A Tour of the Structure-Function Linkage Database
and
Visualizing Sequence-Structure Relationships with Cytoscape

Shoshana Brown, PhD

August 16, 2010

The Structure-Function Linkage Database

[Homepage] Before we start the tutorial, I'm going to take you on a brief tour of the SFLD and cytoscape, just to get you oriented. Feel free to jump in with questions at any time. Since we don't have much time, there will be a lot of stuff I don't go into, but you'll learn a lot more about cytoscape in the tutorial and if you have any questions about the SFLD, feel free to ask me at any point.

The purpose of the SFLD is to link enzyme sequence and structural information to specific functional information. For the Glue grant, we want to provide a repository for the sequences, structures and existing functional information for each of the relevant superfamilies, and help guide target selection and provide context for the proteins that are being studied. As functions are determined by Glue Grant members and collaborators, we will add that information to the SFLD and use it to annotate other sequences that we expect to perform the same functions, based on stringent criteria.

[Sign in to db - then click Superfamily browse so it has time to load] You should all have or be receiving in the not too distant future, logins and passowords to sign into the private version of the SFLD. This will give you access to additional data that's not available through the public SFLD. This is data that's still undergoing curation. Once you're signed in, you'll see a message on the right side of the screen that tells you you're logged in. This will also allow you to curate data in the SFLD--for example, if you experimentally characterize some proteins in your superfamily, etc--via webforms. I'm not going to show those to you today, though, because the web forms are still under development. Anyway, I just wanted to mention that those webforms are in the works, and they'll be one of our top priorities.

Sequences in each SFLD superfamily are organized heirarchically, based on sequence, structure and function. At the top level of the heirarchy is an enzyme superfamily. If you look at this summary information for the superfamily you're interested in and think that there are less sequences and structures than you'd think, that's because we're still in the process of updating some of the GLUE grant superfamilies. If they're not already in the database, we should have the set of sequences and structures that we used to generate the cytoscape networks you'll work with today added to the database shortly. Classifying those sequences into subgroups and families is going to take longer, but you should be able to access basic information about each of the sequences and structures through the SFLD very quickly.

[Enolase SF] To show you the sort of information that we have in the database, I'm going to use the enolase superfamily as an example, because that is currently our best curated superfamily. So be aware that some of the information you see won't immediately be available for other glue grant superfamilies. At the top of the superfamily view we have basic information about the superfamily, including information about the chemical capability common to the superfamily, and an example superfamily structure.

[Expand toolbox] By clicking on the triangle next to toolbox, you can access things like a download link for all the sequences in the superfamily. You can also download a cytoscape network of the superfamily sequences. The problem is that right now, this network will contain a node for every sequence in the superfamily, and for large superfamilies, that can give you a network that's too big even to open. We're working on changing this so that you'll get what we call a metanode network, where each node is a group of closely related sequences, and that should be working in a few months. So in the relatively near future you should be able to download a functional, updated network of your superfamily via the SFLD using the link here. We're also going to add a link so that you'll be able to download structure networks as well. You can also see how a sequence you're interested in aligns with some representative members of the superfamily by pasting it in this box, or you can view a curated alignment of some representative superfamily members.

[Enolase SF alignment] Here's an alignment of some representative members of the enolase superfamily. The metal binding residues required for the partial reaction catalyzed by all superfamily members are highlighted on the alignment.

[Enolase SF] At the next level down in our organizational heirarchy we have the subgroup. The enzymes in the enolase superfamily are currently divided into seven different subgroups. Enzymes in a subgroup are more closely related to each other in sequence space than enzymes in a superfamily. Because the superfamilies we're working with are quite large--in some cases, over 20,000 sequences, it can be overwhelming to look at the entire superfamily, so in many cases, even if you want to examine a sequence of interest within a larger context, you may want to look at it in the context of its subgroup rather than the entire superfamily.

[Mandelate racemase subgroup; Expand toolbox] As you can see, at the subgroup level, we have all the same options you saw for the superfamily, including the ability to download cytoscape networks.

[Enolase SF] Now I'm back to the superfamily view. And you can see that at the level below the subgroup, we have enzymes classified into families. For the purposes of the SFLD, we've defined a family as a group of enzymes that catalyze the same overall reaction according to the same mechanism.

[galactarate dehydratase family] Here I'm showing you the galactarate dehydratase family page. The enzymes in this family are all thought to catalyze the reaction shown here, the dehydration of galactarate, using the active site residues shown in this picture.

[Chimera of Active Site] Clicking on the active site picture will open up the structure in Chimera, a structure visualization program developed by the RBVI at UCSF. In chimera, you can play around with the structure to get a better feel for how the important catalytic residues are positioned in the active site. We don't have time to go into all the features of chimera, but if you're not familiar with it and you're interested in it, I can give you a link to the RBVI website where there are tutorials and manuals.

[galactarate dehydratase family - expand toolbox] Just like at the superfamily and subgroup level, we can look at an alignment or download a cytoscape network. At the botton of the page, you'll see a list of the individual enzymes in the family.

[Check box--Make sure Modbase link and Operon link are selected; Sort by Family Assignment E Code] The checkbox section above the table controls what information is displayed and how it's sorted. I'll just go through a few things here so you get an idea of the type of information we have.

[Click on modbase link] For example, you can follow this link to see the modeled structures in modbase, from the Sali lab, corresponding to a sequence of interest.

[Click on SEED link] We also have a link to operon context information in the SEED database. And we're planning to add links to additional databases with operon context information like Microbes Online and IMG.

[X-ray structure count] You can also see which sequences have crystal structures.

[Click on assoc EFD] So, for example, if you were interested in getting more information about those structures, you could click on the entry for the associated sequence, where you'll see information regarding the catalytic residues in those crystal structures, and links so that you can easily open the structure in chimera. Also on the enzyme display page is information regarding the reaction catalyzed by the enzyme. Notice that there's an evidence code that descibes what sort of information was used to assign a particular reaction to a particular enzyme.

[E Code explanation page] Here the evidence code is IDA, meaning that the enzyme has been shown experimentally to catalyze the relevant reaction.

[Back to EFD page] When enzymes have been experimentally characterized, you'll see a link beside the evidence code that will get you to the associated literature reference. We have evidence codes and references for other things besides reaction assignment, like family assignment etc.

[Click on reaction] We also have information regarding the partial reactions associated with overall reactions.

Now, there are other ways to get into the database besides simply browsing through the heirarchy, like I've just shown you.

[Search by enzyme] You can search by GI number or PDB ID, or you search by amino acid sequence, using either BLAST or a search against the SFLD HMMs which are created based on the curated alignments we store for each family, subgroup, and sueprfamily.

So, I've hopefully shown you that the SFLD has a lot of useful information in it, but how can you best use that information for things like target selection or to get a quick overview of the biological context of a sequence or group of sequences you're interested in? In some cases you might want to go directly to the SFLD, but in other cases it might be nice to be able to look at your superfamily using cytoscape.

[Cytoscape] Cytoscape can be used to visualize any dataset that contains nodes and edges. The networks we're going to be looking at for most of the tutorial are going to be sequence networks where nodes represent a single sequence or a group of closely related sequences and the edges are blast e-values. For some superfamilies, we also have structural similarity networks, where each node represents a protein structure and the edges are N-scores from the FAST program.

There's a lot more to say about cytoscape, but I think the best way to learn about how to use it is to actually use it, so I'm going to let you jump into the tutorial now.

Visualizing Sequence-Structure Relationships with Cytoscape

Introduction

Many of the superfamilies we work with contain thousands of sequences and hundreds of structures. Given the size and diversity of these superfamilies, it can be difficult to use traditional methods, like multiple alignments and phylogenetic trees, to get an overview of the sequence/structure relationships in the superfamily. Network visualization via cytoscape allows the visualization of a much larger set of sequence/structure relationships than is possible using traditional methods. In addition, networks can be manipulated interactively as well as colored and filtered according to annotation information.

In this tutorial, we will primarily be looking at sequence similarity networks, where nodes represent either single protein sequences or groups of fairly closely related protein sequences and edges represent blast e-values.

For some superfamilies, we also have structural similarity networks available for use in this tutorial. In these networks, nodes represent protein structures and edges represent N-scores from the FAST structural superposition program.

In these exercises, you will examine a sequence similarity network for your superfamily of choice. The tutorial focuses on how sequence similarity networks may be useful for target selection, but networks can also be useful in many other contexts, so feel free to ask if you have particular questions/applications in mind.

Before the Tutorial

1. Please download and install Cytoscape on your laptop (This tutorial is written based on Cytoscape version 2.7.0. Earlier versions of cytoscape may not have all features and/or

may work slightly differently): http://www.cytoscape.org/download.php. Note: You may need to install an updated Java Runtime Environment before installing cytoscape. Updated JRE can be downloaded from: http://java.sun.com/javase/downloads/index.jsp.

2. Download the network(s) for your superfamily from http://babbittlab.compbio.ucsf.edu/glue.html. See the Information About Networks section for information about the network(s) available for your superfamily.
3. Because many of the networks are large, you may have to increase the memory allocation for cytoscape to open/manipulate the networks.
    a. On windows:
        i. Download the cytoscape.vmoptions file from http://babbittlab.compbio.ucsf.edu/glue.html and add to the same directory as the cytoscape.bat file.
        ii. Edit the cytoscape.vmoptions file according to the amount of memory you wish to allocate to cytoscape (see http://cytoscape.wodaklab.org/wiki/How_to_increase_memory_for_Cytoscape for more information) but be sure not to insert any carriage returns.
        iii. Start cytoscape by double clicking cytoscape.bat NOT Cytoscape.exe or the cytoscape icon.

    b. On mac/linux:
        i. In the Finder, right-click on the Cytoscape icon and select Show Package Contents.
        ii. Go to the Contents folder and open the file info.plist.
        iii. In the Property List Editor, expand the Root directory, then Java, and modify the VMOptions value. You may put multiple options separated by spaces here. You probably have to right-click the "VMOptions" entry and select "Value type -> Array", then click on the triangle in front of the "VMOptions" entry so that it points downward and on the icon at the end of the "VMOptions" to create a new item. Add a single JVM option for each item, e.g. "-Xms20M" for "Item 0" and "-Xmx2G" for "Item 1". (See: http://cytoscape.wodaklab.org/wiki/How_to_increase_memory_for_Cytoscape for more information)
        iv. Save and close the file.
        v. Start Cytoscape by double-clicking on the icon.

If you want to play with Cytoscape before the in-person tutorial, try this online tutorial to learn the basics: http://opentutorials.rbvi.ucsf.edu/opentutorials/index.php/Tutorial:Introduction_to_Cytoscape.

Using Cytoscape with Protein Similarity Networks

1. Open your sequence network (File-Open-NetworkFileName). The nodes (representing single protein sequences or multiple fairly closely related sequences) are laid out as a grid--not a particularly useful way to visualize them. We'll correct that in a minute.
2. Each node is associated with annotation information. To see the annotation information, select a node or group of nodes with your mouse, then select the annotation information you'd like to see using the top left button in the Data Panel.

3. In a minute we'll try to lay the nodes out in a more useful way. One of the ways we might judge whether the layout is useful is whether it places nodes in the same family near each other (where a family is defined as a set of evolutionarily related enzymes that catalyze the same reaction according to the same mechanism). The only network that currently has a large proportion of SFLD family classifications is the enolase superfamily network, so for the other networks, we'll use swissprot annotations as a proxy. To color the nodes according to swissprot annotation:
    a. Select the Vizmapper tab in the Control Panel.
    b. In the Current Visual Style box, select the button on the right and choose Create New Visual Style from the popup menu.
    c. Enter a name for your visual style in the Enter Visual Style Name box and press OK.
    d. In the Defaults box, select the default appearance for nodes and edges by clicking on the node or edge and then the property you want to change. One thing you might want to change here is NODE_FILL_COLOR. I like to set this to light gray--the white nodes can be hard to see. Note: Depending on the size of the network you're working with, some of your choices may not be reflected in the network (for example, if you select the default for NODE_SHAPE to be a triangle, you will still see the nodes in your network depicted as squares.) This is due to speed considerations. However, if you go to View-Show Graphics Details you will see the properties as you set them.
    e. In the Visual Mapping Browser section, you can map the appearance of nodes/edges to node/edge attributes. We'll just do one example. Scroll down to Node Color and double click in the right-hand rectangle.
    f. The Node Color entry should move to the top of the Visual Mapping Browser section. Click on Please select a value! and choose Swissprot.
    g. Click on Please select a mapping type and choose Discrete Mapper.
    h. You should see a list of all the unique values for the Swissprot node attribute. You can either choose colors for each of these values manually by clicking on the box to the right of the value, clicking on the "..." box, then choosing a color from the popup menu, or you can have cytoscape choose the colors for all of the values automatically by selecting "Node Color", then right clicking it, choosing Generate Discrete Values, and then choosing either Rainbow 1, Rainbow 2 or Randomize. The nodes should now all be colored according to their Swissprot attribute.
4. Experiment with different layouts to see which look useful for giving you an overview of the sequence relationships in the sequence set you're working with. Some layouts that might be particularly useful are organic (Layout-yFiles-Organic) and force directed (Layout-Cytoscape Layouts -Force Directed Layout-Blast E Value). How do these layouts position the nodes compared to your intuitive understanding of the sequence relationships in the superfamily? (It may be useful to look at the Swissprot coloring scheme or create additional coloring schemes based on other node attributes to address this question). Note: If you're working with a network containing a large number of nodes, even fast layouts like organic may take a minute or two.
5. For the purposes of target selection, we're interested in finding proteins that are distant from those that (1) already have a crystal structure, (2) are currently in the pipeline, and (3) have already been functionally characterized. You can change the visual properties of the network so that these proteins can be more easily identified.

6. We can do this in a couple of different ways:
    a. Color proteins that meet any of the above criteria a single color.
        i. Create a new node attribute to indicate proteins that have any of the three criteria we're interested in: Click on the middle button in the Data Panel, choose Boolean Attribute, and enter a name for the new attribute (ex. hasProp)
        ii. Set the hasProp attribute value to true for all nodes associated with a solved crystal structure.
            1. Select all nodes that are associated with a solved crystal structure:
                a. Select the Filters tab from the Control Panel
                b. In the Options listbox, choose Create New Filter.
                c. Type a name for your filter into the New Filter Name box and click OK.
                d. In the Attribute/Filter listbox, choose the node attribute you're interested in (node.pdbFileName).
                e. Click the "Add" button.
                f. In the listbox, type or choose the attribute value you're interested in. (Here, we want to select all nodes that have any value for this attribute, so we'll use the * wildcard.)
                g. Click the Apply Filter button. The node(s) selected by the filter will be highlighted in yellow on the network and displayed in the Data Panel.
            2. Click the attribute batch editor (Third button from the right in the Data Panel)
            3. In the popup window, choose "Set" in the first listbox, "hasProp" in the second listbox and "true" in the by/to box. Click GO. All nodes that are associated with a PDB structure will now have a value of "true" for their hasProp attribute.
        iii. Repeat step ii for all nodes currently in the pipeline (pipeline attribute) and all nodes that have been functionally characterized (use the swissprot attribute as a proxy.)
        iv. Create a new visual style that colors nodes according to their "hasProp" attribute, following the same process you used in step 3 to color by the "Swissprot" attribute.
    b. OR, if you want to be able to differentiate proteins that already have a crystal structure from those that are currently in the pipeline and those that have already been functionally characterized, you can choose a different visual mapping for each property. (For example, follow the process you used in step 3 to map color to nodes that already have a crystal structure, map size to those that are currently in the pipeline, and map shape to those that have already been functionally characterized.) The drawback to this technique is that size and shape aren't as easy to see on the network as color.
7. After setting up your visual styles, you should save them. Export your current visual styles via: File-Export-Vizmap Property File. This file can then be reloaded (in any network where you have the same node attributes) via File-Import-Vizmap Property File.
8. You should also save a copy of your network periodically: File-Save As

9. You now have a visual style that highlights nodes that have already been crystallized, functionally characterized, or are already in the structure pipeline. You can thus choose additional targets for the pipeline that are far from these nodes. However, how close to or far away nodes are from each other is dependent on the e-value cutoff used to create the network. Depending on your superfamily and the specific questions you're asking, you may want to use different e-value cutoffs. You can change the e-value cutoff used in the network by filtering edges according to their e-value.
    a. Select the "Network" tab in the Control Panel.
    b. Make a copy of the current network to play with: File-New-Network-Clone Current Network. You should see a second copy of the network in the Network browser. (You can toggle between the networks by selecting the network you're interested in.)
    c. Select the "Filters" tab on the Control panel.
    d. Click the "Option" button under "Current Filter" and select "Create new filter".
    e. Enter a name in the dialog box and click "OK".
    f. In the "Attributes/Filter" listbox and select edge.BlastEValue. Click the "Add" button.
    g. Double click the bar between the arrows under "Advanced".
    h. In the menu box, enter 0.0 for "Low bound" and "1e-34" for High bound". Click "OK". Check the box for "Not". Click "Apply". All edges with e-value cutoffs greater than 1e-34 will be selected.
    i. Delete the selected edges: Edit-Delete Selected Nodes and Edges.
    j. Redo your layout (see section 4, above). Note how this version compares to the original network (you can switch between the networks using the "Network" tab). If you'd like, you can examine the way your network changes at a series of increasingly stringent e-value cutoffs by following the protocol above, but substituting a series of more stringent e-values for 1e-34.
10. You may be particularly interested in choosing targets from a subset of the network.
    a. Select a subset of nodes you're interested in examining more closely by clicking and dragging the mouse to manually select a cluster. In choosing a cluster to examine more closely, it may help to:
        i. Browse the swissprot annotations
            1. Select all nodes, either by clicking and dragging the mouse or via Select-Nodes-Select all Nodes.
            2. Select the Swissprot defline attribute via the top left button in the Data Panel.
            3. Sort by swissprot defline by clicking on the header in the data panel.
            4. Scroll through the swissprot annotations via the arrows on the right side of the Data Panel. Clicking on any annotation you're interested in will highlight the corresponding node in green in the network.
        ii. Search for a sequence you're particularly interested in by uniprot identifier (see section 11c)
    b. Create a new network containing only this subset of nodes: File-New-Network-From Selected Nodes, All Edges.
    c. Lay out your new network using your layout algorithm of choice (See section 4).
    d. Some other things you might try:
        i. Filtering the edges to a more stringent e-value cutoff and laying out the network again to look at closer sequence relationships.

ii. Selecting nodes based on species (some species are easier to manipulate in the lab, etc). Due to functionality issues, we will use the old version of filters, accessible via Select-Use Old Filters.

1. In the popup box, click the Create New Filter button.
2. In the Filter Creation Dialog box, choose String Filter and click OK.
3. Under "Select graph objects of type", choose "Node".
4. Under "with a value for text attribute", choose "Species".
5. Under "that matches the pattern" type the attribute value you're interested in. (Note, you can use * as a wildcard. So, for example, if you want to filter for all species of the genus Escherichia, you can type "*escherichia*")
6. Click the Apply selected filter button.
7. The node(s) selected by the filter will be highlighted in yellow on the network and displayed in the Data Panel.
8. Note: You can do more complex searches by creating additional filters (following steps 1-5) and then combining them via the Boolean Meta-filter. For example, to search for all Escherichia proteins with a solved crystal structure, create another string filter (filtering for value "*" for the node attribute pdbFileName) and then create a Boolean Meta-filter and select both the species filter and the pdbFileName filter. Refer to the Information about Networks section for more node attributes that might be interesting for use in filters.

11. Because cytoscape networks are a two-dimensional representation of sequence similarity relationships, they are a simplification of the data and may sometimes be misleading. Thus, we suggest further investigation of any hypotheses made based on the sequence similarity network. One way you might do this is to look at a different type of network. If you have a structural similarity network available, look at it in conjunction with the sequence similarity network. Based on your knowledge of your superfamily, you should be able to rationalize differences between the sequence and structure-based networks and determine if these are artifacts of the relative coverage of the superfamily from structures and sequences or of missing "intermediate" nodes present in the sequence networks but not in the structure networks. Or, if the E-value cutoff you are using to look at the sequence network is not statistically significant, you may have spurious connections drawn in the sequence network for pairwise E-value scores close to that cutoff. Also note that because your networks are generated from custom databases containing only sequences that belong to your superfamily (rather than the whole GenPept database as is usually used to identify potentially homology sequences) the background model for generating the E-values is violated. We have found that for most superfamilies, this results in a skew in the meaning of your E-value of ~4 log units so that an E-value cutoff for your network of 10-5 may more likely represent an E-value of 10-1 if the comparisons were done against the whole GenPept database.
12. If you have time, some additional miscellaneous things you might want to play around with:

a. Cytoscape has many plugins that allow you to access additional functionality.

i. Browse the available Plugins:

        1. Go to: Plugins-Manage Plugins. The Manage Plugins box opens. You can browse to see which plugins are already installed and which are available for install.
        2. Expand the plugin category (ex. Analysis) and click on a specific plugin name to get a description of the plugin.

b. You may want an image of your network to use in a presentation or send to a colleague. You can save an image of your network via File-Export-Network View as Graphics. Many different file formats are available. Keep in mind that resolution may be an issue when using certain file formats (ex. png, jpg). Formats like pdf save the network as a vector graphic, so don't have the same resolution issues. However, for large networks, they can be very large and difficult to work with.

c. Searching for a single node/edge, using the search box at the top of the screen.
    - i. First, configure the search box according to the node/edge attribute you're searching for: Click the icon to the right of the search listbox, and select the attribute of interest in the Select Attribute listbox (for example, pdbFileName).
    - ii. Click Apply.
    - iii. Type the specific attribute name you want to search for in the Search box (for example, a specific PDB ID).
    - iv. Hit Enter. You should see a zoomed-in view of the network, with your sequence of interest, highlighted in yellow, in the middle. It will also be shown in the Data Panel at the bottom of the screen.

Information About Networks

1. Node attributes:
    - a. Sequence similarity networks:
        - i. ID: For nonmetanode networks, list the uniprot identifier or gi number corresponding to the node; for metanode networks, use the sequenceIDs node attribute to get this information
        - ii. swissprotDefline: Functional annotation from SwissProt (not available for enolase--instead, use family node attribute)
        - iii. family: SFLD family classification (only available for enolase)
        - iv. species: genus and species
        - v. pdbFileName: PDB ID(s)
        - vi. kingdom
        - vii. lineage2: kingdom and the next level of the NCBI taxonomy tree (useful for coloring networks according to lineage when the full lineage is too specific and the kingdom is too general
        - viii. lineage: The full lineage given by NCBI taxonomy (minus genus and species, which are found in the species node attribute)
        - ix. pipeline: Indicates proteins that are in pipeline for structure determination
        - x. rcdFromSGX: Indicates protein samples received from SGX.
        - xi. pfamFam: The name of the pfam family
        - xii. pfamFamAddl: The name of the pfam family corresponding to a domain that is not a member of the superfamily of interest (available for AH only)

xiii. cogs: The NCBI COG classification (available for AH only)
xiv. length: Number of amino acids of the full-length sequence
xv. sequenceIDs: Lists uniprot identifiers for all sequences represented by a given metanode (for metanode networks only)

b. Structural similarity networks:
   i. ID: PDB ID (or PDB ID.chain ID)
   ii. species: genus and species
   iii. annotation: Annotation from Swissprot or PDB
   iv. cath: The first 6 numbers of the CATH classification (available only for IS)
   v. capType: The cap type (as listed in Aravind et al 2006; for HAD only)
   vi. uniprotID or sfldGI: The uniprot identifier or NCBI GI number corresponding to the node

Note that for metanode networks, a given node attribute is a concatenated list of all unique attributes of the member sequences.

2. Typical edge attributes:
   a. Sequences similarity networks:
      i. BLAST e-value
   b. Structural similarity networks:
      i. FAST N-score
3. Superfamily specific information:
   a. Amidohydrolase
      i. ahClan.pfamDoms.1e-9.meta80.cys: Overall domain metanode sequence similarity network. The domain sequences defined by Pfam as part of the AH clan (January 2010; Note--this does not include the uronate isomerases) were downloaded. The sequence set was then filtered to remove identical sequences. An all-by-all BLAST of the remaining sequences was performed at an e-value cutoff of 1e-9. Sequences connected to each other with an e-value of 1e-80 or less were combined into a single metanode. Nodes represent one or more sequences connected via BLAST with an e-value of 1e-80 or less. Edges represent the most significant BLAST connection between any sequence in metanode X and metanode Y.
      ii. ah.fastNScore1.5.cys: Overall structure similarity network. An all-by-all comparison of each AH structure available in March 2010 was performed via FAST. Nodes are structures (single chain only). Edges are FAST N-scores of 1.5 or greater.
   b. Enolase
      i. allExceptEnolSub.1e-14.meta120.cys: Overall metanode sequence similarity network. Each full-length sequence from the SFLD Enolase superfamily (except sequences in the enolase subgroup) was downloaded. An all-by-all BLAST was performed at an e-value cutoff of 1e-14. Sequences connected to each other with an e-value of 1e-120 or less were combined into a single metanode. Nodes represent one or more sequences connected via BLAST with an e-value of 1e-120 or less. Edges represent the most significant BLAST connection between any sequence in metanode X and metanode Y.
   c. Haloacid dehalogenase

    i. HAD.filt.1e-14.meta80.cys: Overall metanode sequence similarity network. Each full-length sequence from the Pfam HAD clan (April 2010) was downloaded. The sequence set was then filtered (Sequences annotated in swissprot were filtered to remove duplicates. Remaining sequences were filtered to 80% ID). An all-by-all BLAST of the remaining sequences was performed at an e-value cutoff of 1e-14. Sequences connected to each other with an e-value of 1e-80 or less were combined into a single metanode. Nodes represent one or more sequences connected via BLAST with an e-value of 1e-80 or less. Edges represent the most significant BLAST connection between any sequence in metanode X and metanode Y.
    ii. HAD.fastNScore1.5.cys: Overall structure similarity network. An all-by-all comparison of each HAD structure available in May 2010 was performed via FAST. Nodes are structures (single chain only). Edges are FAST N-scores of 1.5 or greater.

d. Isoprenoid synthase
    i. IS.noPrenyltrans.nr.1e-14.meta120.cys: Overall metanode sequence similarity network. Each full length sequence in the Pfam Terpene_synth, Terpene_synth_C, TRI5, polyprenyl_synt, and SQS_PSY families (May 2010) was downloaded. The sequence set was then filtered to remove identical sequences. An all-by-all BLAST of the remaining sequences was performed at an e-value cutoff of 1e-14. Sequences connected to each other with an e-value of 1e-120 or less were combined into a single metanode. Nodes represent one or more sequences connected via BLAST with an e-value of 1e-120 or less. Edges represent the most significant BLAST connection between any sequence in metanode X and metanode Y.
    ii. IS.fastNScore1.5.cys: Overall structure similarity network. An all-by-all comparison of each IS structure (class I and class II) available in May 2010 was performed via FAST. Nodes are structures (single chain only; for structures containing multiple chains with different CATH superfamily classifications, a representative chain for both superfamilies is included in the network). Edges are FAST N-scores of 1.5 or greater.

e. GST
    i. GST.above150.filt.ref.1e-14.cys: Overall sequence similarity network. Each full length sequence in the Pfam GST_N and GST_C families (July 2010) was downloaded. Fragments (sequences < 150 AA) were removed from the set. The set was then filtered (swissprot sequences were filtered to remove identical sequences, remaining sequences were filtered to 60% ID). An all-by-all BLAST of the remaining sequences was performed at an e-value cutoff of 1e-14. Nodes represent sequences. Edges represent BLAST e-values of 1e-14 or more significance.

Appendix

1. Additional reading about sequence/structure similarity networks and the SFLD:

a. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC. PLoS One. 2009;4(2):e4345.
b. An atlas of the thioredoxin fold class reveals the complexity of function-enabling adaptations. Atkinson HJ, Babbitt PC. PLoS Comput Biol. 2009 Oct;5(10):e1000541.
c. Structural diversity within the mononuclear and binuclear active sites of N-acetyl-D-glucosamine-6-phosphate deacetylase. Hall RS, Brown S, Fedorov AA, Fedorov EV, Xu C, Babbitt PC, Almo SC, Raushel FM. Biochemistry. 2007 Jul 10;46(27):7953-62.
d. Target selection and annotation for the structural genomics of the amidohydrolase and enolase superfamilies. Pieper U, Chiang R, Seffernick JJ, Brown SD, Glasner ME, Kelly L, Eswar N, Sauder JM, Bonanno JB, Swaminathan S, Burley SK, Zheng X, Chance MR, Almo SC, Gerlt JA, Raushel FM, Jacobson MP, Babbitt PC, Sali A. J Struct Funct Genomics. 2009 Apr;10(2):107-25.

2. The FAST structural superposition algorithm: Zhu J, Weng Z. FAST: a novel protein structure alignment algorithm. Proteins. 2005 Feb 15;58(3):618-27.