ET TUTORIAL

RHONALD LUA

SHIVAS AMIN

ANGELA WILKINS

SERKAN ERDIN

OLIVIER LICHTARGE

Department of Molecular and Human Genetics

Baylor College of Medicine

Houston, Texas

Prepared for the Fundación Instituto Leloir

1. INTRODUCTION

This manual discusses the different ways to create, access and exploit Evolutionary Tracing of proteins (ET). The inputs to ET are an alignment, a tree, and optionally, and much better for interpretation when it is available, a structure. The output is a file with a rank of importance for each residue. Displaying this residue ranking hierarchy onto the structure brings the power of evolutionary analysis into full view.

We have created for ourselves and for others a series of tools to run ET and see its results. The goal here is to tell you about these tools so that you can use them for your own projects.

First, we will look at the simplest way to run ET, the ET_Report_Maker, which generates a trace and a manuscript describing its results automatically; then the ET_Viewer which is a downloadable java molecular display program to view and graphically manipulate ET results, but which is able to generate new traces with full user control via its ET_Wizard interface. Next, we use PyMOL which is a beautiful and elegant Molecular Graphics display created by Warren DeLano and that accepts python plug-in modules. PyETV is such a plug-in module; it brings the full functionality of a powerful graphics system to ET viewing, including viewing interfaces and complexes. UET, which stands for Unified ET, is a further step. It helps compare and contrast alternative distributions of importance ranks, typically arising by running ET on different selections on input sequences to easily generate Difference ET results and it benefits from Archeopteryx, a public tree viewer that we adapted for Difference ET. In very large protein families, special alignments might be required, such as in GPCR (G protein-coupled receptors).

These tools were built over many years by the members of the Lichtarge Lab. The code primary developers were Anne Philippi, Srini Madabushi, Ivana Mihalek, David Kristensen, Hui Yao and Angela Wilkins. Ivana Mihalek specifically wrote the Report-Maker, Dan Morgan the ET_Viewer, and Rhonald Lua the PyETV plug-in, which he is still enhancing. Rong Yao developed the Difference ET tool in UET.

Nearly all of this demonstration was developed by Rhonald Lua, with assistance from Angela Wilkins for the GPCR section and from Serkan Erdin for the ET annotation section, and with the input of Shivas Amin.

2. SIMPLE PRE-RUN ET

Zero. **ET REPORT MAKER**

This program is the simplest way to run real-value rvET. It will rank the importance of protein residues, identify significant clusters and putative functional surfaces on a structure, and it will suggest mutations that may impact function to various degrees. It also gathers background information about a protein from various sources: sequence, structure and elementary annotation, and superimpose inferences on the evolutionary behavior of individual residues on that background.

The sole input to the ET Report_Maker is a *Protein Data Bank identifier* or *UniProt accession number*, and it returns a human-readable document in PDF format, supplemented by the original data needed to reproduce the results quoted in the report. Dr. Ivana Mihalek designed this program.

TO RUN THE PROGRAM

A. Go to Internet URL: http://mammoth.bcm.tmc.edu/report_maker
B. Enter the PDB ID or UniProtID of interest as the *sole* input
C. Download the main output, which is the Report_Maker pdf document
D. You can also download the zip file that contains auxiliary files.

Report Maker usually has pre-generated reports that reside on the ET server for most PDB ID inputs. The coverage of UniProt IDs is much more spotty and requests may thus often trigger de novo trace runs by Report Maker—with no guarantee of success if the server does not automatically find the expected files in its internet queries. For demo purposes we will try sticking to PDB inputs, except for the one UniProt ID example below.

One. Photoactive yellow protein — example of a PDB ID input:.

i. Enter PDB ID "2phy" (photoactive yellow protein).
ii. Hit or click "Submit".
iii. Open link to PDF document "2phy_report.pdf"
iv. **PDF report includes**:
    1. Overview of the protein and the multiple sequence alignment.
    2. Residue rankings of relative evolutionary importance. Color scheme is described in the appendix.
    3. Geometric estimates of functional residues
        a. Figure 4 shows chromophore binding region
    4. Heuristic suggestions of disruptive mutations
    5. Possible novel functional surfaces
    6. List of sources, references, and data files accompanying the report

v. Download 2phy.zip, the link to the ancillary data files used in the report. The extracted directory once unzipped (2phy_report) should contain the following files:

    1. **2phy_report.pdf** -- the report in pdf format
    2. **2phyA.complex.pdb** -- coordinates of 2phyA with all of its interacting partners. This file can be viewed with PyMOL.
    3. **2phyA.etvx** -- ET viewer input file for 2phyA

4. **2phyA.cluster_report.summary** -- Cluster report summary for 2phyA
5. **2phyA.ranks** -- Ranks file in sequence order for 2phyA. The data in this file can be mapped to a PDB structure with the PyETV plugin on PyMOL.
6. **2phyA.clusters** -- Cluster descriptions for 2phyA
7. **2phyA.msf** -- the multiple sequence alignment used for the chain 2phyA. This alignment file, in GCG/MSF format, is derived from an HSSP alignment (Homology-derived StructureS of Proteins; R. Schneider, A. de Daruvar, C. Sander, 1996)
8. **2phyA.descr** -- description of sequences used in 2phyA msf
9. **2phyA.ranks_sorted** -- full listing of residues and their ranking for 2phyA (see notes below).
10. **2phyA.2phyHC469.if.pml** -- Pymol script
11. **2phyA.cbcvg** -- used by other 2phyA -- related pymol scripts

Which residues are most important?

Where are the functional sites?

What binds that site?

What is the ET rank of the amino acids at that site?

Which other (cognate) residues may occupy these same positions during evolution?

Can you guess at severely disruptive mutations of these residues?

Can you identify other functional sites? If so, where? Which residues?

Coverage: Trace results are commonly expressed in terms of coverage: the residue is important if its "coverage" is small - that is if it belongs to some small top percentage of residues [100% is all of the residues in a chain], according to the trace. The ET results are presented in the form of a table, usually limited to top 25% percent of residues (or to some nearby percentage), sorted by the strength of the presumed evolutionary pressure. (I.e., the smaller the coverage, the stronger the pressure on the residue.) Starting from the top of that list, mutating a couple of residues should affect the protein somehow, with the exact effects to be determined experimentally.

Known substitutions: One of the table columns is "substitutions" - other amino acid types seen at the same position in the alignment. These amino acid types may be interchangeable at that position in the protein, so if one wants to affect the protein by a point mutation, they should be avoided. For example if the substitutions are "RVK" and the original protein has an R at that position, it is advisable to try anything, but RVK. Conversely, when looking for substitutions which will not affect the protein, one may try replacing, R with K, or (perhaps more surprisingly), with V. The percentage of times the substitution appears in the alignment is given in the immediately following bracket. No percentage is given in the cases when it is smaller than 1%. This is meant to be a rough guide - due to rounding errors these percentages often do not add up to 100%.

Surface To detect candidates for novel functional interfaces, first we look for residues that are solvent accessible (according to DSSP program) by at least 10 $Å^2$**,** which is roughly the area needed for one water molecule to come in the contact with the residue. Furthermore, we require that these residues form a "cluster" of residues which have a neighbor within 5 Å from any of their heavy atoms.

Note, however, that, if our picture of protein evolution is correct, the neighboring residues which are not surface-accessible might be equally important in maintaining the interaction specificity - they should not be automatically dropped from consideration when choosing the set for mutagenesis. (Especially if they form a cluster with the surface residues.)

Number of contacts: Another column worth noting is denoted "noc/bb"; it tells the number of contacts heavy atoms of the residue in question make across the interface, as well as how many of them are realized through the backbone atoms (if all or most contacts are through the backbone, mutation presumably won't have strong impact). Two heavy atoms are considered to be "in contact" if their centers are closer than 5 Å.

Annotation: If the residue annotation is available (either from the pdb file or from other sources), another column, with the header "annotation" appears. Annotations carried over from PDB are the following: site (indicating existence of related site record in PDB), S-S (disulfide bond forming residue), hb (hydrogen bond forming residue, and sb (for salt bridge forming residue).

Mutation suggestions: Mutation suggestions are completely heuristic and based on complementarity with the substitutions found in the alignment. Note that they are meant to be disruptive to the interaction of the protein with its ligand. The attempt is made to complement the following properties: small [AV GST C], medium [LP NQDEMIK], large [WF Y HR], hydrophobic [LP V AMWF I], polar [GT CY]; positively [KHR], or negatively [DE] charged, aromatic [WF Y H], long aliphatic chain [EKRQM], OH-group

possession [SDET Y], and NH2 group possession [NQRK]. The suggestions are listed according to how different they appear to be from the original amino acid, and they are grouped in round brackets if they appear equally disruptive. From left to right, each bracketed group of amino acid types resembles more strongly the original (i.e. is, presumably, less disruptive) These suggestions are tentative - they might prove disruptive to the fold rather than to the interaction. Many researchers will choose, however, the straightforward alanine mutations, especially in the beginning stages of their investigation.

File formats: Files with extension "ranks sorted" are the actual trace results. The fields in the table in this file:

- alignment# number of the position in the alignment
- residue# residue number in the PDB file
- type amino acid type
- rank rank of the position in older version of ET
- variability has two subfields:
    1. number of different amino acids appearing in in this column of the alignment
    2. their type
- rho ET score - the smaller this value, the lesser variability of this position across the branches of the tree (and, presumably, the greater the importance for the protein)
- cvg coverage - percentage of the residues on the structure which have this rho or smaller
- gaps percentage of gaps in this column

Two. DNA glycosylase — example of UniProt ID input:

i. Enter "P05523" (DNA glycosylase)
ii. Hit or click "Submit".
iii. Open link to PDF document "P05523_report.pdf"

In this example, Report Maker finds a close structural model for the sequence: PDB 1k82A in region 1-268 with 94% sequence identity match.

iv. Can you tell
    1. What the function of the protein is?
    2. Some statistics on the alignment?
    3. What are the most important residues?
    4. Where are functional surfaces?
    5. Can you suggest useful disruptive mutations

6. Are there potential novel functional surfaces

v. Data files in P05523.zip:
   1. P05523_report.pdf -- the report in pdf format
   2. P05523.seq -- the query (target) sequence; P05523 in fasta format
   3. 1k82A.complex.pdb -- coordinates of 1k82A with all of its interacting partners
   4. 1k82A.etvx -- ET viewer input file for 1k82A
   5. 1k82A.cluster_report.summary -- Cluster report summary for 1k82A
   6. 1k82A.ranks -- Ranks file in sequence order for 1k82A
   7. 1k82A.clusters -- Cluster descriptions for 1k82A
   8. 1k82A.msf -- the multiple sequence alignment used for the chain 1k82A
   9. 1k82A.descr -- description of sequences used in 1k82A msf
   10. 1k82A.ranks_sorted -- full listing of residues and their ranking for 1k82A
   11. 1k82A.1k82AZN450.if.pml -- Pymol script
   12. 1k82A.cbcvg -- used by other 1k82A -- related pymol scripts
   13. 1k82A.1k82I.if.pml -- Pymol script
   14. 1k82A.1k82E.if.pml -- Pymol script

Three. Other examples: PDB 1a09, 1a22, 1fin, 16pk, 1nvt, 2pbl. UniProt P23284.

B. Some advantages
   i. Simple and convenient only needs a PDB ID or UniProt ID
   ii. Identifies relevant data from web
   iii. Integrates it with ET
   iv. Predicts key residues, functional sites, identifies ligands if possible, tallies substitutions and their frequencies, suggest a scale of mutations with decreasing impact.

C. Some limitations
   i. ET parameters pre-determined and fixed (i.e. no user input).
   ii. Cannot use own custom alignments. Alignments for PDB ID inputs default to HSSP-derived alignments.
   iii. Not yet and integrated with other tools in the Lichtarge lab.

3. THE ET VIEWER (ETV)

ETV provides a one-stop environment in which to run, visualize and interpret Evolutionary Trace (ET) predictions of functional sites in protein structures.
It runs real-value ET or integer ET and their displays results as a color map of the structure showing which residues are ranked among the top n^{th} percentile, where *n* is adjustable, and whether they cluster in a statistically significant manner (with a clustering *z*-score above 2). A multiple sequence alignment viewer and phylogenetic tree viewer display the underlying data.

Unlike the Report_Maker, ETV enables user controlled tracing via the ET Wizard, which takes a PDB identifier, or file, for input, and it outputs ranks of evolutionary importance for every sequence position in the protein. Through the ET Wizard, all trace parameters may be adjusted; custom alignments and phylogenetic trees may be used.

TO RUN THE PROGRAM

Launching the ET Viewer with a precomputed trace so that the structure and ranks are loaded automatically.

a. **Shikimate dehydrogenase — precomputed trace on *1nvt***

A. Go to http://mammoth.bcm.tmc.edu/traceview/index.html
B. Enter "1nvt" (**SHIKIMATE DEHYDROGENASE**). Click "Submit".
C. **Open** Click the link to the jnlp file ("1nvt") under "View Trace of"
   a. Allow ET Viewer to be launched by Java Web Start when prompted.
   b. Allow "ETViewer_2" access to your computer when prompted.

The output webpage will display links to pre-generated traces of UNIQUE chains in the PDB structure. All chains of the PDB are not necessarily displayed. For example, if the PDB is a homo-hexamer, only one chain (e.g. chain A) will be displayed.

D. To examine the structure and the trace
   i. Rotate molecule: Left-mouse-click on the graphics/structure viewer window and drag the mouse (or drag finger on mousepad) to rotate the structure
   ii. Vary Evolutionary Trace rank: Drag the slider (horizontal scrollbar above the graphics area) to vary the selection of ET residues. The ET rank threshold (Rho) and percent coverage change in this process.

Observe the variation in the values of the z-scores at the top of the ET Viewer window. This z-score is a measure of the spatial clustering of the selected trace residues. High z-scores correlate with good overlap between binding-site/core residues with the ET residues.

iii. Change representation: Under the **View** menu, you may also:
1. Select **Gobstopper Color** to assign rainbow/prismatic coloring of evolutionary importance
2. Select **Color by cluster** to color trace residues in order of cluster size. The largest cluster is in red, while individual residues are colored black
3. Select **Backbone** to show the structure's backbone
    a. *Select **Mode -> Bonds** and set the slider to low coverage to view the small tight cluster of ET residues.*
    b. Select **Residue Ranks** to show the ranks data file in a new window
    c. Select **ET Tree** to open the **ET Tree Viewer** and view the phylogenetic tree used to compute the trace
4. Selections of residues may be made through the **Edit -> Residue Selection** menu, or by making mouse clicks over the appropriate residues in the structure.

iv. Can you tell
1. What the function of the protein is?
2. Some statistics on the alignment?
3. What are the most important residues?
4. Where the functional surfaces are?
5. Can you suggest useful disruptive mutations?
6. Can you examine the underlying tree, msf, ranks?
7. Can you define an evolutionary core?
8. Are there potential novel functional surfaces?
9. Can you display ligands or dimers?

b. thioesterase — precomputed trace on *2pbl*

E. **For another example**, go back to http://mammoth.bcm.tmc.edu/traceview/index.html and enter "2pbl" (thioesterase), and repeat the above steps.

4. *DE NOVO* TRACE WITH ET WIZARD

REMINDER: Do not let attendees run a new trace during your demo, since your request would then get queued.

1. Requests to produce a *de novo* Evolutionary Trace analysis can be made using the **Utils -> ET Wizard** menu:

    a. Follow the steps as instructed in the series of dialogue boxes
    b. *You will have a chance to upload a PDB file, or specify a PDB ID*. When prompted for a **Structure File**, select **Download PDB file**, and enter as **PDB code** "1a09A". This fetches that file from the PDB, and extracts the correct chain (chain A).

Users must specify a chain indicator or chain letter. An ET analysis can be performed only for one specified chain in the PDB structure.

    c. Accept the defaults on the **Custom sequence list** and **Multiple sequence alignment** by clicking on **Next >**.

If you wish to use your own alignment, you may upload an alignment file in GCG/MSF format. The sequence of the structure you are mapping the results to MUST BE INCLUDED in the alignment.

    d. Enter the download path which is the directory location to save when prompted in the step **Provide download path**
    e. In the **Run Evolutionary Trace** step, you are about to submit your request.
    f. Click **Advanced** to view the various options available for several parameters that can be varied, including:

        i. the number of protein sequences (default is 500)
        ii. the e-value cutoff (default 0.05) of BLAST search (default 500)
        iii. the maximum number of homologs to use from the BLAST sequences (default is 500)
        iv. the minimum percent identity (default is 28%; for lower percent identity, the structures may diverge)
        v. the maximum percent identity (default is 98%).

    g. Click **Cancel** to close this window and accept the defaults, or click OK to accept your changes. Click **Finish** to start the trace
        i. You will receive a notification in the last dialogue box when the trace is complete
        ii. **Locate the zip file and unpack/extract its contents**:

1. View the trace with the ET Viewer: Select **File -> Open ETV Results**, and search for **1a09A.etvx**.
2. View the multiple sequence alignment by selecting **File -> Open Alignment File**, and locate **1a09A.msf**
    a. Clicking on sequence names in the **MSF Viewer** highlights the corresponding sequences in the **ET Tree Viewer**
    b. Selected sequences may be saved into a file. This file may be supplied as input to a new trace request (to ET Wizard or UET).
    c. Dragging the slider in the structure viewer highlights the evolutionarily importance columns in the **MSF Viewer**

h. The ET Viewer user manual is at http://mammoth.bcm.tmc.edu/traceview/HelpDocs/ETViewerManual_2.pdf

i. Advantages

i. **Easy installation and launching of the ET Viewer using any internet browser with Java Web Start.**
ii. **Integrates structure visualization with the multiple sequence alignment and the phylogenetic tree used to compute the trace.**
iii. **Provides an option to perform *de novo* ET analysis via the ET Wizard menu.**

j. Disadvantages

i. **Structure visualization is rudimentary (e.g. does not display a surface or a cartoon (helices and sheets) representation).**
ii. **Only one chain of a PDB structure can be displayed at any one time.**
iii. **Only one trace (or one ET rank file data) can be mapped to a structure. Not suitable for difference ET analysis.**
iv. **Molecules like DNA and other types of non-protein structures and ligands cannot be displayed.**

ET Wizard cannot be used without specifying a PDB ID, or supplying a PDB structure file. If no structure is available, the user may create and supply a bogus/fake/synthetic PDB structure file representing the sequence. There are programs from molecular dynamics packages (like myPresto) that can do this. Or, a program can easily be written or requested from the Lichtarge lab. However, a webservice to be described below (UET) run traces on sequence input.

The webserver that processes the requests made in ET Wizard can run at most 3 traces at the same time. The rest (at most 10) will be queued.

5. PyETV

PyETV is a PyMOL plugin for viewing, analyzing and manipulating predictions of evolutionarily important residues and sites in protein structures and their complexes. It seamlessly captures the output of the Evolutionary Trace server, namely ranked importance of residues, for multiple chains of a complex. It then yields a high resolution graphical interface showing their distribution and clustering throughout a quaternary structure, including at interfaces. Together with other tools in the popular PyMOL viewer, PyETV thus provides a novel tool to integrate evolutionary forces into the design of experiments targeting the most functionally relevant sites of a protein.

Download and install

a. To PyMOL, visit http://www.pymol.org/
b. To install the PyETV plugin, visit: http://mammoth.bcm.tmc.edu/pyetv
c. Instructional youtube videos are available at: http://www.youtube.com/user/EvolutionaryTrace

RUN

One. Basic viewing of SH2 domain with PDB *1a09A*

using the downloaded files from the previous example involving the ET Viewer/Wizard.

LOAD STRUCTURE

i. Run PyMOL, making sure that it appears with a Tcl/Tk window (a Python console/command line) that supports Plugins, and that PyETV is already installed.
ii. **1a09A**
iii. From the File menu, select **File -> Open** and locate the file "**query_1a09A.pdb**" in your directory. The PDB structure should appear in the PyMOL graphics window.
iv. **Hide (H) lines** and **show (S) cartoon** (or select spheres, etc).
v. Open the PyETV plugin window by selecting from the menu, **Plugin -> PyETV**. The PyETV window can be resized/enlarged by clicking on the edges and dragging the mouse (useful if not all input boxes and buttons are visible)

LOAD ET RANKS

i. On the **ET1** page of PyETV, enter "**query_1a09A**" into the "**structure to use**" box.
ii. Click on "**ET ranks file path:**" and locate the file "**ET_1a09A.ranks**".

Note that when the mouse-pointer hovers over the labels and controls Tool Tips appear that let you know more about what these boxes do.

VIEW ET RANKS

iii. Click the "**Map ranks to structure**" button. Bands of solid red color appear and indicate top-ranked ET residues on the secondary structure.

v. Adjust ET rank threshold: *click and drag the slider* just below the "**Map ranks to structure**" button

vi. Using the "**Show ET residues**" drop-down combo box, select **Prismatic**. Drag the slider all the way to the right.

vii. Using the "**Show ET residues**" drop-down combo box, select **as Spheres**. Drag the slider to the 30% rank threshold.

viii. Click the "**Compute z-scores**" button ONCE, then be patient and wait.

VIEW LIGAND

ix. Download the full PDB for 1a09 by selecting from the Plugin menu, **Plugin -> PDB Loader Service**, and entering 1a09. Two chains of 1a09 should appear together with any ligands.

x. **View the amino acid sequence** by clicking the "S" button on the bottom right of the PyMOL graphics window.

xi. Find and select the sequence for the ligand (it should be "ACE PTH E DIP"). Display the selected ligand as spheres (S -> spheres)

Find the surface active site

Find the ligand

Identify which residue the ligand contacts

What is the rank/coverage of the contact residues?

Two. Automatic download of a pre-traced structure: PDB *1a09A*

xii. On the ET2 page, under the "**Load structure and precomputed trace**" box, enter 1a09A.

xiii. Click the button **Load trace**. The structure file and the rank data file will be downloaded (from our ET server) and the appropriate boxes on PyETV will be filled automatically. The structure and the trace are ready to be examined (as in the above example).

Three. Automatic download of a pre-traced structure: PDB *1a09A*

xiv. Close the previous PyMOL session (optional).

xv. Visit http://mammoth.bcm.tmc.edu/ETserver.html, enter "1a09", and click "Submit"
xvi. Click on the link "**1a09**" next to "**All results with PyMol:**". This link is a PyMOL script (**1a09.pml** file). Download and open in PyMOL by clicking on the pml file, or from the PyMOL menus, select **File -> Run**, and locate the **1a09.pml** file.

Four. Automatic download of a pre-traced protein-protein complex: PDB *1fin*

i. Close the previous PyMOL session (optional).
ii. Visit http://mammoth.bcm.tmc.edu/ETserver.html, enter "1fin" (CDK2-cyclinA structure), and click "Submit"
iii. Click on the link "**1fin**" next to "**All results with PyMol:**". This link is a PyMOL script (**1fin.pml** file). Download and open in PyMOL by clicking on the pml file, or from the PyMOL menus, select **File -> Run**, and locate the **1fin.pml** file.
iv. Open the PyETV plugin (**Plugin -> PyETV**). ET rankings are automatically shown in prismatic mode, with ET rank threshold set to 100. Pages are allocated for **1finA** and **1finB**, and one named **Zcoupling**.
v. Examine the structure by rotating it, and showing it in cartoon or surface mode. Hide either 1finA or 1finB by clicking on its name on the right panel of the PyMOL graphics window.
vi. Drag the top slider to 30%. ET rank threshold can also be set independently using the sliders within the **1finA** and **1finB** page.
vii. Find the residues that are at the interface of **1finA** and **1finB** by selecting the **Zcoupling** tab next to 1finB. Make sure the right chains are selected in the two drop-down combo boxes. Click "**Mark interface**" only ONCE, then be patient and wait.
viii. Compute the ET coupling z-score between 1finA and 1finB by clicking "**Compute coupling z-score**" just ONCE, then be patient and wait for the numbers to appear.

Are important residues from one face near those of the other?

What is the statistical significance of this proximity?

Five. Viewing a predicted biological complex from PISA: PDB *1ihf*

i. With PyMOL and the PyETV window open, select the "**Assembly**" tab in PyETV.
ii. Enter "1ihf" for the pdb code (IHF is Integration Host Factor/DNA complex).
iii. Click "**Load Assembly**" ONCE, then be patient and wait.

iv. Tabs will be created for each chain, if a precomputed trace matching the sequence was found. Examination and manipulation of the traces is similar to the previous example with 1fin. Verify that top-ranked ET residues cluster near the protein-protein and protein-DNA interfaces.
v. "**View PISA Page**" opens a webpage to the PISA entry for the PDB structure. (Useful for checking if a predicted biological assembly does indeed exist.)
vi.

ADVANTAGES

vii. PyETV leverages the power and flexibility of the PyMOL molecular graphics system, enabling the high-quality display of many different kinds of structures and ligands (protein, DNA, RNA, small molecules, drugs, polar and hydrogen bonds, etc.) in a wide range of styles (surface, cartoon, stereo, transparencies, etc.) and colors. PyMOL also provides many useful functions and plugins that enable users to do structural alignment, electrostatics calculations, distance measurements, amino acid replacements, animations, etc.
viii. Any number of structures can be displayed at the same time (addressing the primary weakness of the Java ET Viewer).
ix. Any number of traces can be mapped to any structure. ***You can open as many PyETV windows as you want in a single PyMOL session.***
x. Calculations involving protein-protein and protein-ligand interfaces are available.
xi. PyETV is easily extendible to provide other functionality, such as providing new measures of clustering of ET residues.

DISADVANTAGES

xii. No integrated tree viewer,
xiii. no MSF viewer,
xiv. no ET Wizard.

6. DIFFERENCE EVOLUTIONARY TRACE SERVER : UET

This web service brings the technology behind the ET Wizard to the web. While the ET Wizard requires a PDB structure or ID as input, UET can also perform an ET analysis starting from a sequence specified by a UniProt ID or the actual amino acid sequence in FASTA format.

Zero. Difference ET in Regulator of G protein signaling: PDB *1rgs_*

a. Internet URL: http://mammoth.bcm.tmc.edu/uet/

REMINDER: Do not let your students run a new trace before you do, otherwise, your request will be placed in the queue.

1. On http://mammoth.bcm.tmc.edu/uet/, select **PDB ID and chain identifier**, then click **Next**
2. On this page (**Step 2: Enter input**) enter a string of five characters consisting of a ***PDB ID and chain identifier***, e.g. "1rgs_". Click **Next**
3. If a precomputed trace is available, a downloadable link appears for the data files in zip format (**1rgs_.zip**). The amino acid sequence with prismatic coloring indicating relative evolutionary importance should also appear. Place the mouse-pointer over an amino acid symbol to view the residue number.
4. If you choose to perform a difference ET analysis starting from the precomputed trace, click "**continue with difference ET on 1rgs_**".

5. If you want to run a new trace, click "**Run a new trace for 1rgs_**".

CUSTOMIZE ET

You may also perform a difference trace after running a new trace. For demo purposes: don't !

6. You may also click on "**Show Advanced Options**" to view the parameters that can be varied for this trace. A custom multiple sequence alignment file (in GCG/MSF format) may be provided near the bottom of the Advanced Options for upload. If you do run a new trace, be patient, and the browser will indicate progress by updating every

few seconds. Do not close the browser before the trace is complete.

7. When the trace is complete, the new trace results will be available for download as a zip file. A prismatic coloring of the amino acid sequence will also appear when the trace is successful (if it is not successful, a link to the trace log file will appear).

DIFFERENCE ET

8. Clicking on the link "**Perform a difference trace**" may start a difference trace analysis. If you click on "**Perform a difference trace**", the following ensue:
    1. Your browser is redirected to a simple web page containing a summary of the multiple sequence alignment. A phylogenetic tree viewer is also launched. Allow the tree viewer (***Archaeopteryx***) to run.
    2. On the tree viewer, examine the tree by using the Zoom features (buttons on the left panel)
    3. On the tree viewer, click "**Click on Node to:**" and select "**Choose Superclass**". Click on a node to select a superfamily. Choose a color when prompted and click OK. A window will appear listing some information about the node (such as the number of sequences under it). Close this window.
    4. Go back to "**Click on Node to:**" and select "**Choose Subclass**". Click on a node to select a subfamily.
    5. Click on "**DiffET Submit**" to start the traces based on the superfamily and subfamily alignments. Be patient and wait. A new webpage will open to indicate the progress of the trace. Close the small message box stating that your request had been submitted. The number of submissions through "**DiffET Submit**" is limited.
    6. The final output page presents a table containing links to the data files. Included in the page is a prismatic coloring of the amino acid sequence indicating relative evolutionary importance based on the superfamily (SUPERCLASS) and subfamily (SUBCLASS) alignments. The evolutionary importance can also be visualized as histograms by clicking "**Visualize ET ranks as histograms**".
    7. Click on the icon (black and green) next to "Difference trace" in the table. This icon is a link to a

PyMOL script (1rgs_trace.pml) that loads the structure (1rgs) and the superfamily and subfamily traces. Download this link and run it in PyMOL. (If your browser can associate the .pml extension with PyMOL, the browser can launch PyMOL directly. Otherwise, open PyMOL, then select in the menu **File -> Run** and locate 1rgs_trace.pml)

8. From the PyMOL menus, select **Plugin -> PyETV**. On the **Diff ET** page of the PyETV window, click the button "**Map ranks to structure**". On the 1rgs structure in the PyMOL structure viewer:
    a. red indicates the subfamily-only trace residues,
    b. blue indicates the superfamily-only trace residues,
    c. orange indicates the ET residues common to both families, at the ET percentile rank threshold specified by the sliders.
9. Adjust the ET rank threshold by moving the slider for the superfamily.
10. The slider for the subfamily can be adjusted independently by un-checking "Synchronize Subfamily ranks slider to Superfamily."

    Is the superfamily trace significant?

    Is the subfamily trace significant?

    Where do they overlap?

    Are some sites specific to the subfamily?

Advantages

1. UET makes the technology behind the ET Wizard available on a web browser. (No need to launch an ET Viewer).
2. Compared with ET Wizard and Report Maker, the user can directly provide an amino acid sequence (in FASTA format) to UET as input for a *de novo* trace analysis.
3. UET provides options to compute other "flavors" of ET (Real-value ET (rvET), Pair-rank ET (prET), Optimization and automatic sequence selection, Position-specific gap reducing rvET. These options are available for queries with PDB structure.)

Disadvantages

1. Similar to the ET Wizard, only a very limited number of trace requests (1 for UET) can be processed by the webserver at any one time. The rest (up to 10) of the requests will be queued and the next one processed when the webserver completes the previous computation.

7. DIFFERENCE ET SERVER FOR GPCR CLASS A RECEPTORS

A. Internet URL: http://mammoth.bcm.tmc.edu/gpcr/diff_GPCRpaper.html
B. Example
    a. Choose two sets of sequences for analysis. Options are Class A, Class A minus Olfactory, Amine, Peptide, Hormone, Rhodopsin, Olfactory, Prostanoid, Glycoprotein, Viral, and Lysosphingolipid GPCRs. Class A vs Amines is a good example.
    b. Hit or click "Submit".
    c. New page returns (this should be only a few seconds) with new options. From left to right the new links are as follows:
        1. Difference ET in pymol session on PDB ID 1f88
            - Once link (green molecule) is clicked, a Pymol script downloads (Trace_1F88.pml).
            - Opening the script (I just clicked on it) brings up Pymol with the GPCR structure 1f88.
            - Open PyETV plugin.
            - Map ET results to structure 1f88 transmembrane with plugin.
            - Switch to spheres to see the difference ET results.
        2. Difference ET in Pymol session on PDB ID 2rh1. Same as previous step with structure 2rh1 instead of 1f88.
        3. Alignments for the two sets of sequences. Download alignment in GCG/MSF format.
        4. ET rank files for each set of sequences. Link for viewing the ET results in ETViewer. ET Viewer opens with relevant structure and the default trace for the relevant set of sequences (can not see the difference ET here, just normal ET).
        5. Zip file containing complete set of ET results (result.zip).
            - Alignments.
            - Structures.
            - Pymol sessions.
            - Pymol scripts.
            - Readme file.
            -

Can you map out ligand binding sites unique to:

- *bioamine receptors?*
- *Chemokine receptors?*
- *Visual receptors?*
- *Olfactory receptors?*

Are they similar or different?

8. ET ANNOTATIONS

ETA suggests the function of protein structures. It starts with a structure of unknown function, such as those from structural genomics, and with no prior knowledge of its mechanism uses the phylogenetic Evolutionary Trace (ET) method to extract key functional residues and propose a function-associated 3D motif, called a ***3D template***. ETA then searches previously annotated structures for geometric template matches that suggest molecular and thus functional mimicry. In order to maximize the predictive value of these matches, ETA next applies distinctive specificity filters -- evolutionary similarity, function plurality and match reciprocity. In large scale controls on enzymes, prediction coverage is moderate 43-60% but the positive predictive value rises to 92%, thus minimizing false annotations. Users may modify any search parameter, including the 3D templates. ETA thus expands the ET suite for protein structure annotation, and may contribute to annotation metaservers.)

a. Internet URL: http://mammoth.bcm.tmc.edu/eta/
b. ETA analysis starts with the PDB code of the protein structure of unknown function, including a 1-digit chain identifier
    - 3h04A
    - 1nvtA
    - 2eerA
    - 1bjwA.
c. Click “Submit”. An ET analysis then provides information on the evolutionary importance of each residue. If this ET analysis is cached, the server goes to step 2. If not, it launches automatically a new trace with default parameters. One may gain control over this process by uploading a custom ET analysis that was run before through the ET Wizard. Clicking “Browse” to locate such an ET file and “Upload” to submit it to the ETA server.
d. Next, the server predicts a functional site template by identifying a cluster of evolutionarily important residues on the surface of the protein, picking the six most important ones. It renders an image of the template. This template can be explored in depth by clicking on the image to download a PyMOL session file. The template may be customized if alternate choices of residues are of interest. Click “Submit Template” to continue with the analysis.
e. The server next identifies possible amino acid types for each template residue based on the multiple sequence alignment used by ET. Each unique combination is listed, along with the number of times it occurs in the alignment. Combinations may be turned on or off using their check boxes. Custom amino acid labels can also be added. Click “Find Matches” to begin the template search.

f. The results page contains GO and EC predictions based on reciprocal matches (highly reliable) and non-reciprocal matches (less reliable). The GO terms and EC numbers are hyperlinked to web pages containing more information about that GO term or EC number.

A note about the 3h04A example ETA matches: The DALI algorithm, which performs whole domain three-dimensional structural alignments, reveals similarities to the same carboxylesterases that ETA matches. The catalytic triad of chain 2c7b is known to be Ser154, Asp251, and His281, and these residues are aligned with a corresponding serine, aspartic acid, and histidine in chain 3h04A, suggesting functional importance for these residues. All three residues of this triad were included in the reciprocal ETA template.

What is the G.O. annotation of your protein?

Does it have an E.C. annotation?

Download the PyMOL session on these templates.

Are they completely on the surface?

Are they sequential in the sequence?

What is their evolutionary importance?

What proteins are being matched?

What is the sequence identity of these proteins?

What is the RMSD of these matches

Download the PyMOL session for these matches

Compare and contrast the template and the matches

Advantages:

- ETA provides a set of residues that are necessary for the function, stability or fold of a protein.
- In test cases, many of these residues have been shown to participate in substrate interaction.
- Experimentally validated

Disadvantages:

- ETA requires a structure to build templates: The Lichtarge lab is currently doing research to extend the ETA application to protein sequences that lack structure information.

9. REFERENCES

GENERAL REVIEWS

1. Lichtarge, O., Sowa, M.E. (2002) Evolutionary predictions of binding surfaces and interactions. ***Curr. Op. Struct. Biol.*** 12 : 21-27.
2. Lichtarge O. and Wilkins A.D. (2010) Evolution: a guide to perturb protein function and networks. ***Curr. Op. Struct. Biol***. 20(3):351-9.
3. Erdin S, Lisewski, A.M., and Lichtarge O. (2011). Protein Function Prediction: Towards Integration of Similarity Metrics. ***Curr. Op. Struct. Biol***. 21(2):180-8.

TOOLS

4. Lua RC and Lichtarge O. (2010). PyETV: a PyMOL evolutionary trace viewer to analyze functional site predictions in protein complexes. **Bioinformatics**. 26(23):2981-2.
5. Ward, R.M., Venner E., Daines B., Murray S., Erdin S., Kristensen D.M. and Lichtarge O. (2009) Evolutionary Trace annotation server: automated enzyme function prediction in protein structures using 3D templates. ***Bioinformatics*** 25:1426-7.
6. Morgan, D.H., Kristensen, D.M., Mittleman, D., Lichtarge, O. (2006). ET Viewer: an application for predicting and visualizing functional sites in protein structures. ***Bioinformatics*** 22:2049-50.
7. Mihalek, I., Lichtarge, O. Evolutionary Trace Report Maker: a new type of service for comparative analysis of proteins (2006). ***Bioinformatics*** 22:1656-7

THEORY

8. Lichtarge, O., Bourne H.R., Cohen F.E. (1996). The Evolutionary Trace method defines the binding surfaces common to a protein family. ***Journal of Molecular Biology*** 257:342-358. **Cover art.**
9. Lichtarge O., Yamamoto, K.R., Cohen F.E., (1997). Identification of functional surfaces of the zinc binding domains of intracellular receptors. ***Journal of Molecular Biology*** 274:325-337.
10. Madabushi, S., Yao, H., Marsh, M., Philippi, A., Kristiansen, D., Sowa, M.E., Lichtarge, O. *****, (2002). Structural clusters of Evolutionary Trace residues are statistically significant and widespread in proteins ***J. Mol. Biol.*** 316:139-153
11. Yao, H., David M. Kristensen, D.M., Mihalek, I., Sowa, M.E., Shaw, C., Kimmel, M., Kavraki, L., Lichtarge, O**.** (2003) An accurate, scalable method to identify functional sites in protein structures. ***J. Mol. Biol.*** 326:255-261.

12. Mihalek, I., Res, I., Yao, H., Lichtarge, O. (2003). Combining inference from evolution and geometric probability in protein structure evaluation. ***J. Mol. Biol*** 331:263-279.
13. Mihalek, I., Res, I., Lichtarge, O. (2004). A family of evolution-entropy hybrid methods to rank the importance of protein residues. ***J. Mol. Biol***. 336:1265-1282.
14. Raviscioni, M., Gu, P., Sattar, M., Cooney A.J., Lichtarge, O. (2005). Correlated evolution at molecular interfaces is common and guides the rational engineering of protein-DNA binding specificity. ***J. Mol. Biol***. 350: 402-15.
15. Mihalek, I., Res, I., Lichtarge, O. (2006) Evolutionary and structural feedback on selection of sequences for comparative analysis of proteins. ***Proteins*** 63:87-99.
16. Yao, H., S., Mihalek, I., Lichtarge, O. (2006) Rank Information: a structure-independent measure of Evolutionary Trace quality that improves identification of protein functional sites. ***Proteins*** 65:111-23.
17. Mihalek, I., Res, I., Lichtarge, O. (2006) A structure and sequence guided monte-carlo sequence selection strategy for multiple alignment-based analysis of proteins. ***Bioinformatics*** 22:149-56.
18. Mihalek, I., Res, I., Lichtarge, O. (2007). Background frequencies for residue variability estimates: BLOSUM revisited. ***BMC Bioinformatics*** 8:488.
19. Wilkins A.D., Lua R., Erdin S., Ward R.M. and Lichtarge O. (2010) Sequence and structure continuity of Evolutionary emportance pmproves protein functional site discovery and annotation. ***Protein Sci.*** 19(7):1296-311.

APPLICATIONS

20. Lichtarge, O., Bourne H.R., Cohen F.E. (1996). Evolutionarily conserved $G_{\alpha\beta\gamma}$ binding surfaces support a model of the G protein-receptor complex. ***Proc. Natl. Acad. Sci. U.S.A.*** 93:7507-7511.
21. Sowa, M.E., Wei He, Wensel, T.G. and Lichtarge, O. (2000) Identification of a general RGS-effector interface. ***Proc. Natl. Acad. Sci. U.S.A.*** 97:1483-1488.
22. Sowa, M.E., Wei He, Slep, K.C., Kercher, M.A., Lichtarge, O*., Wensel, T.G. (2001) Prediction and confirmation of an allosteric pathway for regulation of RGS domain activity. ***Nature Struct. Biol.*** 8:234-237.
23. Quan XJ, Denayer T, Yan J, Jafar-Nejad H, Philippi A, Lichtarge O, Vleminckx K, Hassan BA. (2004) Evolution of neural precursor selection: functional divergence of proneural proteins. ***Development*** 131:1679-89.
24. Madabushi, S., Gross, A., Philippi, A., Meng, E.C., Wensel, T.G., Lichtarge, O. (2004). Signaling determinants reveal functional subdomains in the transmembrane region of G protein-coupled receptors. ***J. Biol. Chem***. 279:8126-8132.
25. Shenoy S.K., Drake M.T., Nelson C.D., Houtz D.A., Xiao K., Madabushi S., Reiter E., Richard T. Premont, Lichtarge O. and Lefkowitz R.J. (2006) β-Arrestin-

dependent, G protein-independent ERK1/2 activation by the β2-adrenergic receptor. ***J. Biol. Chem***. 281:1261-73.

26. Raviscioni, M., Qiang He, Salicru, E.M., Smith C.L. and Lichtarge, O. (2006). Evolutionary identification of a subtype specific functional site in the ligand binding domain of steroid receptors. ***Proteins*** 64:1046-57
27. Kristensen D. M., Ward R. M., Lisewski A. M., Erdin S., Chen B. Y., Fofanov V. Y., Kimmel M., Kavraki L. E., and Lichtarge O. (2008) Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. ***BMC Bioinformatics*** 9:17.
28. Kobayashi, H., Ogawa K., Yao R., Lichtarge O. and Bouvier M. (2009) Functional rescue of beta1-adrenoceptor dimerization and trafficking by pharmacological chaperones. ***Traffic*** 10(8):1019-33.
29. Rodriguez GJ, Yao R., Lichtarge O.* and Wensel T*. (2010). Evolution-guided discovery and recoding of allosteric pathway specificity determinants in psychoactive bioamine receptors. ***Proc. Natl. Acad. Sci. USA*****.** 107(17):7787-92
30. Häberle J, Shchelochkov OA, Wang J, Katsonis P, Hall L, Reiss S, Eeds A, Willis A, Yadav M, Summar S, Lichtarge O, Rubio V, Wong LJ, Summar M; and the Urea Cycle Disorders Consortium. (2010). Molecular Defects in Human Carbamoyl Phosphate Synthetase I: Mutational Spectrum, Diagnostic and Protein Structure Considerations **Hum. Mutat.**
31. Venner E, Lisewski AM, Erdin S, Ward RW, Amin S and Lichtarge O. (2010). Accurate protein structure annotation through competitive diffusion of enzymatic functions over a network of local evolutionary similarities. ***PLoS One*** 5(12):e14286.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge support from the National Institute of Health, NIH GM079656 and GM066099, and from the National Science Foundation, NSF, CCF 0905536.

Pymol is a user-sponsored molecular visualization system on an open-source foundation being developed and supported by Schrodinger LLC. PyMOL was originally developed by Warren L. DeLano. For more information about Pymol see http://www.pymol.org/. Visit the **PyMOLWiki** for tutorials, scripts, answers to frequently asked questions, and more. A user-maintained knowledge base, the PyMOLWiki is full of helpful information.