Figures and figure supplements

Inferring joint sequence-structural determinants of protein functional specificity

Andrew F Neuwald et al
Figure 1. BPPS-SIPRIS analysis of the GNAT superfamily and Gna1-family based on structural coordinates for Gna1 (pdb: 4ag9) (Dorfmueller et al., 2012). SIPRIS clearly associates Gna1-residues with the substrate and homodimeric interfaces ($p=8.5 \times 10^{-7}$). Color scheme: homodimer subunits A and B, green and blue backbones, respectively; BPPS-defined Gna1-family residues in subunits A and B, magenta and red sidechains, respectively (glycine residues are shown as C$_a$ atom spheres); GNAT superfamily residues, yellow sidechains; ligands, cyan. Lys116 (shown in light red) is outside of the SIPRIS defined cluster, but forms a hydrogen bond to a CoA phosphate group. BPPS-SIPRIS spherical clustering identified the GNAT superfamily residues shown ($p=1.7 \times 10^{-5}$). The following figure supplement and source data are available for Figure 1.

DOI: https://doi.org/10.7554/eLife.29880.003
Figure 1—figure supplement 1. Applying SIPRIS to the Gna1 protein in conjunction with various methods. Residue sets were defined using BPPS and three other programs (with default parameter settings). Residue color schemes: BPPS: NAT superfamily, yellow; Gna1-family, red; FRpred: conserved, yellow; subtype, red; CLIPS-1D: structurally important, yellow; ligand binding, orange; catalytic, red; ET: residues of high functional importance, orange; CoA and substrate, cyan. The SIPRIS predefined clustering p-values corresponding to the homodimer/substrate interface are indicated below each image.

DOI: https://doi.org/10.7554/eLife.29880.004
Figure 2. BPPS-SIPRIS analysis of R4 P-loop GTPases. Bound guanine nucleotide (shown in cyan) allows orientation of each subfigure relative to the others. (A) BPPS-defined hierarchical relationships among the GTPases examined here. (B) Entamoeba histolytica Rho1 GTPase (pdb: 3refB) (Bensch et al., 2011). Color scheme: R4-specific residues forming a BPPS-SIPRIS-defined hydrogen-bond network (p=8.3 × 10^-3), red sidechains; residues conserved in P-loop GTPases and interacting with bound guanine nucleotide, yellow sidechains; atoms forming hydrogen bonds, CPK coloring. Modeled hydrogen atoms were generated using the Reduce program (Word et al., 1999). (C) Rab4 bound to GTP and to the Rab-binding domain of Rabenosyn (pdb: 1z0kA [Eathiraj et al., 2005]). BPPS-SIPRIS-defined residues distinctive of R4 (red sidechains) and Rab (orange) have core and Rabenosyn-contacting predefined cluster p-values of 2.6 × 10^-6 and 2.9 × 10^-6, respectively. The sensor threonine (Thr40) has substantial van der Waals contact with Glu44; Thr40 is a R4-specific (red) residue outside of the SIPRIS-defined cluster. (D) Rab8a in complex with the GTP analog, GNP, and with Ocr1 (residue 540–678) (pdb: 3qbtA (Hou et al., 2011)). Residues distinctive of Rab GTPases (orange) and of the Rab8 subgroup (green) are enriched at the Ocr1 interface (p=5.2 × 10^-7 and 6.1 × 10^-6, respectively). (E) Rab8a homodimeric complex (pdb: 4hwAB) (Guo et al., 2013). Rab-specific residues (orange) are enriched at the homodimeric interface (p=8.7 × 10^-7). The following source data are available for Figure 2.

DOI: https://doi.org/10.7554/eLife.29880.006
Figure 3. BPPS-SIPRIS analysis of translation-associated P-loop NTPases. (A). *Thermus aquaticus* EF-Tu complexed with the antibiotic enacyloxin IIA, a GTP analog, and Phe-tRNA (pdb: 1ob5) (Parmeggiani et al., 2006). Color scheme: BPPS-SIPRIS defined GTPase-, TF- and EF-Tu/CysN-specific
residues, yellow, red, and orange sidechains, respectively; GTPase domain backbone, green; C-terminal β-barrel domains, gray; phe-tRNA, teal; 5’ end nucleotide bases, light cyan; guanine nucleotide, cyan; enacyloxin IIA, greenish-cyan. Spheres indicate glycine Cα atoms. (B). BPPS-SIPRIS cluster of EF-Tu TF-residues centered on EF-Ts Phe81 at the EF-Tu/EF Ts interface (pdb: 1efu) (Kawashima et al., 1996). Regions in EF-Ts conserved between E. coli and cow are shown in cyan both in the figure and in the corresponding alignment below it. (C). P. aeruginosa EF-Tu bound to the Tse6 toxin domain (pdb: 4zv4) (Whitney et al., 2015). EF-Tu His20, which corresponds to His19 in (B), appears to form a salt bridge with Glu291 of Tse6. In light pink are regions of Tse6 contacting EF-Tu. Spherically clustered residues (p=0.0060) centered on Glu291 of Tse6 are shown with red sidechains. (D). Spherically clustered EF-Tu/CysN residues (orange; p=6.3 × 10⁻⁶) within the CysND complex (pdb: 1zun) (Mougous et al., 2006). (E). Spherically clustered EF-Tu/CysN-residues in EF-Tu (pdb: 1ob5) (p=1.0 × 10⁻⁶). (F). Human eIF4AIII bound to RNA, ADP, and the γ-phosphate transition state mimic AlF₃ (pdb: 3e7) (Nielsen et al., 2009). Color scheme: eIF4AIII N- and C-terminal domains, violet and green, respectively; RNA and ADP, cyan; AlF₃, light cyan; superfamily-conserved catalytic residues, yellow sidechains; RNA helicase-specific residues clustered on (light cyan-colored) RNA bases 4–5, red; other RNA helicase-specific residues, light red; C-terminal catalytic residues, bright green. The following source data are available for Figure 3.

DOI: https://doi.org/10.7554/eLife.29880.008
Figure 4. BPPS-SIPRIS analysis of synaptojanin/EEP domains. (A). The two major groups of the BPPS-defined EEP hierarchy examined here. (B). Human APE1 phosphorothioate substrate complex (pdb: 5dfi) (Freudenthal et al., 2015). Replacement of the phosphodiester bond with phosphorothioate prohibits cleavage by APE1 at the abasic site (circled). Cys310, which is nitrosated, is indicated. Color scheme: APE1 backbone trace, green; DNA strand containing the abasic site, cyan; complementary strand, marine blue; the BPPS-SIPRIS-defined residues distinctive of the EEP superfamily and of the exoIII-AP-endo family, yellow and red sidechains, respectively; basic residues within a loop interacting with the major groove of DNA, purple. (C). Close up of the APE1 active site. EEP-specific residues forming a hydrogen-bond network are shown with yellow sidechains. For clarity, only a few of the EEP- and exoIII-AP-endo-specific residues in the network are shown. The following source data are available for Figure 4.

DOI: https://doi.org/10.7554/eLife.29880.010
Figure 5. BPPS-SIPRIS analysis of synaptojanin/EEP domains within INPP5 proteins. Color code: EEP-residues, yellow sidechains; INPP5 residues, red sidechains; INPP5B-, INPP5E- and SHIP2-subfamily residues, orange sidechains; ligands, cyan; atoms involved in hydrogen bonds, CPK coloring. (A). Figure 5 continued on next page.
Figure 5 continued

Human INPP5B in complex with phosphatidylinositol 3,4-bisphosphate (pdb: 4cml) (Tréaugues et al., 2014), which is associated with cytosolic and mitochondrial membranes (Speed et al., 1995). BPPS-SIPRIS results: EEP spherical cluster, \( p = 5.8 \times 10^{-13} \); INPP5 spherical cluster, \( p = 3.9 \times 10^{-7} \); INPP5B spherical cluster, \( p = 0.0021 \). (B) INPP5 hydrogen bond network within human INPP5B (pdb: 3mtc) (unpublished). (C) View of INPP5-residues (in 3mtc) that bind the 4-phosphate group required for substrate recognition. (D) Human INPP5B with phosphate bound to a possible membrane interaction or allosteric site (Mills et al., 2016). (E) Human INPP5B Ocr1 with glycerol bound to the same site as indicated in (D) (Tréaugues et al., 2014). (F) INPP5 subgroups within the BPPS-defined hierarchy. (G) Human INPPSE (pdb: 2xsw) (unpublished), which is associated with the primary cilium, an organelle involved in signal transduction (Jacoby et al., 2009) (spherical cluster, \( p = 3.6 \times 10^{-7} \)). (H) Human SHIP2 (pdb: 4a9c) (Mills et al., 2012), which is associated with membrane ruffle formation (Hasegawa et al., 2011) (spherical cluster, \( p = 0.30 \)). The following source data are available for Figure 5.

DOI: https://doi.org/10.7554/eLife.29880.012
Figure 6. BPPS-SIPRIS analysis of DNA glycosylases. (A). Thymine DNA glycosylase (TDG) family (red sidechains) and metazoan subfamily (orange sidechains) residues forming a significant hydrogen bond network ($p = 3.5 \times 10^{-8}$) within human TDG (pdb: 5hf7) (Pidugu et al., 2016). (B). TDG H-bond network consisting of residues distinctive both of all TDGs (red sidechains) and of metazoan TDGs (orange sidechains). This network includes hydrogen bonds to DNA oxygen atoms on either side of the thymine base to be excised (cyan); note that Phe238 and Tyr235 appear to position the N-terminus of their helix to hydrogen bond to substrate backbone oxygens; another such hydrogen bond involves Ser273, a residue generally conserved in the entire superfamily. The water molecule shown may act as the nucleophile in the reaction. For clarity, not all of the BPPS-SIPRIS-defined residues are shown. (C). TDG hydrogen-bond network residues may help position basic residues (green sidechains) interacting with the minor and major grooves of DNA. (D). TDG family-specific hydrogen-bond network residues surrounding a proposed catalytic water molecule (red sphere with dot cloud). (E). A BPPS-SIPRIS-defined H-bond network ($p = 1.7 \times 10^{-7}$) distinct from that of TDG within Thermus thermophilus uracil DNA glycosylase (UDG) (pdb: 2dp6). The following source data are available for Figure 6.

DOI: https://doi.org/10.7554/eLife.29880.014
**Figure 7.** Overview of BPPS-SIPRIS analysis. (A) Steps required for a BPPS-SIPRIS analysis. The fatax program adds phylum-annotations to database sequences. MAPGAPS detects and aligns database sequences containing the domain defined by a cma-formatted MSA or hiMSA. (MAPGAPS can also convert an MSA from fasta- to cma-format.) This creates an MSA that step 1 of BPPS then partitions hierarchically into subgroups based on discriminating pattern residues, as illustrated schematically in (B). Step E of BPPS checks for consistency between BPPS step 1 runs. Step 2 of BPPS adjusts the sub-alignment for each subgroup to align and possibly assign pattern residues to regions uniquely conserved in that subgroup, thereby creating a hiMSA. Step 3 of BPPS creates, for each node in the hiMSA, lineage-specific ‘contrast alignments’, as is illustrated schematically in (C), and a corresponding input file to SIPRIS, which identifies statistically significant structural interaction networks associated with pattern residues. For further descriptions, see text. (B) Schematic diagram of the node eight contrast alignment. Sequences assigned to node 8’s subtree (green subfamily nodes in (C)) constitute a ‘foreground’ partition; sequences assigned to the other nodes of the subtree rooted at the parent of node 8 (gray subfamily nodes in (C)) constitute a ‘background’ partition, and the remaining sequences constitute a non-participating partition. Green horizontal bars in (B) represent foreground sequences. The green vertical bars in (B) represent conserved foreground residue patterns (as shown below each bar), these diverge from (or contrast with) the background compositions at those positions (white vertical bars). Red vertical bars above quantify the degree of divergence. (C) Schematic diagram of a BPPS-3-generated set of ‘contrast alignments’ corresponding to the node 9 lineage of the sequence hierarchy in (A). Within a hiMSA, there is one such lineage for each leaf node. Horizontal lines represent aligned sequences and are colored by level in the hierarchy. Thin light gray horizontal lines represent non-homologous and deleted regions. Vertical lines represent the contrasting pattern positions upon which the hierarchy is based and are similarly colored by levels. The trees shown correspond to each subgroup along the lineage. The colored, gray and white nodes in each tree correspond, respectively, to their alignment foreground, background and non-participating partitions. The background for the entire superfamily (lower right) consists of standard amino acid frequencies at each position.

DOI: https://doi.org/10.7554/eLife.29880.016
Appendix 1—figure 1. Eleven haloacid dehalogenase sequences that the SFLD assigned to SG1129, but that are more closely related to SG1130 sequences. The Venn diagram shows the overlap between the subgroups BSG15, SG1129 and SG1130 with the numbers of sequences indicated. The table gives the mean pairwise gapped BLAST scores for the 11 sequences assigned to both SG1129 and BSG15 versus the sequence sets shown; this analysis indicates that the 11 sequences should be reassigned from SG1129 to SG1130. Similar analyses indicate that four other sequences in SG1129 should be reassigned to SG1135 (based on mean scores of 27 versus 139) and that a sequence in SG1136 should be reassigned to SG1137 (based on a mean score of 8 versus 149).

DOI: https://doi.org/10.7554/eLife.29880.022