Interactive analysis of functional residues in protein families

Morgan N. Price and Adam P. Arkin
Lawrence Berkeley National Lab

Abstract

A protein's function depends on functional residues that determine its binding specificity or its catalytic activity, but these residues are typically not considered when annotating a protein's function. To help biologists investigate the functional residues of proteins, we developed two interactive web-based tools, SitesBLAST and Sites on a Tree. Given a protein sequence, SitesBLAST finds homologs that have known functional residues and shows whether the functional residues are conserved. Sites on a Tree shows how functional residues vary across a protein family by showing them on a phylogenetic tree. These tools are available at http://papers.genomics.lbl.gov/sites.

Introduction

A protein can be thought of as a three-dimensional scaffold that places key functional residues, for binding or for catalysis, in the correct locations. Although these functional residues are critical for proteins' functions, they are not considered by most of the widely-used tools for annotating protein functions (i.e., (National Center for Biotechnology Information 2013; Overbeek et al. 2014; Seemann 2014; Haft et al. 2018)). The only exception that we are aware of is UniProt's UniRule, which records the active site residues for many protein families, and adds a warning if any of them are altered (Zaru et al. 2020).

If a protein is of particular interest, then manual analysis of its functional residues, as inferred from experimental studies of homologous proteins, is more effective than an automated approach. In particular, a human analyst can often find information in research articles or protein structures that is not represented in the annotation databases. However, doing these manual analyses is laborious. First, it's not obvious which homologs have experimental data about their functional residues. Second, once experimental data about key residues in a homolog is found, it can be quite cumbersome to identify the corresponding residue in the protein of interest. For proteins with structures, there's usually two different residue numberings, corresponding to the natural sequence and the portion of the sequence that was crystallized and whose structure was resolved; when reading a manuscript, it's not always obvious which coordinate system is being used. If multiple members of a family have been studied, a manuscript may use residue numbers from a reference protein instead of from the protein being studied. And key residues are sometimes shown as highlighted columns in alignments, but without residue numbers.

If the functional residues are partially conserved, it's helpful to see how those residues vary across the family. This is particularly useful for similar proteins with known function, as this can reveal if a change to a functional residue is likely to lead to a change in function. However, functional residues are often far apart in the sequence, so alignment viewers do not make it easy to view functional residues across a protein family.

To make it easier to examine the functional residues of a protein or a protein family, we developed SitesBLAST and Sites on a Tree. SitesBLAST compares a protein of interest to a large database of proteins with known functional residues. Sites on a Tree shows key residues across a protein family, along with a phylogenetic tree to show how the sequences are related to each other.

Results

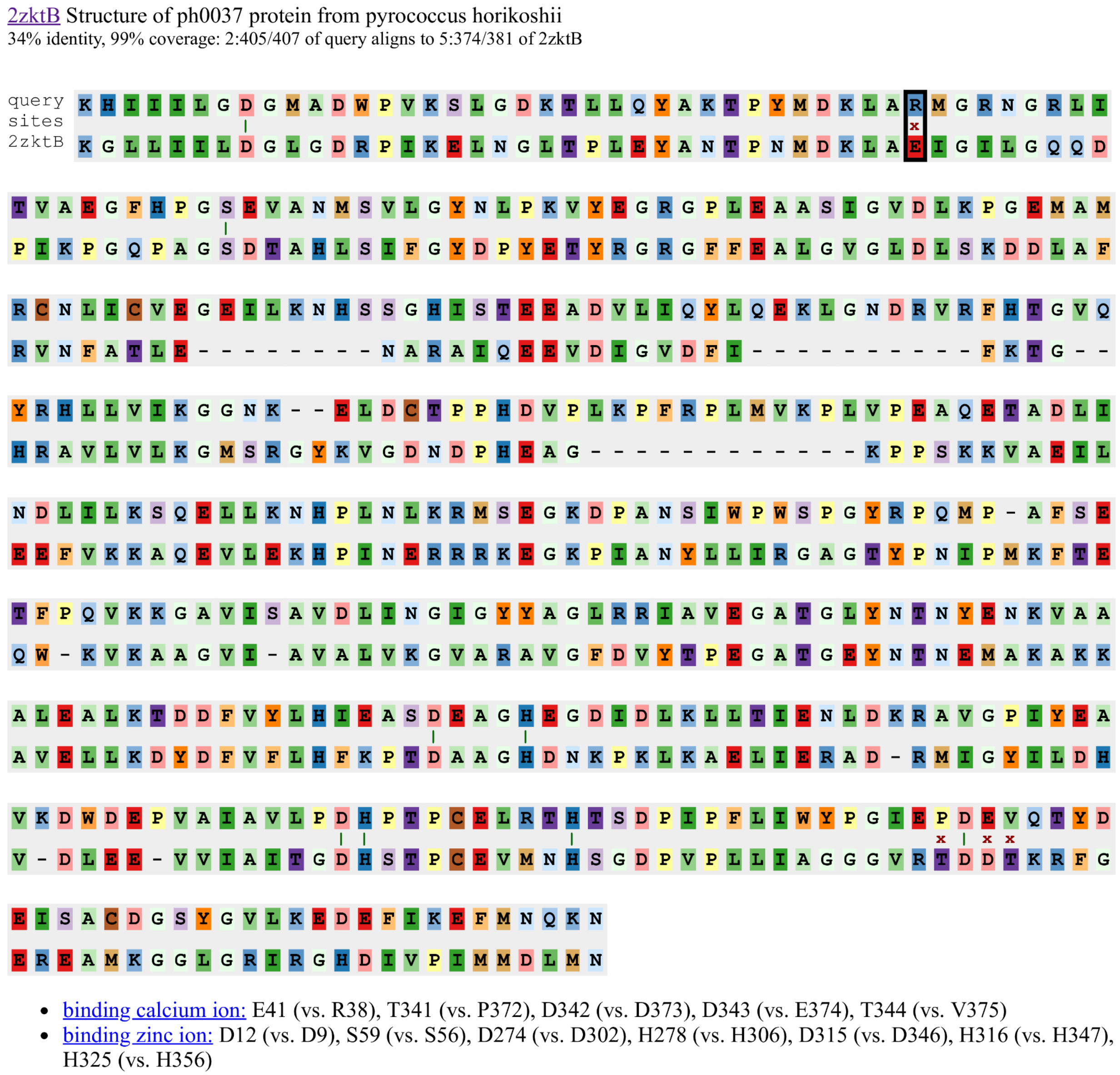
Experimentally-identified functional residues for 100,000 proteins

As of April 2022, SitesBLAST's database includes functional residues for 125,195 distinct protein sequences. SitesBLAST relies on two sources of functional residues: BioLiP (Yang et al. 2013) and Swiss-Prot. BioLiP incorporates protein-ligand interactions from protein structures in the Protein Data Bank (PDB), with biologically irrelevant ligands removed. BioLiP also includes active site residues if they are annotated in the PDB entry. In SitesBLAST's database, 94,655 distinct protein sequences have information from BioLiP. Most of the functional residues from BioLiP (97%) are involved in binding.

Swiss-Prot is the curated subset of UniProt (UniProt Consortium 2019) and includes many different kinds of "sequence features," along with evidence codes. SitesBLAST's database only incorporates features from Swiss-Prot if they have experimental evidence. Many of the features from Swiss-Prot (45%) indicate a covalent modification to the protein. These are included in SitesBLAST's database because they are often important for a protein's function, but many of them are not, strictly speaking, functional residues. Another 38% of the features from Swiss-Prot describe experimentally-mutated residues. These are annotated regardless of whether mutating the residue had an effect, so not all of the mutated residues are important for function. If modified and mutated sites are ignored, then the number of distinct protein sequences from Swiss-Prot with functional sites drops from 31,131 to 8,442.

SitesBLAST

At the SitesBLAST website, you can enter a protein's sequence or identifier. SitesBLAST will compare the query to its database, using protein BLAST, and will show up to 20 alignments. For example, as shown in Figure 1, the alternative homoserine kinase BT2402 is similar to the B chain from a crystal structure of a phosphoglycerate mutase. The structure includes two zinc ions and a calcium ion. As shown at the bottom of Figure 1, the zinc binding sites are fully conserved (for instance, D12 aligns to D9 in BT2402). In contrast, the calcium binding site is not conserved (i.e., E41 vs. R38).



2zktB Structure of ph0037 protein from pyrococcus horikoshii
34% identity, 99% coverage: 2:405/407 of query aligns to 5:374/381 of 2zktB

```
query K H I I I L G D G M A D W P V K S L G D K T L L Q Y A K T P Y M D K L A R M G R N G R L I
sites               |                                                         x
2zktB K G L L I I L D G L G D R P I K E L N G L T P L E Y A N T P N M D K L A E I G I L G Q Q D

T V A E G F H P G S E V A N M S V L G Y N L P K V Y E G R G P L E A A S I G V D L K P G E M A M
                  |
P I K P G Q P A G S D T A H L S I F G Y D P Y E T Y R G R G F F E A L G V G L D L S K D D L A F

R C N L I C V E G E I L K N H S S G H I S T E E A D V L I Q Y L Q E K L G N D R V R F H T G V Q
R V N F A T L E - - - - - - - - N A R A I Q E E V D I G V D F I - - - - - - - - - - F K T G - -

Y R H L L V I K G G N K - - E L D C T P P H D V P L K P F R P L M V K P L V P E A Q E T A D L I
H R A V L V L K G M S R G Y K V G D N D P H E A G - - - - - - - - - - - - K P P S K K V A E I L

N D L I L K S Q E L L K N H P L N L K R M S E G K D P A N S I W P W S P G Y R P Q M P - A F S E
E E F V K K A Q E V L E K H P I N E R R R K E G K P I A N Y L L I R G A G T Y P N I P M K F T E

T F P Q V K K G A V I S A V D L I N G I G Y Y A G L R R I A V E G A T G L Y N T N Y E N K V A A
Q W - K V K A A G V I - A V A L V K G V A R A V G F D V Y T P E G A T G E Y N T N E M A K A K K

A L E A L K T D D F V Y L H I E A S D E A G H E G D I D L K L L T I E N L D K R A V G P I Y E A
                                    |       |
A V E L L K D Y D F V F L H F K P T D A A G H D N K P K L K A E L I E R A D - R M I G Y I L D H

V K D W D E P V A I A V L P D H P T P C E L R T H T S D P I P F L I W Y P G I E P D E V Q T Y D
                            | |                 |                               x | x x
V - D L E E - V V I A I T G D H S T P C E V M N H S G D P V P L L I A G G G V R T D D T K R F G

E I S A C D G S Y G V L K E D E F I K E F M N Q K N
E R E A M K G G L G R I R G H D I V P I M M D L M N
```

- binding calcium ion: E41 (vs. R38), T341 (vs. P372), D342 (vs. D373), D343 (vs. E374), T344 (vs. V375)
- binding zinc ion: D12 (vs. D9), S59 (vs. S56), D274 (vs. D302), H278 (vs. H306), D315 (vs. D346), H316 (vs. H347), H325 (vs. H356)

Figure 1: An example alignment from the SitesBLAST website. Each aligned functional residue is highlighted with a dark green vertical bar if the two sequences match, or with a red x otherwise.

To make it easier to relate the binding sites to the alignment, SitesBLAST has interactive highlights. Hovering on a binding site at the bottom such as “E41 (vs. R38)” will highlight the corresponding location in the alignment (black box in the top row of Figure 1). Conversely, hovering on a functional residue in an alignment will highlight the same site at the bottom (not shown). Also, hovering on any alignment position will show the residue numbering in both sequences.

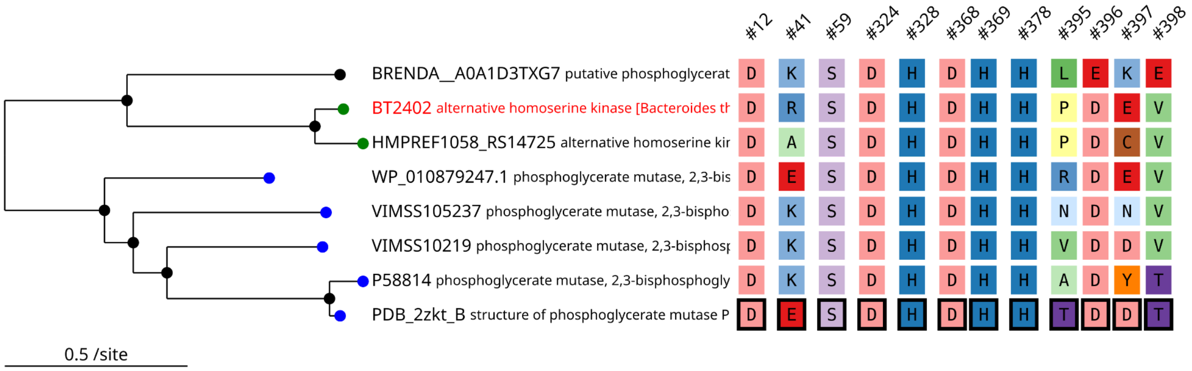
Overall, SitesBLAST takes just a few seconds to highlight potential functional residues, and whether or not they are conserved.

The coverage of SitesBLAST's database

To estimate the coverage of SitesBLAST, we selected 1,000 proteins at random from UniProt's reference proteomes and compared them to SitesBLAST's database using protein BLAST. 56% of the queries had hits ($E \leq 10^{-3}$), and 49% had hits with 30% identity or higher. (Homologs at 30% identity or higher are likely to have similar functions, and the alignments are likely to be accurate.) We checked a random sample of 40 of the proteins with hits of at least 30% identity. All but one of these had functionally-informative hits with known active site residues, residues that bind to substrates or other biologically-relevant ligands, or residues whose mutation leads to a loss of function. The final case (A0A2T6DQW1) was ambiguous: there are protein structures of homologs in complex with inhibitors or with ligands whose biological relevance is not proven. Overall, given a random protein that was predicted from a genome's sequence, SitesBLAST can identify potential functional residues about half of the time.

Sites on a Tree

Where SitesBLAST compares two sequences at a time, Sites on a Tree shows multiple sequences in a family. When considering how functional residues vary within a family and determine a protein's function, the most informative sequences are for proteins whose function is known. So, given a protein of interest, Sites on a Tree can identify homologs that have known functional sites (as in SitesBLAST) or whose function is known. The analyst can also add other proteins of interest to the list. Given these proteins, the website builds an alignment with MUSCLE 3 (Edgar 2004) and infers a phylogenetic tree with FastTree 2 (Price et al. 2010). Each of these steps usually takes a few seconds. Alternatively, the analyst can perform any of these steps themselves and upload unaligned sequences, an alignment, or a tree. Sites on a Tree supports sequences in fasta format, alignments in fasta, clustal, or Stockholm format, and trees in newick format.



	#12	#41	#59	#324	#328	#368	#369	#378	#395	#396	#397	#398
BRENDA__A0A1D3TXG7 putative phosphoglycerat	D	K	S	D	H	D	H	H	L	E	K	E
BT2402 alternative homoserine kinase [Bacteroides th	D	R	S	D	H	D	H	H	P	D	E	V
HMPREF1058_RS14725 alternative homoserine kir	D	A	S	D	H	D	H	H	P	D	C	V
WP_010879247.1 phosphoglycerate mutase, 2,3-bis	D	E	S	D	H	D	H	H	R	D	E	V
VIMSS105237 phosphoglycerate mutase, 2,3-bispho	D	K	S	D	H	D	H	H	N	D	N	V
VIMSS10219 phosphoglycerate mutase, 2,3-bisphosp	D	K	S	D	H	D	H	H	V	D	D	V
P58814 phosphoglycerate mutase, 2,3-bisphosphogly	D	K	S	D	H	D	H	H	A	D	Y	T
PDB_2zkt_B structure of phosphoglycerate mutase P	D	E	S	D	H	D	H	H	T	D	D	T

Figure 2: Sites on a Tree results for the alternative homoserine kinase BT2402. We used Sites on a Tree to automatically select characterized homologs, build a tree and alignment, and select known functional sites (highlighted with boxes). Other positions in the alignment are omitted. Each functional residue has hover text to describe its role. In the tree, leaf nodes are color coded by the protein's function.

Sites on a Tree can show the known sites (from SitesBLAST's database), or the analyst can choose which sites to show. As shown in Figure 2, when showing known sites, SitesBLAST highlights them with boxes. Sites on a Tree can identify known sites in uploaded sequences if, after removing gaps, the sequence is identical to a sequence in SitesBLAST's database.

Alternatively, the analyst can choose which residues to show by entering alignment positions or positions within the "anchor" sequence. (By default, the original query is the anchor.) To help the analyst find the correct residue number, Sites on a Tree can list all of the matches, across all of the sequences in the alignment, for subsequences or patterns such as NSG, CxxC, or DEA[DH].

The analyst can also customize the view. For example, in Figure 2, sequences with the same function have the same coloring in the tree; these colors were set by uploading a table with a color for each protein identifier. The uploaded table can also contain a description and a web link for each protein. Alternatively, there's a link to download the tree+sites graphic in SVG format, which can be edited in tools such as Inkscape or Adobe Illustrator.

Visualizing functional sites across hundreds of sequences

The tree+sites graphic (such as shown in Figure 2) works well for up to a few dozen sequences, but what if the analyst is studying a large family? If the analyst has chosen which residues to include, Sites on a Tree shows a more compact view (Figure 3). If a sequence position is conserved across a clade in a tree, then the amino acid code is drawn just once. If a single sequence or a small clade has a variant residue, and the clade is too small to draw the amino acid code, then only the color is shown. The compact view is only used when the analyst is choosing which sites to show, as there is no way to highlight the known functional residues in specific sequence in the compact view.

To ensure that variant residues can be seen in the compact view, Sites on a Tree needs to use a different color for every amino acid. Unfortunately, none of the standard color schemes for amino acids do this. We used the RColorBrewer library to select 12 paired colors and interpolated within each pair to get additional colors for each group of similar amino acids. The groups are: negatively charged residues (DE); small polar residues (ST); positively charged residues (NQKRH); aromatic residues (FWY); small hydrophobic residues (GAVLI); and other residues (PMC). Gaps in the alignment are shown in gray. For the larger groups (NQKRH and GAVLI), we altered the lighter color at the end of the scale, to make them easier to discriminate. Despite our best efforts, it can still be difficult to identify a residue by its color alone, but Sites on a Tree provides mouseover text and a zoom feature. Clicking on an internal node in the tree will navigate to a new view for just that clade, with enough vertical space to label each variant residue.

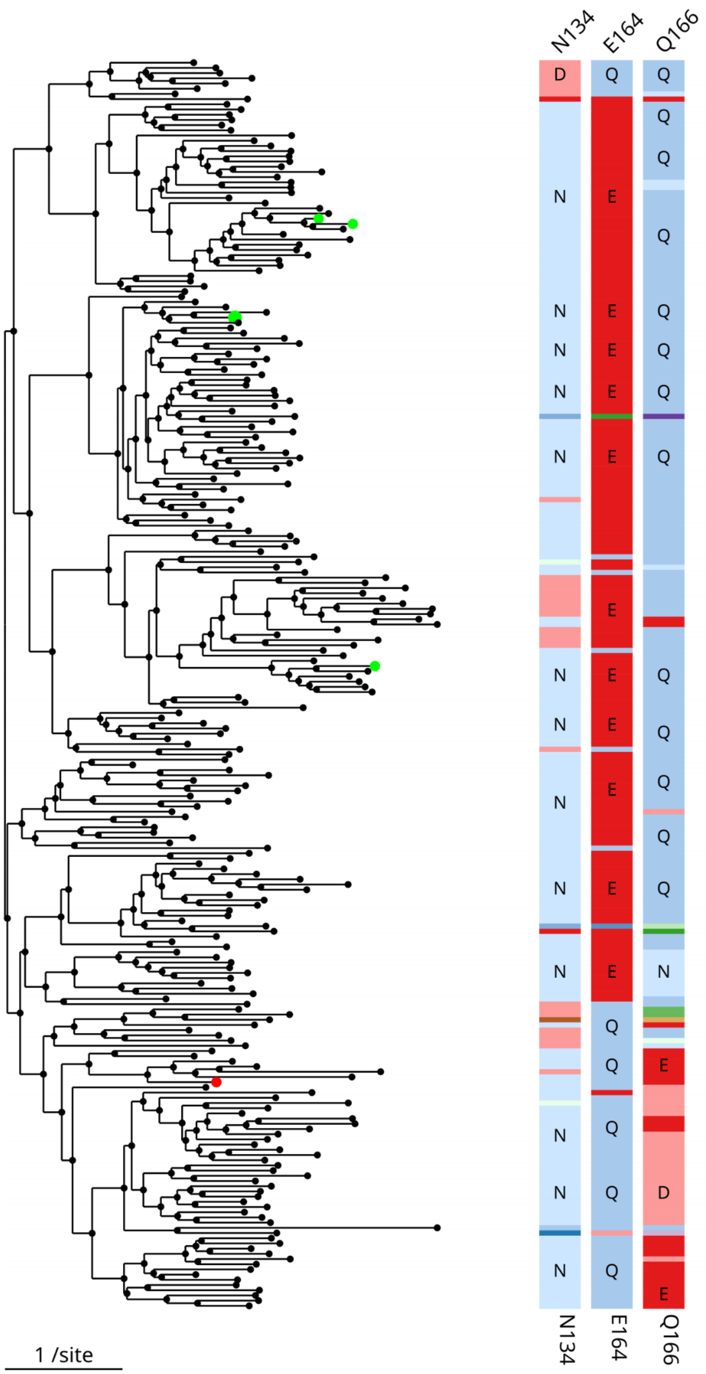


Figure 3: Putative active site residues for 240 sequences from the 3-ketoglycoside hydrolase family. Characterized members of the family are highlighted (colored nodes). The label of each column is based on the anchor sequence (BT3761).

If there are more than 30 sequences, then the compact view does not have space for the protein's identifiers. Instead, clicking on a leaf node will show the protein's identifier. Then, hovering on the identifier will show the description, and clicking on it will navigate to a page which includes sequences, links, and a list of known functional sites (if any). To locate family members of interest within the tree, Sites on a Tree has a search feature which highlights proteins whose identifiers or descriptions match the query. (The search feature was inspired by the ATV viewer (Zmasek and Eddy 2001).)

The example in Figure 3 has 240 sequences, which is about as much as will fit in one computer screen. For larger families, we recommend using CD-HIT or usearch (Li and Godzik 2006; Edgar 2010) to remove highly-similar sequences and hence to reduce the size of the tree. In fact, for our example, we began with all of the sequences from the 3-ketoglycoside hydrolase family from MicrobesOnline (Dehal et al. 2010) that have an alignment score of 80 bits or more against PFam PF06439.11 (Finn et al. 2014). We added a few more sequences for characterized proteins. This gave 646 sequences, or 2-3 screenfulls if viewed in Sites on a Tree. We clustered these at 60% identity to get a more manageable visualization of 240 sequences (Figure 3).

Discussion

Although functional residues are rarely taken into account during automated annotation, homologs with known functional sites are available for about half of all proteins. As far as we know, SitesBLAST is the first automated tool for finding this information. However, there's often more knowledge about the functional residues in the papers than in the databases. Conversely, ligand binding sites in protein structures may not be important for function inside the cell. So, when using SitesBLAST, it's important to read the paper that describes the protein structure (if there is one). We also recommend looking for additional relevant papers, for instance using PaperBLAST, which finds papers about a protein and its homologs (Price and Arkin 2017).

Sites on a Tree makes it easy to see how selected residues vary across a protein family. Sites on a Tree scales to hundreds of sequences, and can help the analyst select functional residues. Sites on a Tree may also be useful for visualizing putative functional residues that were identified by automated tools (i.e., (Laurie and Jackson 2005)).

SitesBLAST and Sites on a Tree are tools for exploration: they won't necessarily indicate a protein's function. For enzymes, if all of the key active site and substrate-binding are conserved, then the function is probably conserved as well. But it is often difficult to be sure that all of the key residues have been identified.

Methods

Data sources

Swiss-Prot and BioLiP were downloaded in April 2022. Sites on a Tree also uses a database of over 100,000 characterized proteins, taken from the characterized subset of the PaperBLAST database ((Price and Arkin 2019); we used the April 2022 release). UniProt reference proteomes were downloaded in May 2020.

SwissProt sequence features

SwissProt describes many different types of sequence features and not all of them are included in SitesBLAST's database. For protein modifications, we used CARBOHYD, CHAIN, CONFLICT, CROSSLNK, DISULFID, INIT_MET, LIPID, MOD_RES, NON_CONS, NON_STD, PEPTIDE, PROPEP, SIGNAL, TRANSIT, UNSURE, VAR_SEQ, and VARIANT features. But VARIANT features were ignored if the feature comment contains only a gene name, a strain name or a dbSNP reference. For binding, we used BINDING, CA_BIND, DNA_BIND, METAL, and NP_BIND features. For other functional features, we used ACT_SITE, MOTIF, REGION, and SITE features. MUTAGENESIS features were stored as a separate category. Sequence features of any type were only included if they were based on experimental evidence (evidence code ECO:0000269).

Phylogenetic trees

When inferring a phylogenetic tree, Sites on a Tree trims the alignment to remove columns that are ≥50% gaps or that have more lower-case than upper-case letters. (HMMer's hmmalign uses lower case for positions that are not actually homologous.) Either of these trimming steps can be disabled. The trimmed alignment is used to infer the phylogenetic tree, but is not used elsewhere (the site only shows the untrimmed alignment). After inferring a tree with FastTree 2, Sites on a Tree uses midpoint rooting to select the root of the tree.

If the analyst uploads a tree, then Sites on a Tree will treat the tree as rooted. Note that most tree inference tools produce unrooted trees, with the tree represented with an arbitrary root. (In a fully resolved tree, the root node has two children if the tree is rooted and three if the tree is unrooted; Sites on a Tree allows multifurcations, so it will accept either type.) Trees can be rerooted with tree editors such as FigTree, MEGA4, or phylip's retree.

Software and software settings

SitesBLAST's database is stored using sqlite3 (in the same database file as PaperBLAST's database) and as a protein BLAST database. SitesBLAST and Sites on a Tree are implemented in perl (version 5.16.3) and HTML 5. Sites on a Tree uses JavaScript for interactive highlighting. For Sites on a Tree, the tree and the aligned residues are rendered using SVG (scalable vector graphics).

SitesBLAST uses protein BLAST (version 2.2.18) with an E-value cutoff of 0.001 and filters the query sequence for lookup only (-F “m S”). Sites on a Tree uses the same settings but only reports homologs that cover at least 70% of the query. Also, by default, Sites on a Tree only includes homologs in alignments if they are at least 30% identical to the query. The identity cutoff ensures that sequence alignments are likely to be accurate. Also, more distantly-related sequences may have unrelated functions, in which case aligning functional residues may not be useful. If more distant sequences are included, we recommend using a structure-aware aligner such as MAFFT-DASH (Rozewicki et al. 2019) or an HMM-based aligner such as HMMer's hmmalign.

SitesBLAST uses MUSCLE v3.8.31 with fast options (-maxiters 2 -maxmb 1000) and FastTree 2.1.11 with default settings.

Availability of data and code

The code is available as part of the PaperBLAST code base (https://github.com/morgannprice/PaperBLAST). The SitesBLAST and PaperBLAST databases are updated every two months. The April 2022 database is archived at figshare (https://doi.org/10.6084/m9.figshare.20022590.v1). Instructions for downloading the current database are at https://github.com/morgannprice/PaperBLAST#Download.

Acknowledgements

This material by ENIGMA- Ecosystems and Networks Integrated with Genes and Molecular Assemblies (http://enigma.lbl.gov), a Science Focus Area Program at Lawrence Berkeley National Laboratory is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Biological & Environmental Research under contract number DE-AC02-05CH11231.

Bibliography

Dehal, P. S., M. P. Joachimiak, M. N. Price, J. T. Bates, J. K. Baumohl, D. Chivian, G. D. Friedland, K. H. Huang, et al. 2010. MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Research* 38: D396-400. doi:10.1093/nar/gkp919.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792–1797. doi:10.1093/nar/gkh340.

Edgar, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26: 2460–2461. doi:10.1093/bioinformatics/btq461.

Finn, R. D., A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, et al. 2014. Pfam: the protein families database. *Nucleic Acids Research* 42: D222-30. doi:10.1093/nar/gkt1223.

Haft, D. H., M. DiCuccio, A. Badretdin, V. Brover, V. Chetvernin, K. O'Neill, W. Li, F. Chitsaz, et al. 2018. RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Research* 46: D851–D860. doi:10.1093/nar/gkx1068.

Laurie, A. T. R., and R. M. Jackson. 2005. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* 21: 1908–1916. doi:10.1093/bioinformatics/bti315.

Li, W., and A. Godzik. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22: 1658–1659. doi:10.1093/bioinformatics/btl158.

National Center for Biotechnology Information, F. T.-N., Alexander Souvorov, Terence Murphy, Michael DiCuccio, Paul Kitts. 2013. Eukaryotic Genome Annotation Pipeline.

Overbeek, R., R. Olson, G. D. Pusch, G. J. Olsen, J. J. Davis, T. Disz, R. A. Edwards, S.

Gerdes, et al. 2014. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research* 42: D206-14. doi:10.1093/nar/gkt1226.

Price, M. N., and A. P. Arkin. 2019. Curated BLAST for genomes. *mSystems* 4. doi:10.1128/mSystems.00072-19.

Price, M. N., P. S. Dehal, and A. P. Arkin. 2010. FastTree 2 — approximately maximum-likelihood trees for large alignments. *Plos One* 5: e9490. doi:10.1371/journal.pone.0009490.

Rozewicki, J., S. Li, K. M. Amada, D. M. Standley, and K. Katoh. 2019. MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Research* 47: W5–W10. doi:10.1093/nar/gkz342.

Seemann, T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30: 2068–2069. doi:10.1093/bioinformatics/btu153.

UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* 47: D506–D515. doi:10.1093/nar/gky1049.

Yang, J., A. Roy, and Y. Zhang. 2013. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Research* 41: D1096-103. doi:10.1093/nar/gks966.

Zaru, R., M. Magrane, S. Orchard, and UniProt Consortium. 2020. Challenges in the annotation of pseudoenzymes in databases: the UniProtKB approach. *The FEBS Journal* 287: 4114–4127. doi:10.1111/febs.15100.

Zmasek, C. M., and S. R. Eddy. 2001. ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics* 17: 383–384. doi:10.1093/bioinformatics/17.4.383.