Published in final edited form as:

Curr Protoc Bioinformatics. ; 48: 2.10.1–2.10.16. doi:10.1002/0471250953.bi0210s48.

Using the Structure-Function Linkage Database to Characterize Functional Domains in Enzymes

Shoshana Brown¹ and Patricia Babbitt^{1,2,3}

¹Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA 94158-2330

²Chemistry, School of Pharmacy, University of California, San Francisco, San Francisco, CA 94158-2330

³the California Institute for Quantitative Biosciences, University of California, San Francisco, San Francisco, CA 94158-2330

Abstract

The Structure-Function Linkage Database (SFLD; <http://sfld.rbvi.ucsf.edu/>) is a web-accessible database designed to link enzyme sequence, structure and functional information. This unit describes the protocol by which a user may query the database to predict the function of uncharacterized enzymes and to correct misannotated functional assignments. It is especially useful in helping a user discriminate functional capabilities of a sequence that is only distantly related to characterized sequences in publicly available databases.

Keywords

protein superfamily analysis; protein sequence analysis; structure-function relationships; protein function prediction; annotation transfer

The Structure-Function Linkage Database (SFLD) is a web-accessible database designed to link enzyme sequence, structure and molecular function (Akiva et al., 2014; Pegg et al., 2006). Within the database, enzyme sequences are classified hierarchically into superfamilies, subgroups, and families (Figure 2.10.1). At the top of the hierarchy, distantly related enzymes within the same superfamily catalyze different overall reactions but are linked by a pattern of conserved amino acid residues in the active site that mediate a common chemical capability such as a partial reaction (Gerlt and Babbitt, 2001). At the

Corresponding Author: Patricia C. Babbitt. babbitt@cgl.ucsf.edu; Phone: (415) 476-3784.
Contact Information: Shoshana Brown. sbrown@cgl.ucsf.edu; Phone: (415) 502-1248

Internet Resources

<http://sfld.rbvi.ucsf.edu/>

The Structure-Function Linkage Database

<http://www.ncbi.nlm.nih.gov/gene/>

Get genome context information for a specific gene.

<http://www.microbesonline.org/>

Get operon context information for a specific gene.

http://www.theseed.org/wiki/Main_Page

Get operon context information for a specific gene.

bottom of the hierarchy, enzymes within the same family exhibit more sequence and structural similarities than proteins at the superfamily or subgroup level, and catalyze the same overall reaction via the same mechanism, mediated by a common set of catalytic residues.

The SFLD also provides sequence similarity networks, useful for providing large-scale summaries of the relationships within large groups of related enzymes, and for making hypotheses regarding the function of uncharacterized proteins. Such hypotheses can be further investigated using SFLD data and tools, for example, the mapping of specific chemical capabilities to specific sequence and structural motifs. (See Basic Protocol.)

Queries are currently limited to twelve superfamilies in the core SFLD (networks and detailed curation information available) and thirty-five additional superfamilies in the extended SFLD (networks and light curation information available). Additional superfamilies will be added as they can be curated.

BASIC PROTOCOL: USING THE SFLD TO PREDICT THE FUNCTION OF AN UNCHARACTERIZED ENZYME

The SFLD may be used as a tool for predicting the function of an uncharacterized enzyme by suggesting likely chemical capabilities for that enzyme based on overall sequence similarity to SFLD superfamilies, subgroups, or families and conservation of specific amino acid motifs associated with specific catalytic functions (Figure 2.10.1). If an uncharacterized protein can be confidently classified into an SFLD family, the overall reaction catalyzed by family members, including substrate specificity, can generally be assigned to it. Even if the uncharacterized protein can only be confidently classified at the superfamily level, the chemical capability conserved across the superfamily can still be used as a starting point to predict the overall reaction catalyzed by the protein.

Necessary Resources

Hardware—A computer running any common operating system (ex. Unix, Windows, Macintosh OS)

Software—An up-to-date web browser

Files—An amino acid sequence for the enzyme of interest in simple text or FASTA format (see Appendix 1B for a description of FASTA format). The amino acid sequence for the protein used as an example in this protocol corresponds to the National Center for Biotechnology Information (NCBI; <http://www.ncbi.nlm.nih.gov/>) protein database entry with gi number 6469462. To download this sequence, select “Protein” in the “All Databases” listbox at the top of the NCBI homepage, type the gi number into the adjacent search box, and click the “Search” button. In the results page that appears, select “FASTA” from the “Display Settings” listbox. A FASTA format sequence will be displayed in your web browser. Alternately, the fasta page for the sequence may be directly accessed by typing this URL into your web browser: <http://www.ncbi.nlm.nih.gov/protein/6469462?report=fasta>. The sequence may then be copied (by highlighting the sequence with your

mouse and selecting the “Copy” option under the “Edit” heading in the menu at the top of your web browser) and pasted into the SFLD as part of the protocol described below.

1. Open the SFLD homepage in your web browser using the following URL: <http://sfld.rbvi.ucsf.edu/>.

The SFLD homepage is displayed, as shown in Figure 2.10.2.

2. *Open the Search by Enzyme page.* Click the “Search by Enzyme” link at the top of the SFLD homepage.

The Search by Enzyme page is displayed.

3. Search for superfamilies, subgroups, and families homologous to a given protein. Paste the protein sequence (one letter amino acid code) into the Protein Sequence box of the Search by Enzyme page, click the “HMM” radiobutton (see Figure 2.10.3), and click the Search button below the sequence box.

The sequence is searched against a library of hidden Markov models (HMMs) constructed from hand-curated sequence alignments for superfamilies, subgroups, and families in the database. The search results appear in tabular format, sorted according to the SFLD hierarchy as shown in Figure 2.10.4 (rather than by the HMMER (Eddy, 1998) E-value). A lower E-value indicates a more significant match. E-values greater than 10^{-10} should be viewed as on the border of statistical significance.

For the sequence used in this example, the most significant match at the superfamily level is Enolase, the most significant match at the subgroup level is mandelate racemase, and the most significant match at the family level is D-galactonate dehydratase.

4. Check the protein for sequence conservation of superfamily, subgroup, or family-specific catalytic residues. Choose a family, subgroup, or superfamily of interest, and click the corresponding “Align to this...” link from the “Alignment to HMM” column. (The example given below represents the information obtained by clicking on the “Align to this Family” link for the D-galactonate dehydratase family, the most significant match according to E- value.)

The alignment appears in the middle panel of the display (Figure 2.10.5). The query sequence is the bottom sequence in the alignment. If there are more sequences than fit in the alignment panel, you may scroll down in the alignment panel to see the remaining sequences. If the alignment is too wide to fit on the screen, you may scroll to the right in the alignment panel to view the remaining alignment positions. The alignment coloring scheme can be changed by choosing an alternate selection from the “Use color scheme” listbox in the top panel of the screen.

Superfamily/subgroup/family-specific catalytic residues are listed in the bottom panel of the screen, along with their associated function, evidence code, and literature reference links. The evidence code, based on those developed by the Gene Ontology consortium (<http://www.geneontology.org/>)

www.geneontology.org/GO.evidence.shtml), describes the type of evidence used to determine the function of a given catalytic residue. Clicking on the associated comment bubble provides a popup with a detailed explanation of what the evidence code means. Clicking on the Reference icon provides a popup with the literature reference that was used, in part, to determine the function of the associated catalytic residue.

Scrolling to the right through the alignment shows that alignment positions corresponding to catalytic residues are automatically highlighted (Figure 2.10.5), allowing quick determination of whether a query sequence contains the machinery required for the superfamily/subgroup/family-specific functionalit.

In the example given, the protein sequence contains all the catalytic residues associated with the D-galactonate dehydratase family. This, in conjunction with the highly significant E-value match to the D-galactonate dehydratase family HMM, suggests that the enzyme may be a true member of the D-galactonate dehydratase family, catalyzing the D-galactonate dehydratase reaction.

5. Use superfamily, subgroup, or family functional information as a basis for predicting the function of the protein. Click the family name hyperlink at the top of the alignment display (Figure 2.10.5). (If you are looking at a subgroup or superfamily alignment, the link will display the appropriate subgroup or superfamily name.)

The SFLD page for the superfamily/subgroup/family is displayed (Figure 2.10.6 shows the page for the D-galactonate dehydratase family). At the top of the display, a “breadcrumb” trail of hyperlinks allows easily navigation between the different levels of the superfamily-subgroup-family hierarchy. In the middle left of the upper display panel, a crystal structure for a representative member of the superfamily/subgroup/family is displayed for groups containing one or more member enzymes with solved crystal. To the right of the structure box is summary information regarding the number of sequences, structures, and reactions in the SFLD for the superfamily/subgroup/family. The sequence, structure, and reaction counts are all hyperlinks. Clicking a hyperlink will display a table of the associated sequences, structures, or reactions. Figure 2.10.7 shows part of the sequence table for the D-galactonate dehydratase family.

The information shown in the sequence table may be modified by clicking the “Toggle Columns” tab, then selecting which annotation information to display or undisplay by clicking the associated button. (The annotation information associated with buttons with a yellow background color is set to display, while the annotation information for buttons with a white background color is not set to display.) The sequence table may be sorted by select annotation information by clicking the “Sort Set” tab, then clicking the button for the annotation you wish to sort by. The sequence

table may also be filtered by species name by clicking the “Filter Set” tab, entering a species name into the box, and clicking the “Filter” button. All annotation available for display in the sequence table may be downloaded by clicking the “TSV File” button on the top right of the display. The fasta sequences only may be downloaded by clicking the “FULL FASTA” button.

Back on the family/subgroup/superfamily display page (Figure 2.10.6), under the summary information and structure image is a short text description of the superfamily/subgroup/family. Clicking the “References” tab gives a list of literature references that may be useful as general references for the group. The final section of tabs in this top panel allows users to download sequence similarity networks (by clicking the “Download Network” tab, then clicking the button for the desired network), view a sequence alignment of the group (by clicking the “View Alignment” tab, then clicking the “View” button), align a sequence of interest to the curated group alignment (by clicking the “Align Sequences(s)” tab, pasting a sequence into the alignment box, and clicking the “Align” button), or download sequences and/or annotations for the group (by clicking the “Download Data Set” tab and clicking the button for the desired data set).

In the next panel, found only on pages for families with at least one solved crystal structure, an image of the active site of an example structure is shown, with family-specific catalytic residues labeled. Clicking on the active site image will open the image using the molecular visualization software package Chimera (Pettersen et al., 2004). (Chimera must be installed on a user's machine before this feature can be used. See <http://www.cgl.ucsf.edu/chimera/> for Chimera documentation.)

In the final display panel, found only on family pages, the overall reaction catalyzed by the family is displayed. (Subgroup and superfamily pages, in contrast, show a table listing subgroups and families found within the group.)

Because the example enzyme described in steps 1-4 above has been classified into the D-galactonate dehydratase family based on the earlier steps of the protocol, its putative function may be read directly from the reaction panel of the D-galactonate dehydratase family page (bottom panel, Figure 2.10.6). Putative catalytic residues may be identified by examining the highlighted residues in the query sequence when aligned to the D-galactonate dehydratase family (step 4 above).

See Guidelines for Understanding Results for further discussion about inferring the function of an uncharacterized protein using the SFLD.

ALTERNATE PROTOCOL: USING THE SFLD TO CORRECT MISANNOTATED FUNCTIONAL ASSIGNMENTS

Although computational prediction of protein function is required to bridge the gap between the number of sequenced genes and the number of experimentally characterized proteins, such techniques sometimes result in incorrect annotations. One study found that annotation error rates for large databases like GenBank NR and TrEMBL range from 5 - 63% depending on the group examined (Schnoes et al., 2009). Other estimates for annotation error rates range from 8% (Brenner, 1999), to 40% (Devos and Valencia, 2001). When proteins with incorrectly annotated function are used to make computational functional predictions for additional proteins, these inaccuracies may be further propagated (Gilks et al., 2002; Schnoes et al., 2009). The SFLD may be used to correct some instances of misannotation by showing that an enzyme does not share the properties of other characterized enzymes that perform its annotated function and/or by suggesting a more likely function for the enzyme in question.

Necessary Resources

Hardware—A computer running any common operating system (ex. Unix, Windows, Macintosh OS)

Software—An up-to-date web browser

Files—An amino acid sequence for the enzyme of interest in simple text or FASTA format (see Appendix 1B for a description of FASTA format). The amino acid sequence for the protein used as an example in this protocol corresponds to the National Center for Biotechnology Information (NCBI; <http://www.ncbi.nlm.nih.gov/>) protein database entry with gi number 392392608. To download this sequence, select “Protein” in the search listbox at the top of the NCBI homepage, type the gi number into the adjacent search box, and click the “Search” button. In the results page that appears, select “FASTA” from the “Display” listbox. A FASTA format sequence will be displayed in your web browser. Alternately, the fasta page for the sequence may be directly accessed by typing this URL into your web browser: <http://www.ncbi.nlm.nih.gov/protein/392392608?report=fasta>. This sequence may then be copied (by highlighting the sequence with your mouse and selecting the “Copy” option under the “Edit” heading in the menu at the top of your web browser) and pasted into the SFLD as part of the protocol described below.

Retrieve the current annotation for the enzyme of interest from the NCBI GenPept database

- 1 Open the NCBI homepage in your web browser using the following URL: <http://www.ncbi.nlm.nih.gov/>.
The NCBI homepage is displayed.
- 3 Select “Protein” from the “All Databases” listbox at the top of the screen and type the gi or accession number for the protein of interest into the adjacent search box (See Figure 2.10.8). Click the “Search” button.

A summary page for the protein of interest is displayed, as shown in Figure 2.10.9. In the example given, the NCBI GenPept database lists the function of gi|392392608 as “L-lysine 2,3-aminomutase.” In the remaining steps in this protocol, the SFLD will be used to either support or refute this functional assignment.

- 4 Open the SFLD homepage in your web browser using the following URL: <http://sfld.rbvi.ucsf.edu/>.

The SFLD homepage is displayed, as shown in Figure 2.10.2.

Browse to the SFLD family with the function assigned to the enzyme in question

- 5 Click the “Browse by Reaction” link at the top of the SFLD homepage.

A table listing the reaction catalyzed by each family in the database is displayed, as shown in Figure 2.10.10.
- 6 Scroll down on the Reaction Browse page until the reaction of interest is found. (Alternately, select the “Find” option under the “Edit” heading in the menu at the top of your web browser, and type the full or partial name of the reaction in the search box. For example, “lysine.”) Click the link for the corresponding family (“L-lysine 2,3-aminomutase”), in the 5th column from the right on the Reaction Browse page.

The family page is displayed. (See Figure 2.10.6 for an example family page.)

Align the enzyme of interest to the family

- 7 Click the “Align Sequences(s) tab at the bottom of the first display panel.

The “Align Sequences(s)” panel is displayed, as shown in Figure 2.10.11.
- 8 Paste the amino acid sequence for the enzyme of interest (single letter amino acid code) into the box. Click the “Align” button.

An alignment of the query sequence with a representative subset of the family is displayed. See Basic Protocol step 4 for an explanation of the alignment panel.
- 9 Check the alignment for conservation of family-specific catalytic residues.

Because the enzyme chosen for this example is annotated as “L-lysine 2,3-aminomutase” in the NCBI protein database, it is aligned against the L-lysine 2,3-aminomutase family. As shown in Figure 2.10.12, the example sequence (the bottom sequence in the alignment) appears to be missing the two aspartic acids required for binding the lysine substrate, indicating that it may not belong to the L-lysine 2,3-aminomutase family.

- 10 Search the SFLD for a more likely function for your protein using the Basic Protocol.

Although the results of this further analysis are not given here to avoid redundancy with the Basic Protocol, performing this analysis suggests the sequence is a member of the Glutamate 2,3-aminomutase family rather than the L-lysine 2,3-aminomutase family.

GUIDELINES FOR UNDERSTANDING RESULTS

When attempting to classify an uncharacterized sequence into an SFLD family, subgroup, or superfamily using the Search by Enzyme feature, the HMMER E-value may only be used as an approximate indication of whether a sequence might belong to a particular grouping because the sequences in different families/subgroups/superfamilies may have very different divergence rates. For example, the *o*-succinylbenzoate synthase family within the enolase superfamily contains sequences with quite distant relationships (Glasner et al., 2006) — thus a query sequence may belong in this family while exhibiting a relatively poor (high) E-value match to the family HMM. Other families within the same superfamily may contain sequences that are more similar over a comparable group of organisms — thus an E-value of the same magnitude may not indicate membership to another family.

Although the HMMER E-value cannot be used to determine the definitive family/subgroup/superfamily classification for a protein, the SFLD provides additional information that may be used to increase a user's confidence in such classifications. For example, a query sequence may be examined in the context of a family/subgroup/superfamily alignment and easily checked for the presence of catalytic residues, as described in Basic Protocol step 4. The position of an enzyme sequence within a family/subgroup/superfamily sequence similarity network may also be examined, as described in the Suggestions for Further Analysis section. Again, these analyses do not provide a definitive answer as to whether an enzyme is correctly classified — in particular, closely related families within the same superfamily may cluster tightly together in a sequence similarity network and have a similar or identical pattern of catalytic residues — but they do provide additional evidence that may help support or refute putative classifications.

Once a user is sufficiently confident that a protein has been correctly classified, the functional information stored in the SFLD may be used to infer a putative function for the protein. If the protein has been classified into an SFLD family, this inference is as simple as reading the family-specific reaction from the SFLD family page. In many cases, however, an uncharacterized enzyme may be classified into an SFLD superfamily but not a family. Here, ancillary information may be used along with the functional information stored in the SFLD to predict a function for the enzyme in question, in a process termed superfamily analysis. (See Superfamily Analysis in the Commentary section below.)

The increased accuracy of functional information obtained via the SFLD can be appreciated by comparing the results of the basic protocol for example sequence 6469462 to the corresponding NCBI GenPept annotation (<http://www.ncbi.nlm.nih.gov/protein/6469462>).

GenPept annotates the sequence as a putative isomerase, however the SFLD analysis suggests it is actually a D-galactonate dehydratase.

COMMENTARY

Background Information

Theories of enzyme evolution—Several theories have been advanced to describe the complex process of enzyme evolution. Horowitz suggested that ligand binding is the dominant constraint guiding enzyme evolution (Horowitz, 1945; Horowitz, 1965). According to his theory, biochemical pathways evolved backwards. When the substrate for an enzyme in the pathway was depleted, a new enzyme evolved from this enzyme by gene duplication to produce the needed substrate from an available precursor. While the reaction mechanism of the new enzyme diverged from that of its precursor to produce a new reaction, the ability to bind the common substrate/product was retained. Although this theory may apply to some groups of enzymes (Gerlt and Babbitt, 2001), it does not appear to be the dominant mechanism governing enzyme evolution (Teichmann et al., 2001; Todd et al., 2001). Chemistry-driven evolution (Babbitt and Gerlt, 1997; Jensen, 1976; Petsko et al., 1993), an alternate theory which appears to represent a substantial portion of enzymes (Rison et al., 2002), states that an underlying chemical capability is the dominant constraint guiding enzyme evolution. According to this theory, a newly evolved enzyme retains a fundamental chemical capability of its progenitor, while altering its ligand binding interactions and some step(s) in its overall mechanism to enable it to catalyze a completely different overall reaction (Gerlt and Babbitt, 2001).

Groups of enzymes which evolved according to chemistry-driven evolution, termed mechanistically diverse superfamilies, pose a particularly difficult problem for computational functional prediction, because related proteins may share only a single step or chemical capability rather than catalyzing the same overall reaction. Thus simple annotation transfer methods may lead to erroneous functional predictions. The SFLD offers users the tools required to perform a more refined functional prediction for members of such superfamilies via superfamily analysis.

Superfamily analysis—Superfamily analysis refers to the use of superfamily-specific functional information, along with ancillary information, to predict the function of an uncharacterized protein. An early example of the successful use of this strategy was the prediction that an uncharacterized open reading frame (ORF) in *Escherichia coli* encodes galactonate dehydratase — a prediction that was subsequently verified by experimental characterization of the enzyme in question (Babbitt et al., 1995). The analysis was performed roughly as follows. First, the uncharacterized ORF was determined, based on overall sequence similarity, to be a member of the enolase superfamily. Second, the superfamily membership of the ORF was verified based on examination of an alignment of the sequence to other superfamily members, which showed the presence of some superfamily-specific catalytic residues. Because the uncharacterized ORF had been classified as an enolase superfamily member, the reaction catalyzed by the associated enzyme could then be assumed to involve the abstraction of a proton attached to a carbon

adjacent to a carboxylic acid group to form an enolate ion intermediate — the chemical capability conserved across the enolase superfamily. This information was used, along with ancillary analyses suggesting the ORF is part of an operon involved with the utilization of acid sugars, to predict its function.

The SFLD is designed to facilitate the initial steps of superfamily analysis by:

1. Providing a quick indication of which families, subgroups or superfamilies in which a given protein may be a member, by matching the protein sequence to a library of HMMs.
2. Providing alignments of the protein to the families, subgroups, and superfamilies to which it might belong, with catalytic residues automatically highlighted, to facilitate the determination of whether or not the protein conserves the catalytic residues necessary to perform a particular molecular function.
3. Providing information about the chemical capability conserved across a given family or superfamily.
4. Providing links to databases that may provide ancillary information useful in determining protein function, such as genome or operon context. (See Suggestions for Further Analysis section.)

Critical Parameters and Troubleshooting

The SFLD currently contains twelve superfamilies in the core (highly curated) section and thirty-five additional superfamilies in the extended section (less extensively curated). If your enzyme sequence is not a member of one of the superfamilies in the database, you will not be able to use the database in the functional annotation of your protein. If your enzyme sequence is a member of the extended SFLD rather than the core SFLD, not all the data mentioned in the protocol may be available for your protein. In particular, you may find it more useful to utilize the BLAST search instead of the HMM search described in Basic Protocol step 3, by selecting the BLAST radio button rather than the HMM radio button (see Figure 2.10.3).

Also note that links given in this protocol may change over time. If a link is no longer valid, a Google search (<https://www.google.com/>) for the resource in question may help.

Suggestions for Further Analysis

Sequence Similarity Networks—As mentioned in the Guidelines for Understanding Results section above, sequences similarity networks provide an intuitive way to examine the relationships within a large group of related proteins. Placing an uncharacterized protein within a superfamily network that has been colored according to experimentally characterized proteins, for example, may suggest whether the sequence is likely to have the same or a similar function as a previously characterized protein (clusters tightly with characterized protein) or may represent a new function (is found far from any characterized proteins).

SFLD sequence similarity networks include a rich variety of annotation information, including species, SFLD family assignment, SFLD family assignment evidence code, Protein Databank ID, and SwissProt annotation. Networks may be painted with this annotation information, facilitating further analysis.

Although a protocol for the use of similarity networks and a discussion of their interpretation is beyond the focus of this article, tutorials describing the use of SFLD sequence similarity networks may be accessed by pasting the following URL into your web browser: http://sfld.rbvi.ucsf.edu/django/web/tutorial_links/.

Genomic Context—As mentioned in the Superfamily Analysis section above, ancillary information may be used along with information in the SFLD to infer the function of an uncharacterized protein. One particular type of ancillary information that has proven especially useful is genome or operon context information. This information can be found in several publicly accessible databases, including the NCBI Entrez Genomes database, the Microbes Online database, and the SEED database (see Internet Resources, below). SFLD enzymes with genome context are linked to this information in the SEED database and Microbes Online database.

For the sequence used as an example in the Basic Protocol given above (NCBI gi number 6469462), examination of genome context (see <http://pubseed.theseed.org/seedviewer.cgi?page=Annotation&feature=fig|100226.1.peg.3435>) shows a homolog of 2-deoxy-D-gluconate-3-dehydrogenase and a homolog of 2-dehydro-3-deoxyphosphogalactonate aldolase (which functions in the catabolism of galactonate (Deacon and Cooper, 1977)) located near the gene of interest, supporting the assignment of the galactonate dehydratase function to the corresponding protein.

ACKNOWLEDGEMENT

This work was supported by NIH R01-GM60595, NIH P41-RR01081, NIH P01-GM071790, NIH U54-GM093342, and NSF DBI-0640476.

Literature Cited

- Akiva E, Brown S, Almonacid DE, Barber AE 2nd, Custer AF, Hicks MA, Huang CC, Lauck F, Mashiyama ST, Meng EC, Mischel D, Morris JH, Ojha S, Schnoes AM, Stryke D, Yunes JM, Ferrin TE, Holliday GL, Babbitt PC. The Structure-Function Linkage Database. *Nucleic Acids Res.* 2014; 42:D521–530. [PubMed: 24271399] [Describes the SFLD.]
- Babbitt PC, Gerlt JA. Understanding enzyme superfamilies. *Chemistry As the fundamental determinant in the evolution of new catalytic activities.* *J Biol Chem.* 1997; 272:30591–30594. [PubMed: 9388188]
- Babbitt PC, Mrachko GT, Hasson MS, Huisman GW, Kolter R, Ringe D, Petsko GA, Kenyon GL, Gerlt JA. A functionally diverse enzyme superfamily that abstracts the alpha protons of carboxylic acids. *Science.* 1995; 267:1159–1161. [PubMed: 7855594] [Describes the use of superfamily analysis to elucidate the function of an uncharacterized ORF in *Escherichia coli*.]
- Brenner SE. Errors in genome annotation. *Trends Genet.* 1999; 15:132–133. [PubMed: 10203816]
- Deacon J, Cooper RA. D-Galactonate utilisation by enteric bacteria. The catabolic pathway in *Escherichia coli*. *FEBS Lett.* 1977; 77:201–205. [PubMed: 324806]
- Devos D, Valencia A. Intrinsic errors in genome annotation. *Trends Genet.* 2001; 17:429–431. [PubMed: 11485799]

- Eddy SR. Profile hidden Markov models. *Bioinformatics*. 1998; 14:755–763. [PubMed: 9918945]
- Gerlt JA, Babbitt PC. Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu Rev Biochem*. 2001; 70:209–246. [PubMed: 11395407] [Describes various mechanisms of enzyme evolution, including chemistry-driven evolution of mechanistically diverse superfamilies. Several mechanistically diverse superfamilies are discussed in detail.]
- Gilks WR, Audit B, De Angelis D, Tsoka S, Ouzounis CA. Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*. 2002; 18:1641–1649. [PubMed: 12490449]
- Glasner ME, Fayazmanesh N, Chiang RA, Sakai A, Jacobson MP, Gerlt JA, Babbitt PC. Evolution of structure and function in the o-succinylbenzoate synthase/N-acylamino acid racemase family of the enolase superfamily. *J Mol Biol*. 2006; 360:228–250. [PubMed: 16740275]
- Horowitz NH. On the Evolution of Biochemical Syntheses. *Proc. Natl. Acad. Sci. USA*. 1945; 31:153–157. [PubMed: 16578152]
- Horowitz, NH. The evolution of biochemical syntheses - retrospect and prospect.. In: B., V.; V.H., J., editors. *Evolving Genes and Proteins*. Academic Press; New York: 1965. p. 15-23.
- Jensen RA. Enzyme recruitment in evolution of new function. *Annu Rev Microbiol*. 1976; 30:409–425. [PubMed: 791073]
- Pegg SC, Brown SD, Ojha S, Seffernick J, Meng EC, Morris JH, Chang PJ, Huang CC, Ferrin TE, Babbitt PC. Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. *Biochemistry*. 2006; 45:2545–2555. [PubMed: 16489747]
- Petsko GA, Kenyon GL, Gerlt JA, Ringe D, Kozarich JW. On the origin of enzymatic species. *Trends Biochem Sci*. 1993; 18:372–376. [PubMed: 8256284]
- Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera—a visualization system for exploratory research and analysis. *J Comput Chem*. 2004; 25:1605–1612. [PubMed: 15264254]
- Rison SC, Teichmann SA, Thornton JM. Homology, pathway distance and chromosomal localization of the small molecule metabolism enzymes in *Escherichia coli*. *J Mol Biol*. 2002; 318:911–932. [PubMed: 12054833]
- Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS computational biology*. 2009; 5:e1000605. [PubMed: 20011109]
- Teichmann SA, Rison SC, Thornton JM, Riley M, Gough J, Chothia C. The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli*. *J Mol Biol*. 2001; 311:693–708. [PubMed: 11518524]
- Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol*. 2001; 307:1113–1143. [PubMed: 11286560]

Figure 2.10.1.

SFLD classification hierarchy. Sequences of unknown function are classified at a given level of the hierarchy only if sufficient evidence exists to support that classification. A sequence can be a member of a superfamily or subgroup even if its specific reaction family is unknown. This avoids “over-annotation” of a sequence to a more specific function than the available evidence warrants.

Home About SFLD Documentation ▾ Tutorials Contact Us Curator's Entrance

Browse by Superfamily Browse by Reaction Search by Enzyme

Structure Function Linkage Database

YRAAPLC...NGDPDDLIL.KLADHSEKVARNVGL...
 YHTAPLC...YGDPELYA.KLADHSEKVARNVGL...
 PNSLSYCQLLPTGEAALDYL.ASNP...
 RHE...EALTPEAIVRLAESAQERYGQSEKMG...
 RHE...EALTPEAIVRLAESAQERYGQSEKMG...
 RHE...EALTPEAIVRLAESAQERYGQSEKMG...
 RHQ...KAMNSEAVVRLAESAQERYGQSEKMG...

See description in *Nucleic Acids Res.* 2014 Jan 1;42:D521-30 [↗](#)

What is the Structure-Function Linkage Database (SFLD)?

- A hierarchical classification of enzymes that relates specific sequence-structure features to specific chemical capabilities
- A collection of tools and data for investigating sequence-structure-function relationships and hypothesizing function
- The analysis and archive site for superfamilies targeted by the Enzyme Function Initiative [↗](#)
- More...

How can I use the SFLD?
(see the [tutorials](#) for examples)

- **Classify a sequence** using Hidden Markov Models or BLAST search
- **Browse superfamilies in the SFLD**
- **Browse reactions** (overall)
- **Search for a specific enzyme** (by name, sequence database ID, or PDB ID)

What makes the SFLD unique?

- Superfamilies are defined by a conserved chemical capability such as a [partial reaction](#), families by a conserved [overall reaction](#) ([more...](#))
- Conserved [partial reactions](#) are correlated with associated active site similarities
- Large-scale summaries of relationships between and within groups of enzymes are provided as [sequence similarity networks](#)

Projects under Development

- Extended SFLD (XSFLD)
- Identifying Potential Misannotations

Figure 2.10.2.
The SFLD homepage.

Browse by Superfamily Browse by Reaction Search by Enzyme

Search by Sequence

Protein Sequence

```
EHHWQGMYRG
AFYRGGAVLCSAISGIDQALWDITARDLGVPVHALLGGAVRDSVRMYGHARGVTDFAFH
AARRSRSGM
TAIKFGLDGPVRALDTVEYVDROVARVAAREAVGPAFDIALDFHGRTPAMARRLAERL
SPMGPLFIEE
PCLPENVDALAEIARATNVPIAAGERIFTKWGFRELLERRAVSVVQPDLC HAGGISEVRR
IAAMAEAHYV
ALAPHNPLGVSLLAACLOVAASTPNFLIOEYLALDDGADLGAGILAVPFSVENGHIAVPT
GPGLGIEVDE
AQLMDRVYAGDWTTPRFTHQDGS HGDW
```

☐ BLAST ☒ HMM Search

Figure 2.10.3.

The SFLD Search by Enzyme page with a protein sequence pasted into the Protein Sequence box and the HMM radiobutton selected.

HMM hits

Select →	Toggle Columns	Query Sequence
Download	TSV File	Blast this sequence

Name	Level	E-Value	Score	Alignment to HMM
Enolase	superfamily	9.2e-90	296.1	Align to this Superfamily
└ mandelate racemase	subgroup	3.1e-105	346.8	Align to this Subgroup
└ mandelate racemase	family	7.5e-23	76.1	Align to this Family
└ D-tartrate dehydratase	family	2.5e-11	38.4	Align to this Family
└ L-tartrate/galactarate dehydratase	family	9.9e-35	114.8	Align to this Family
└ rhamnonate dehydratase	family	2.3e-46	152.9	Align to this Family
└ gluconate dehydratase	family	9.2e-68	223.7	Align to this Family
└ L-fuconate dehydratase	family	3.7e-22	73.2	Align to this Family
└ L-galactonate dehydratase	family	1.9e-17	57.8	Align to this Family
└ D-arabinonate dehydratase	family	9.9e-53	174.2	Align to this Family
└ D-galactonate dehydratase	family	7.7e-164	539.9	Align to this Family
└ muconate cycloisomerase	subgroup	5.7e-39	129.4	Align to this Subgroup
└ dipeptide epimerase	family	2.6e-19	64.2	Align to this Family
└ N-succinylamino acid racemase 2	family	7.6e-25	82.2	Align to this Family
└ muconate cycloisomerase (syn)	family	6.2e-35	115.9	Align to this Family
└ N-succinylamino acid racemase	family	1.1e-17	58.8	Align to this Family
└ o-succinylbenzoate synthase	family	6e-19	63.4	Align to this Family
└ muconate cycloisomerase (anti)	family	2.7e-18	60.9	Align to this Family
└ chloromuconate cycloisomerase	family	6.8e-33	109.4	Align to this Family
└ mannonate dehydratase	subgroup	1.6e-38	127.4	Align to this Subgroup
└ mannonate dehydratase	family	2e-33	110.7	Align to this Family
└ galactarate dehydratase	subgroup	2.1e-10	34.8	Align to this Subgroup
└ galactarate dehydratase	family	2.1e-10	34.8	Align to this Family
└ glucarate dehydratase	subgroup	1.1e-39	131.2	Align to this Subgroup
└ glucarate dehydratase	family	1.8e-11	38.2	Align to this Family

Figure 2.10.4.

SFLD HMM sequence search results. Results are displayed in tabular format, sorted according to the SFLD hierarchy.

Alignment of Family: D-galactonate dehydratase

Use color scheme: SFLD default view

Open raw MSA file.

This alignment was created by utilizing the hand-curated multiple sequence alignment of this Family and its corresponding Hidden Markov model (HMM). The former was created from a representative set of sequences and highlights the conserved residues that participate in this Family's enzymatic function(s). The table below lists those residues and their functions.

Figure 2.10.5. Alignment of a query sequence with the SFLD D-galactonate dehydratase family. The query sequence is the bottom sequence in the alignment. Family-specific catalytic residues are automatically highlighted, and their function and associated references are given in the table in the bottom panel.

Figure 2.10.6.
The SFLD family page for the D-galactonate dehydratase family.

Top Level

| Superfamily (core)

| Subgroup

| Family

| Sequence Set:

Name

Enolase

mandelate racemase

D-galactonate dehydratase

all sequences

Download 891 entries as

FULL FASTA

|

TSV File

Select →

Toggle Columns

Sort Set

Filter Set

Page 1 of 5 (1-200 of 891)

Next →

Last →

Protein Name	Superfamily	Family	Species ↓	Databases	MicrobesOnline Operon	The SEED	PDB ID	Structures	Updated
D-galactonate dehydratase	Enolase	D-galactonate dehydratase	Sinorhizobium melliloti 1021 ↓	<div>GIREF</div> <div>TrEFI</div> <div>MO</div> <div>TS</div> <div>MB</div>	131807	fig 1286640.3.peg.1937			Jul. 30, 2012
D-galactonate dehydratase	Enolase	D-galactonate dehydratase	Ralstonia solanacearum GMI1000 ↓	<div>GIREF</div> <div>SPEFI</div> <div>MO</div> <div>TS</div> <div>MB</div>	171028	fig 267608.1.peg.2751			Jul. 30, 2012
D-galactonate dehydratase	Enolase	D-galactonate dehydratase	Xanthomonas axonopodis pv. citri str. 306 ↓	<div>GIREF</div> <div>SP</div> <div>MO</div> <div>TS</div> <div>MB</div>	257639	fig 190486.1.peg.1737			Jul. 30, 2012
D-galactonate dehydratase	Enolase	D-galactonate dehydratase	Streptomyces coelicolor A3(2) ↓	<div>GIREF</div> <div>Tr</div> <div>MO</div> <div>TS</div> <div>MB</div>	243770	fig 100226.1.peg.3435			Jul. 30, 2012
D-galactonate dehydratase	Enolase	D-galactonate dehydratase	Thermobifida fusca ↓	<div>GIREF</div> <div>Tr</div> <div>EFI</div> <div>MO</div> <div>TS</div> <div>MB</div>	936817	fig 1169414.3.peg.1870			Jul. 30, 2012

Figure 2.10.7.

The SFLD sequence set table for the D-galactonate dehydratase family.

Figure 2.10.8.

The NCBI homepage with “Protein” selected in the “All Databases” listbox and gi number 392392608 specified as the protein sequence for which to search.

[Display Settings:](#) ☒ GenPept[Send to:](#) ☐

L-lysine 2,3-aminomutase [Desulfitobacterium dehalogenans ATCC 51507]

NCBI Reference Sequence: YP_006429210.1

[FASTA](#) [Graphics](#)[Go to:](#) ☐

LOCUS YP_006429210 414 aa linear CON 11-JUN-2013
DEFINITION L-lysine 2,3-aminomutase [Desulfitobacterium dehalogenans ATCC 51507].
ACCESSION YP_006429210
VERSION YP_006429210.1 GI:392392608
DBLINK Project: [82553](#)
BioProject: [PRJNA82553](#)
DBSOURCE REFSEQ: accession [NC 018017.1](#)
KEYWORDS RefSeq.
SOURCE Desulfitobacterium dehalogenans ATCC 51507
ORGANISM [Desulfitobacterium dehalogenans ATCC 51507](#)
Bacteria; Firmicutes; Clostridia; Clostridiales; Peptococcaceae; Desulfitobacterium.
REFERENCE 1 (residues 1 to 414)
AUTHORS Lucas,S., Han,J., Lapidus,A., Cheng,J.-F., Goodwin,L., Pitluck,S., Peters,L., Ovchinnikova,G., Teshima,H., Detter,J.C., Han,C., Tapia,R., Land,M., Hauser,L., Kyrpides,N., Ivanova,N., Pagani,I., Kruse,T., de Vos,W.M., Smidt,H. and Woyke,T.
CONSRTM US DOE Joint Genome Institute
TITLE Complete sequence of Desulfitobacterium dehalogenans ATCC 51507
JOURNAL Unpublished
REFERENCE 2 (residues 1 to 414)
CONSRTM NCBI Genome Project
TITLE Direct Submission
JOURNAL Submitted (11-JUN-2012) National Center for Biotechnology Information, NIH, Bethesda, MD 20894, USA

Figure 2.10.9.

NCBI GenPept display for the protein corresponding to gi number 392392608.

List of all Reactions

Download TSV File

Reaction Name	EC	Superfamily	Subgroup	Family	Functional Domain#
TDP-4,6-dideoxy-3-keto-D-glucose formation	1.-.-.-	Radical SAM	DesII	DesII Family	11
Formation of TDP-actinospectose	1.-.-.-	Radical SAM	Spectinomycin biosynthesis (SpcY)	SpcY Family	2
[Glycerol dehydratase]-activating enzyme	1.-.-.-	Radical SAM	Activating enzymes, group 2	Glycerol dehydratase activase	3
Ser-type sulfatase maturation enzyme	1.1.99.-	Radical SAM	Dehydrogenase Subgroup	Anaerobic Ser-type sulfatase-maturing enzyme	617
Cys-type sulfatase maturation enzyme	1.1.99.-	Radical SAM	Dehydrogenase Subgroup	Anaerobic Cys-type sulfatase-maturing enzyme	75
2-deoxy-scylo-inosamine dehydrogenase (AdoMet-dependent)	1.1.99.38	Radical SAM	SPASM/twitch domain containing	2-deoxy-scylo-inosamine dehydrogenase (BtrN)	4
Reduction of peroxide	1.11.1.1	Two Dinucleotide Binding Domains Flavoproteins	NADH peroxidase/oxidase and disulfide CoA reductase	NADH peroxidase	1617
Addition of a molecule of oxygen into sulfur and nitrogen based compounds	1.14.13.8	Two Dinucleotide Binding Domains Flavoproteins	monooxygenase	flavin containing monooxygenase	3647
Reduction of mercuric ion	1.16.1.1	Two Dinucleotide Binding Domains Flavoproteins	disulfide reductase	mercuric reductase	326
Reduction of adrenodoxin	1.18.1.2	Two Dinucleotide Binding Domains Flavoproteins	adrenodoxin reductase	adrenodoxin reductase	393
Reduction of ferredoxin	1.18.1.3	Two Dinucleotide Binding Domains Flavoproteins	NADH ferredoxin reductase	NADH ferredoxin reductase	2728
Botryococcus squalene synthase	1.3.1.96	Isoprenoid Synthase Type I	Squalene/Phytoene Synthase Like	botryococcus squalene synthase	1

Figure 2.10.10.

The SFLD Browse by Reaction page. Note that only part of the table is displayed.

Select Task → Download Network View Alignment **Align Sequence(s)** Download Data Set

Multiple Sequence Alignment

Align one or multiple sequences to the alignment of a representative set of sequences of this family.
Enter FASTA formatted sequence(s)

and view results using SFLD viewer

Curation
Curator: meng | Status: valid | Classification: public | Updated: Oct. 1, 2013 [EDIT](#)

Figure 2.10.11.
The SFLD Align Sequence(s) panel.

Position	Amino acid	Function	Curator notes	Evidence Code	References
210	Tyr (Y)	PLP binding	Tyr cited in MACIE. Also, PLP enzym (more +)	IEA	
213	Arg (R)	PLP binding		IEA	
222	Cys (C)	Binds [4Fe-4S]-AdoMet cluster		IDA	
226	Cys (C)	Binds [4Fe-4S]-AdoMet cluster		IDA	
229	Cys (C)	Binds [4Fe-4S]-AdoMet cluster		IEA	
231	Arg (R)	Binds lysine substrate		IEA	
297	Arg (R)	PLP binding		IEA	
393	Asp (D)	Binds lysine substrate		ISS	
430	Asp (D)	Binds lysine substrate		ISS	
437	Lys (K)	PLP binding, reaction nucleophile		IDA	

Figure 2.10.12.

Alignment of a query sequence with the SFLD L-lysine 2,3-aminomutase family. The query sequence is the bottom sequence in the alignment. Family-specific catalytic residues are automatically highlighted, and their function and associated references are given in the table in the bottom panel.