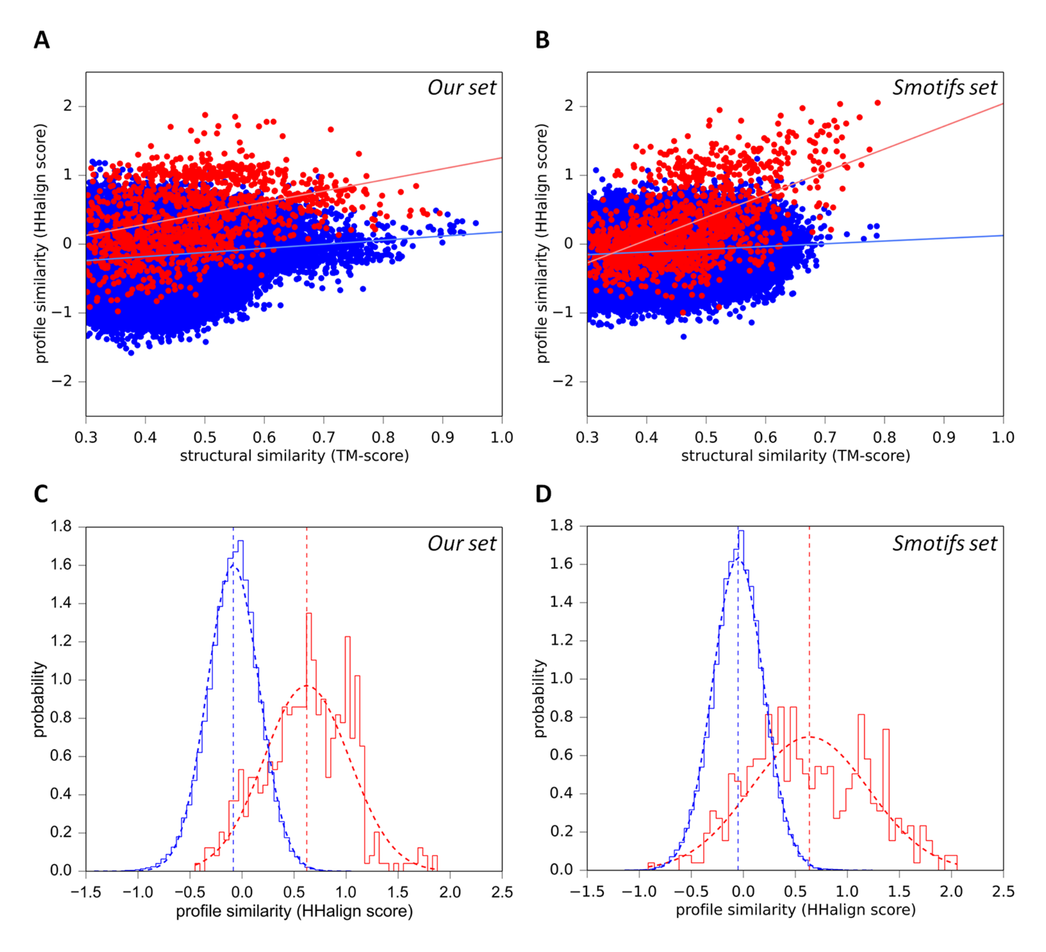


Figure 5. Sequence similarity of our fragments cannot be explained by structural constraints. (**A**) For each occurrence of any of our 40 fragments, we searched for structural matches in SCOPe30 and plotted the TM-align score versus the profile-similarity score for the fixed alignment given by TM-align. The putatively homologous matches to occurrences of the same fragment in another superfamily are shown in red. Matches to fragments outside the list of folds in which the query fragment was found to occur (i.e. non-homologous matches) are blue. (**B**) Same as A, but using the Smotifs reference fragments as queries instead of our set. Matches within superfamilies (homologs) are shown in red, matches between fragments from different folds (analogs) are shown in blue. For both sets, sequence and structure similarity scores are significantly correlated for presumably homologous matches (our set: r=0.38; Smotifs: r=0.56, see linear regression lines) but not for analogous matches (our set: r=0.14; Smotifs: r=0.12). (**C, D**) Distribution of profile similarity scores for matches with a TM-score $\geq$ 0.5, for the homologous and analogous distribution in the plots (A) and (B), respectively. The means of the Gaussian fits are exactly the same in C and D.

DOI: http://dx.doi.org/10.7554/eLife.09410.010

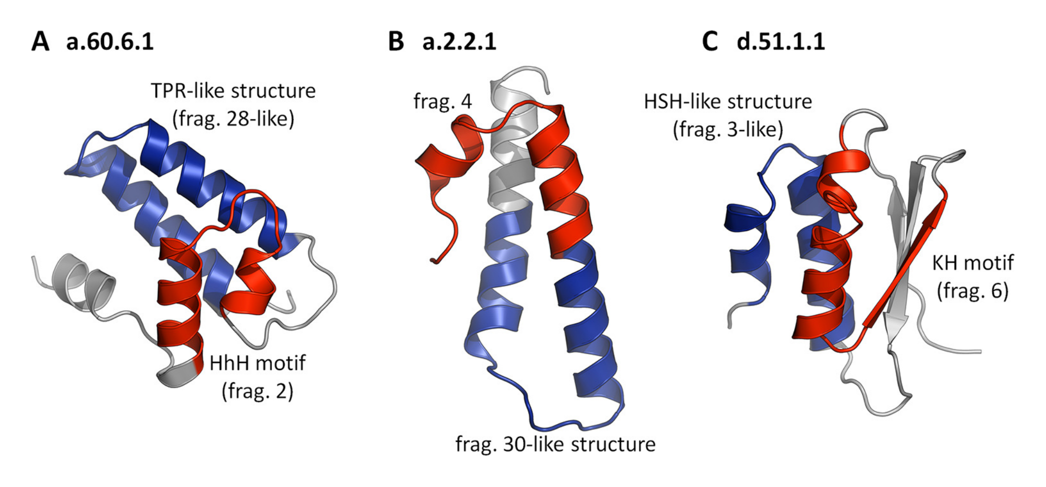


Figure 6. Folds showing two nonidentical fragments, one of which is which is not significant by our criteria. No SCOP fold combines two of our fragments at the cutoffs used in this study (TM-score ≥ 0.5 and HHsearch probability ≥ 70%). If we however omit the sequence cutoff entirely for the second fragment, combinations become apparent. In these three examples, the 'significant' fragments are colored in red, the 'nonsignificant' ones in blue, and the remainder in gray. The structures are: (**A**) a.60.6.1, N-terminal domain of polymerase β (4KLI, A: 10-91), (**B**) a.2.2.1, 50S ribosomal protein L29 (1VQ8, V: 1-65), and (**C**) d.51.1.1, KH domain-like hypothetical protein APE0754 (1TUA, chain A: 1-84).

DOI: http://dx.doi.org/10.7554/eLife.09410.011

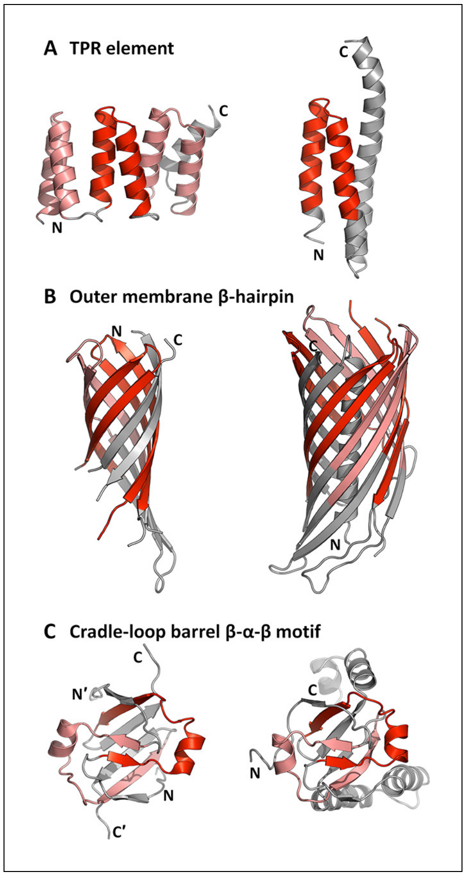


Figure 7. Amplification and accretion are key forces in the emergence of domains. Of the 40 fragments in our set, 14 form folds by repetition. The fragments are coloured in red in the shown structures. (**A**) The TPR element (***Figure 3***: 28) occurs repetitively in the TPR-like superfamily (a.118.8; 1ELW, shown on the left side) and singly in six other folds (e.g., a.7.16, 2CRB, right). (**B**) Outer membrane β-barrels comprise 4–12 homologous copies of a β-hairpin element (***Figure 3***: 25); examples include the eight-stranded OmpA (1QJP, left) and the twelve-stranded NalP (1UYN, right). The entire barrels are formed by repetition, but the strands of the hairpin split by the N- and C- termini are left gray. (**C**) The transcription factor AbrB (1YFB, left) is a homodimer and contains one copy of the β-α-β motif (***Figure 3***: 7) per subunit. MraZ has internal sequence symmetry and contains two homologous copies of the β-α-β motif (1N0E, right).

DOI: http://dx.doi.org/10.7554/eLife.09410.012