Exploring Biomolecular ***Sequence*** ↔ ***Structure*** ↔ ***Function*** Relationships, I: Focusing on Sequence and Taking a Putative Aminotransferase as a Case-study

GOAL: Learn how to explore, discover, identify, analyze and synthesize information from multiple bioinformatic resources in order to address a single question/problem (e.g., enzyme substrate specificity).

OVERVIEW: Biochemistry, alongside most other areas of the biosciences, is increasing becoming a data science. Such is the case because high-throughput DNA sequencing, microarrays, and numerous other massively parallel, 'omics'-scale technologies have been developed—and increasingly matured into the mainstream—since the late 1980s. In particular, there has been a phenomenal increase in the sheer *volume* and *types* of data. For our immediate purposes in this Lab, chief among these types of data are (i) biomolecular ***sequence data***, (ii) three-dimensional (3D) ***structural data***, and (iii) ***functional data*** (e.g., enzymatic activities, ligand-binding assays, and other biochemical activities). When analyzed together, in as integrated a manner as possible, these three types of data position us to glean unprecedented insight into the subtle, complicated relationships that link protein *sequence* ↔ *structure* ↔ *function* (acted upon by over three billion years of evolution). These relationships, in turn, literally *define* much of an organism's molecular physiology. In order to decipher these relationships—and really convert what would otherwise be mundane, sterile data-sets into biochemical insight and new knowledge—we require computational approaches.

That computational methods are required is clear by considering the need for automation, driven by the volume of available data: it is inconceivable that even an army of human beings could manually compare, annotate, curate, etc. the ≈20,000 or more protein-coding genes in humans (and this is to mention nothing of the entire protein universe, considered across all [known] species!). Therefore, biomolecular (DNA, RNA, protein) sequence databases sprung into existence from the dawn of genomics (late-1970s, early-1980s). Similar tales can be told for structural data (the Protein Data Bank [PDB] began in the 1970s, after only the first few protein structures), protein interaction data and other functional data, and so on. In short, computational methods (algorithms) and technical approaches (e.g., workflow management systems) are indispensable in every stage of biomolecular sciences—collecting, organizing, sorting, searching/querying, comparing, and otherwise analyzing biomolecular sequences, structures and functions. Such analyses enable one to generate models to precisely describe structure ↔ function relationships, and ideally these models correspond to *experimentally testable* hypotheses (i.e., in the wet-lab).

As you may suspect by now, the scientific efforts described above largely occur at the interfaces of the ***biosciences*** (biochemistry, cell biology, etc.) and the ***computational sciences*** (computer science algorithms; databases [DB], and other technological approaches). As is generally true for most scientific investigations, an overriding goal in the work that we will pursue in this Lab is to generate a hypothesis/model based on all of the information that is available. In bioinformatics and computational biology, the concept of 'available information' is not as straightforward as it may seem, at least at first. This is because potentially relevant data occur in many different types of databases, in the literature, and so on. This Lab Module aims to introduce you to some of the resources that are available and to encourage you to uncover other resources—on your own. The last part ('*on your own*') is key for several reasons; one of these reasons is not that we seek to make this a difficult Lab, but rather because bioinformatic databases are notorious for being somewhat volatile (web addresses change, websites are no longer maintained or become otherwise outdated, etc.). The scientific independence and resourcefulness that are developed in pursuing bioinformatic studies serve one well in many other areas of life too (beyond purely scientific).

For the purposes of this Lab Module, we will focus on functional characterization—*via bioinformatic analyses*—of a putative aminotransferase. The 3D structure of this protein of interest (POI), corresponding to PDB ID 3FDB (DOI 10.2210/pdb3fdb/pdb), is entitled "Crystal structure of putative PLP-dependent beta-cystathionase (NP_940074.1) from *Corynebacterium diphtheriae* at 1.99 Å resolution".

We will proceed along nine stages. Note that this nine-step workflow is adapted from the general approach described by Mazumder & Vasudevan in "*Structure-Guided Comparative Analysis of Proteins: Principles, Tools, and Applications for Predicting Function*" (*PLOS Computational Biology*, 2008; DOI 10.1371/journal.pcbi.1000151). Briefly, the nine major steps follow:

Step 1. Gather sequence and structure files, and learn about them (what they contain, file formats, conversion methods, etc.)

Step 2. PSI-BLAST against the NCBI non-redundant (nr) database to identify potential homologs

Step 3. Evaluate pairwise alignments against the identified structures from Stage 1

Step 4. Scan against sequences pattern, domain, and family classification DBs

Step 5. Search against structural family DBs for purposes of structural classification

Step 6. Identify potential functionally-relevant residues (NB: This is a very crude and impartial treatment here—just a first pass to what type of work you'll have to do overall [this semester and next].)

Step 7. Search for structural neighbors and perform structure-guided alignments

Step 8. Phylogenetic trees and analyses thereof

Step 9. Evidence-based assignment of biological function for your POI

General notes:

- Ctrl-mouse-clicking (ctrl-left, ctrl-right, right-click, etc.) is your friend in this type of bioinformatic work! (…will open links in new browser tab/window, spawn a context-specific menu, etc.)

Step 1: Gather sequence and structure files and learn about them

1.1 What organism was the POI cloned from?

Corynebacterium diphtheriae

Beyond the structure and the protein sequence, list three other pieces of information you can obtain about your protein from the PDB site?

- The protein is a homo-dimer
- POI is annotated as a "PLP-dependent **beta-cystathionase**"
- The protein is a **mixed a-helix/b-sheet** secondary structure
- POI was xtallized with numerous bound small molecules (chloride, glycerol, etc.)
- SeMet phasing to determine structure, based on "**MSE**" residue
- Looks like **PLP was covalently bound** to the (Schiff base) Lysine side-chain in the structure

1.2 How was the protein expressed?

Using the **"SpeedET" plasmid** in the **host *E. coli***, **strain HK100**…

1.3 Where's the protein **sequence** in the PDB entry? Are there AAs that you don't recognize or changes in the sequence that you didn't expect? If so, be sure to clarify what you don't understand. There are tabs at the top; click on the sequence tab. Read the page. What does is meant by "The structure XYZ has in total # chains. Out of these "1 are sequence-unique."? This will be important later and each of your POIs will be different.
Click on 'sequence' tab (3rd from left)…
Yes… for example, residue MSE (SeMet).

1.4 Based on your understanding of the above, choose the chain which is most appropriate to and click on FASTA sequence in the "Download files" pull-down and save the file—You will be copying and pasting this sequence frequently. What is FASTA format?
OK, this should be fairly straightforward (see Lab Guide)

1.5 Open the file with any program that handles text files (text editor). What do you notice about the N-terminal portion of this sequence, in terms of amino acid composition? (Always remember that this is the AA sequence of the protein that was crystallized; however, it is not the natural sequence.)
OK (see Lab Guide)

1.6 What can we learn from the REMARK fields (in the PDB 'header' or 'preamble') about how the protein was expressed?
We learn this: """"THE CONSTRUCT WAS EXPRESSED WITH A PURIFICATION TAG MGSD-KIHHHHHHENLYFQG. THE TAG WAS REMOVED WITH TEV PROTEASE LEAVING ONLY A GLYCINE (0) FOLLOWED BY THE TARGET SEQUENCE""""

1.7 "To find the natural protein sequence, go to the Entrez cross database and search for your POI using both the PDB id and the GenBank accession number/NCBI Reference sequence (found on the JCSG PSCA target page or in the PDB title………" (See the full Lab Guide for rest of this step/question.)

From http://www.ncbi.nlm.nih.gov/protein/NP_940074.1, we find **NP_940074.1** / **GI:38234307**. This 'GI' business is the GenBank accession #, and the NP_[0-9]* is the NCBI RefSeq accession #. Now, this is interesting: Visit the protein search page (http://www.ncbi.nlm.nih.gov/protein), enter as a query '3fdb', which returns http://www.ncbi.nlm.nih.gov/protein/?term=3fdb, and there we see two results (or 'hits'):

http://www.ncbi.nlm.nih.gov/protein/217035513?report=fasta—The PDB one. It contains a link to the structure on the right-hand side (RHS), explicitly states the PDB code in the FASTA header.

http://www.ncbi.nlm.nih.gov/protein/38200570?report=fasta—This is the 'natural' one… because many of the various bioinformatic databases are 'smart' and in continuous (automated) information exchange/updating with one another, this entry also contains a link to the PDB structure.

– Now check-out the N-termini of these two, see how they differ (a "cloning artefact"…)?
- Both searches, PDB ID and GenBank accession, yielded a 'hit' to the 3D structure (the PDB entry).
- If search with either 'NP_940074.1' or 'GI:38234307', then we get just 1 unique hit.
- If search with '3fdb', then we get the two hits (URLs above).

Here is the naturally-occurring (*wild-type*) sequence, 376 residues (**note this convention**: a `fixed-width` font for biopolymer sequences, which gets everything nicely in-register and facilitates comparing strings of characters):

```
  1 mqfpsiedlr arntmkwtry gqgvlplwva esdfstcpav lqaitdavqr eafgyqpdgs
 61 llsqataefy adrygyqarp ewifpipdvv rglyiaidhf tpaqskvivp tpayppffhl
121 lsatqregif idatgginlh dvekgfqaga rsillcnpyn plgmvfapew lnelcdlahr
181 ydarvlvdei haplvfdgqh tvaagvsdta asvcititap skawniaglk caqiifsnps
241 daehwqqlsp vikdgastlg liaaeaayry gtdflnqeva ylknnhdfll heipkripga
301 kitpmqatyl mwidfrdtti egspseffie kakvamndga wfgedgtgfc rlnfatsrev
361 leeaidrmak avshht
```

The “structure sequence”, 377 residues (AA seq from PDB file, i.e. from the expression construct):

```
  1 gxqfpsiedl rarntxkwtr ygqgvlplwv aesdfstcpa vlqaitdavq reafgyqpdg
 61 sllsqataef yadrygyqar pewifpipdv vrglyiaidh ftpaqskviv ptpayppffh
121 llsatqregi fidatgginl hdvekgfqag arsillcnpy nplgxvfape wlnelcdlah
181 rydarvlvde ihaplvfdgq htvaagvsdt aasvcitita psxawniagl kcaqiifsnp
241 sdaehwqqls pvikdgastl gliaaeaayr ygtdflnqev aylknnhdfl lheipkripg
301 akitpxqaty lxwidfrdtt iegspseffi ekakvaxndg awfgedgtgf crlnfatsre
361 vleeaidrxa kavshht
```

1.8 Go to BLAST and scroll down to the bottom of the page and click on “Align two (or more) sequences using BLAST (bl2seq)”. Be sure to select the balstp tab and not blastn (the default). What is the difference between blastp and blastn?

- First, perform these steps…
- blast**p** is for peptide (protein) sequence alignment, blast**n** for nucleotide (DNA/RNA) alignment.

1.9 Copy and paste the natural and PDB sequences into the Query and Subject Sequence (respectively) and click-on “Blast” at the bottom of the page. What is different between the protein that was used for determining the structure and the one occurring naturally in the organism? Is the starting amino acid the same? Are there any mutations?

Hint 1: Before clicking the “*Align two or more sequences*” box, it is **very useful** to click on the “**?**” symbol to learn a little bit about what you’re doing—this provides a quick help guide. (We emphasize this because it’s generally true of *all* of these bioinformatic databases/‘web-service’ types of resources—they often have many built-in help features that folks often don’t know to explore…)

Hint 2: *Wanna make life easier?* If you copied the sequences into this MS Word file as plaintext (as above), instead of saving the two sequences as separate FASTA text files on your computer, then an easy way to select & paste the sequences from above, without copying the numbers too (which you would then have to manually remove from the BLAST input field), is to use the trick of the ALT key while highlighting the region you wish to copy. The ALT key selects a ‘vertical’ region of text. (Analogously, if you’re in a UNIX/Linux shell and using the vi/vim editor, `ctrl-v` will get you into ‘column mode’.)

So, after all that rigmarole, finally here (next page) is the result of blastp for these two sequences. Notice the '`X`'s (…more on that below).

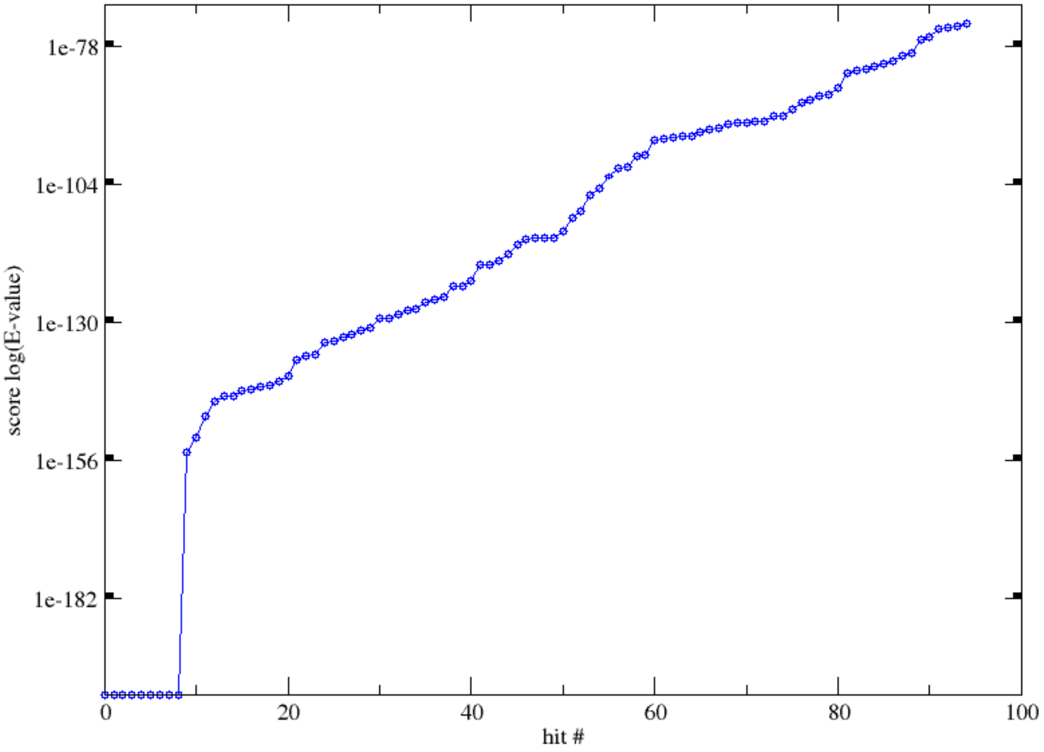
```
Score       Expect  Method               Identities  Positives   Gaps
762 bits(1967) 0.0     Compositional matrix adjust.  368/375(98%)   368/375(98%)   0/375(0%)

Query  2    QFPSIEDLRARNTMKWTRYGQGVLPLWVAESDFSTCPAVLQAITDAVQREAFGYQPDGSL  61
            QFPSIEDLRARNT KWTRYGQGVLPLWVAESDFSTCPAVLQAITDAVQREAFGYQPDGSL
Sbjct  3    QFPSIEDLRARNTXKWTRYGQGVLPLWVAESDFSTCPAVLQAITDAVQREAFGYQPDGSL  62

Query  62   LSQATAEFYADRYGYQARPEWIFPIPDVVRGLYIAIDHFTPAQSKVIVPTPAYPPFFHLL  121
            LSQATAEFYADRYGYQARPEWIFPIPDVVRGLYIAIDHFTPAQSKVIVPTPAYPPFFHLL
Sbjct  63   LSQATAEFYADRYGYQARPEWIFPIPDVVRGLYIAIDHFTPAQSKVIVPTPAYPPFFHLL  122

Query  122  SATQREGIFIDATGGINLHDVEKGFQAGARSILLCNPYNPLGMVFAPEWLNELCDLAHRY  181
            SATQREGIFIDATGGINLHDVEKGFQAGARSILLCNPYNPLG VFAPEWLNELCDLAHRY
Sbjct  123  SATQREGIFIDATGGINLHDVEKGFQAGARSILLCNPYNPLGXVFAPEWLNELCDLAHRY  182

Query  182  DARVLVDEIHAPLVFDGQHTVAAGVSDTAASVCITITAPSKAWNIAGLKCAQIIFSNPSD  241
            DARVLVDEIHAPLVFDGQHTVAAGVSDTAASVCITITAPS AWNIAGLKCAQIIFSNPSD
Sbjct  183  DARVLVDEIHAPLVFDGQHTVAAGVSDTAASVCITITAPSXAWNIAGLKCAQIIFSNPSD  242

Query  242  AEHWQQLSPVIKDGASTLGLIAAEAAYRYGTDFLNQEVAYLKNNHDFLLHEIPKRIPGAK  301
            AEHWQQLSPVIKDGASTLGLIAAEAAYRYGTDFLNQEVAYLKNNHDFLLHEIPKRIPGAK
Sbjct  243  AEHWQQLSPVIKDGASTLGLIAAEAAYRYGTDFLNQEVAYLKNNHDFLLHEIPKRIPGAK  302

Query  302  ITPMQATYLMWIDFRDTTIEGSPSEFFIEKAKVAMNDGAWFGEDGTGFCRLNFATSREVL  361
            ITP QATYL WIDFRDTTIEGSPSEFFIEKAKVA NDGAWFGEDGTGFCRLNFATSREVL
Sbjct  303  ITPXQATYLXWIDFRDTTIEGSPSEFFIEKAKVAXNDGAWFGEDGTGFCRLNFATSREVL  362

Query  362  EEAIDRMAKAVSHHT  376
            EEAIDR AKAVSHHT
Sbjct  363  EEAIDRXAKAVSHHT  377
```

No, there are not any real amino acid (AA) 'mutations' *per se*, only those arising from two sources: (i) a couple AA differences at the *N*-terminus, due to the cloning construct, and also (ii) the Methionine residues are replaces by MSE (selenomethionine; SeMet) residues (the 'X' above).

Step 2: Psi-BLAST against NCBI's non-redundant (nr) database

2.1: OK, done, visit http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins and paste in the sequence. Or, alternatively (there are often multiple ways to achieve any one task in bioinformatics and with DBs!), another approach is to click on the link "Analyze this Sequence" → "Run BLAST" (menu on RHS at http://www.ncbi.nlm.nih.gov/protein/CAE50265.1 [see S1.7 above for origin of this URL]); clicking on that link takes you to the BLAST page with the accession number already filled-in.

2.2: "*What can you gain from the BLAST results? How do you use the information to investigate your protein of interest?*"—The answer to this is ***a lot***. For example, in the *Graphic Summary* section we see *Conserved Domains*, this POI belongs to the "`AAT_I superfamily`" (What does *that word*—'*superfamily*'—mean? An admirable family…? This is worth digging into on your own.) Look at the layout of the major sections of the results page—'*Graphic Summary*', '*Descriptions*', and '*Alignments*' are the three major sections. Collapse them, open each one at a time, explore their contents, see what they contain. In particular, note the Query Coverage, *E*-values and 'Ident' scores in the Descriptions section. For fun, the following is kinda neat (& completely optional!)—let's look at the distribution of *E*-values in rank-ordered hits, given by the blue trace in the graph below. Do you notice any interesting feature(s) ?



2.3: "*There are many links to other databases and a plethora of information. You may need to go to links from this page to answer these questions. Look for as much information as you can. What do you find?*" Again, the answer here is ***lots*** (see response to S2.2 above).

2.4: "*Does your protein have a conserved domain? What is the conserved domains name? Why would we look to conserved residues to try to understand the function of a protein?*"—Yes, see the above answer to S2.2. Looks like the conserved domain is termed `AAT_I superfamily`, `AAT_like` domain. There are interesting things in this view. For example, we can mouse over a region flagged as the "homodimer interface" in the output... Why conserved residues to understand the function? Because AA residues are what literally *define* the functionality of a protein, be it enzymatic (substrate specificity?), ligand-binding (involved in signaling?), structural scaffolding, or whatever.

2.5: "*What residue has the conserved domain database listed as a catalytic residue?...*"—Follow the instructions in the Lab Guide to get to the Conserved Domains page with the following alignment …from which we see that it's **LYS 222** (=234–12) as the putative catalytic residue (will also try PDBsum, in Step 6):

```
Feature 1                                                                  #
1    151 .[3].DIRSY.[24].FSIFVLH.[2].AHNPTGTD.[17].LFPFFDSAYQ.[23].LFCAQSFSKN.[4].NERVGNL 270
query125 .[2].EGIFI.[18].ARSILLC.[1].PYNPLGMV.[17].ARVLVDEIHA.[21].CITITAPSKA.[4].GLKCAQI 234
1IJ  121 .[2].ECRTV.[19].VKVVYVC.[1].PNNPTGQL.[16].AIVVADEAYI.[17].LAILRTLSKA.[4].GLRCGFT 226
1A   134 .[2].KIIEC.[24].NKALLFC.[1].PHNPVGRV.[17].LMLWSDEIHF.[22].TITFTAPSKT.[4].GMGMSNI 250
gi 1 135 .[2].TLVKN.[23].TKMLILC.[1].PHNPVGRV.[17].LLLVSDEIHA.[22].TFTCLAPSKT.[4].GLQTSYC 250
gi 1 125 .[2].TILKN.[24].VKVMVLC.[1].PHNPTGRV.[17].VIVISDEIHS.[22].SIVCISPSKT.[4].GLQVSSL 241
gi 5 134 .[2].EVVAS.[23].VKLFILC.[1].PHNPVGRV.[17].VIVIADEIHC.[22].SITCMAPSKT.[4].GTQTSAI 249
gi 6 132 .[2].EVVYS.[22].CKLLILS.[1].PHNPGGRV.[17].TLVISDEIHA.[22].SLVFMSPSKA.[4].GLASSYA 246
gi 6 129 .[2].AVTEC.[25].NKILIFC.[1].PHNPVGRC.[17].TILISDEIHA.[21].IIVFLSGGKL.[4].GIFSSYV 245
gi 7 145 .[2].VPVAN.[23].TKLLLLC.[1].PHNPVGRV.[17].IVVVSDEVYC.[22].SVSLLSASKM.[4].GLKHSQV 260
```

2.6: "*Are there cofactor binding sites mapped or other conserved features/sites? If so, what is the cofactor and what residues (amino acid and number) are conserved*"—On the same page as above (S2.5), see that there's a "pyridoxal" tab immediately to the right of the "catalytic" tab. Again, remember that these are the bioinformatically-annotated features of the POI, based on similarity to already characterized proteins in the DBs. Also, we can see from the previous page (the BLAST results page) that there is a "pyridoxal 5'-phosphate binding site" annotated in the graphical summary of the Conserved Domains output. This, too, implies cofactor binding.

2.7: "*What is the oligomeric state of the protein? Determine this two ways. At the top of the BLAST results page, you'll see a schematic. Oligomerization interface residues may be noted here. If not a monomer, what is the oligomeric state and what residues are conserved at the oligomeric interface?* "
It looks like a **dimer**. NB: The optimal way to do this is via the PDB's info (see "generated by PISA" in the PDB entry). Going back to the BLAST-based output, here we see that the oligomerization interface (protein/protein interface) is tagged as "Feature 3", and here's a small part of that (analogous to above "Feature 1"):

```
Feature 3                                                #                             ###
hit1     151 .[3].DIRSY.[24].FSIFVLH.[2].AHNPTGTD.[17].LFPFFDSAYQ.[23].LFCAQSFSKN.[4].NERVGNL 270
query    125 .[2].EGIFI.[18].ARSILLC.[1].PYNPLGMV.[17].ARVLVDEIHA.[21].CITITAPSKA.[4].GLKCAQI 234
hit2     121 .[2].ECRTV.[19].VKVVYVC.[1].PNNPTGQL.[16].AIVVADEAYI.[17].LAILRTLSKA.[4].GLRCGFT 226
hit3     134 .[2].KIIEC.[24].NKALLFC.[1].PHNPVGRV.[17].LMLWSDEIHF.[22].TITFTAPSKT.[4].GMGMSNI 250
hit4     135 .[2].TLVKN.[23].TKMLILC.[1].PHNPVGRV.[17].LLLVSDEIHA.[22].TFTCLAPSKT.[4].GLQTSYC 250
etc. etc. etc. (deleted for sake of clarity here)....
```

The following is from the header of the PDB file (interesting parts are highlighted/extracted):

```
...........
REMARK 300 BIOMOLECULE: 1
REMARK 300 SEE REMARK 350 FOR THE AUTHOR PROVIDED AND/OR PROGRAM
REMARK 300 GENERATED ASSEMBLY INFORMATION FOR THE STRUCTURE IN
...........
...........
REMARK 350 COORDINATES FOR A COMPLETE MULTIMER REPRESENTING THE KNOWN
REMARK 350 BIOLOGICALLY SIGNIFICANT OLIGOMERIZATION STATE OF THE
REMARK 350 MOLECULE CAN BE GENERATED BY APPLYING BIOMT TRANSFORMATIONS
REMARK 350 GIVEN BELOW.  BOTH NON-CRYSTALLOGRAPHIC AND
...........
REMARK 350 BIOMOLECULE: 1
REMARK 350 AUTHOR DETERMINED BIOLOGICAL UNIT: DIMERIC
REMARK 350 SOFTWARE DETERMINED QUATERNARY STRUCTURE: DIMERIC
REMARK 350 SOFTWARE USED: PISA
REMARK 350 TOTAL BURIED SURFACE AREA: 9020 ANGSTROM**2
REMARK 350 SURFACE AREA OF THE COMPLEX: 27080 ANGSTROM**2
REMARK 350 CHANGE IN SOLVENT FREE ENERGY: -97.0 KCAL/MOL
```

(Sidenote: Even though these energetic values are purely computational predictions and not yet experimentally validated, that's quite interesting! Assuming the value is at least semi-accurate, can you use what you know from general chemistry to consider what this implies for the monomer ⇆ dimer equilibrium?)

2.8: As requested, return to the BLAST results for "emb|CAE50265.1| (376 letters)". Then, consider the following question:

"*What proteins are similar in sequence and have structures determined? List the accession number, the name, the organism, and the E-score of the top hit.*"

The top hit (highest sequence similarity) of known 3D structure is the POI's PDB code—i.e., 3fdb... That's a 'trivial' hit. After that, the next highest (*bona fide*) hit is **4dgt_a** (which means "*Chain A of PDB entry 4DGT*"). Note: As of oct2013, to see the 'S' codes on the RHS column (the 'Links' column), it may be necessary to revert to the old style formatting of the BLAST output; do this by going to top of page, click on "Formatting options" pull-down menu, then toggle on "Old view", and finally click the blue "Reformat" button to the right. Then, the page of results will be reformatted so that you should see the 'S' for structure link in the table of hits.

(Note: as mentioned in the lab manual, another trick you can try is to re-BLAST the sequence, but this time limit the search database to only known structures [i.e., the PDB], rather than using 'nr'... Go ahead and try this. You should retrieve 3fdb and then 4dgt as the top two hits!—and, notice now that (reassuringly!) *all* of the hits have an 'S' in the RHS column.)

So, back on track: Here's our conclusion: The protein of known structure with greatest sequence similarity to 3fdb is 4dgt, and here's the result for this one:

- Accession number: 4dgt (PDB ID); GI:374978008 (GenPept ID); find it at http://www.ncbi.nlm.nih.gov/protein/374978008 (interesting trick: click on the "Run BLAST" under "Analyze this sequence" on the RHS... Do you recover 3fdb as top hit? This is known as a 'reciprocal' search analysis.)
- Protein name: It's a putative AT, listed as "*Plp-Bound Putative Aminotransferase From Clostridium Difficile 630 Crystallized With Magnesium Formate*"
- Species of origin: *Clostridium difficile* 630
- *E*-score for 3fdb ↔ 4dgt: 1e-59 (query coverage = 98%, ident = 28%)

…and here’s the (pairwise) alignment:

Chain A, Crystal Structure Of Plp-Bound Putative Aminotransferase From Clostridium Difficile 630 Crystallized With Magnesium Formate
pdb|4DGT|A3911
See 3 more title(s)
Chain B, Crystal Structure Of Plp-Bound Putative Aminotransferase From Clostridium Difficile 630 Crystallized With Magnesium Formate
pdb|4DGT|B
Chain A, Crystal Structure Of Plp-Bound Putative Aminotransferase From Clostridium Difficile 630
pdb|4DQ6|A
Chain B, Crystal Structure Of Plp-Bound Putative Aminotransferase From Clostridium Difficile 630
pdb|4DQ6|B
Structure-3D structure displays
Identical Proteins-Proteins identical to the subject

GenPeptGraphics Next Match Previous Match

Alignment statistics for match #1

Score	Expect	Method	Identities	Positives	Gaps
201 bits(510)	1e-59	Compositional matrix adjust.	107/386(28%)	200/386(51%)	15/386(3%)

```
Query  2    QFPSIEDLRARNTMKWT----RYGQG-VLPLWVAESDFSTCPAVLQAITDAVQREAFGYQ  56
             F  I D     + KW+    +YG   +LP+WVA+ DF   P ++ ++ + +++E +GY
Sbjct  7    NFNEIVDRSNNFSSKWSEMEKKYGTNDLLPMWVADMDFKAAPCIIDSLKNRLEQEIYGYT  66

Query  57   PDGSLLSQATAEFYADRYGYQARPEWIFPIPDVVRGLYIAIDHFTPAQSKVIVPTPAYPP  116
                  +++   +   R+ ++ + EW+   P V+  + + I+  T A  K+++  P Y P
Sbjct  67   TRPDSYNESIVNWLYRRHNWKIKSEWLIYSPGVIPAISLLINELTKANDKIMIQEPVYSP  126

Query  117  FFHLLSATQREGIFI------DATGGINLHDVEKGFQAGARSILLCNPYNPLGMVFAPEW  170
            F  ++    RE I        +    ++  D+E   +   +  +LCNP+NP+G V+  +
Sbjct  127  FNSVVKNNNRELIISPLQKLENGNYIMDYEDIENKIK-DVKLFILCNPHNPVGRVWTKDE  185

Query  171  LNELCDLAHRYDARVLVDEIHAPLVFDG-QHTVAAGVSDTAASVCITITAPSKAWNIAGL  229
            L +L D+  +++ +++ DEIH+ ++    +H   A +S       IT  AP+K +NIAGL
Sbjct  186  LKKLGDICLKHNVKIISDEIHSDIILKKHKHIPMASISKEFEKNTITCMAPTKTFNIAGL  245

Query  230  KCAQIIFSNPSDAEHWQQ-LSPVIKDGASTLGLIAAEAAYRYGTDFLNQEVAYLKNNHDF  288
            + + ++  +  D +      + +     +   L+A EA+Y  G  +L   + YL++N DF
Sbjct  246  QSSYVVLPDEKDYKLLDDAFTRIDIKRNNCFSLVATEASYNNGESWLESFLEYLESNIDF  305

Query  289  LLHEIPKRIPGAKITPMQATYLMWIDFRDTTIEGSPSE-FFIEKAKVAMNDGAWFGEDGT  347
             +  I + +P  K+   + TYL+W+DF    +     E   ++K KVA+N G  FG  G+
Sbjct  306  AIKYINENMPKLKVRKPEGTYLLWVDFSALGLSDEELESILVQKGKVALNQGNSFGIGGS  365

Query  348  GFCRLNFATSREVLEEAIDRMAKAVS  373
            G+ R+N A  R +LEEA+ R+  A++
Sbjct  366  GYQRINLACPRSMLEEALIRIKNAIN  391
```

Notice the highlighted residue above (K222 in the *query*—the putative catalytic residue based on analysis from above).

Step 3: Evaluate pairwise alignment with the identified structures from Stage 1

3.1: "*Still in BLAST, scroll down or (click on the max score hyperlink to jump next to the hit with a structure) to display the sequence alignment with the sequence identified in 2.8 (the best E-score and a known structure).*"

Done—See step 2.8 above.

3.2: "*Copy and paste the alignment in a document for later. Study the alignment. Find the amino acid that were identified as a catalytic residue(s) (you need to compare the sequences; the numbering will not be the same between the two proteins). Does the "hit" sequence have this residue?*
What about the other conserved features/sites that you listed in 2.5 and 2.6? "

Done—See step 2.8 above. Yes, the hit sequence has this residue (K222 in the query [the POI]).

3.3: *"Are there other alignments that are to proteins with known structures? If so, copy and paste these alignments into the saved document as well. Record the PDB IDs as well."*

Yes, there are some, especially if we search against the PDB-only sequence database. Here they are (PDB ID / E-value; see the local file with which we are providing you "`4T11U2T6016-Alignment2.txt`"):

```
4DGT / 1e-59
1C7N / 7e-50
3T32 / 3e-49
3B1C / 2e-46
1D2F / 1e-34
```

3.4: Go to ProFunc and compare results (w/ above results from BLAST).

Nice, now look at what's stated at http://www.ebi.ac.uk/thornton-srv/databases/profunc/index.html:

> """"*The aim of the ProFunc server is to help identify the likely biochemical function of a protein from its three-dimensional structure. It uses a series of methods, including fold matching, residue conservation, surface cleft analysis, and functional 3D templates, to identify both the protein's likely active site and possible homologues in the PDB.*"""

That's just what we want to do!

In the ProFunc output page, under the section "Matches to existing PDB structures", we see the following:

```
Sequence search vs existing PDB entries. Chain A
94 matching sequences found by FASTA search
  PDB code      E-value    %-tage id Overlap  Name
1. 4dgt(A)      2.3e-38    28.608    388  Crystal structure of plp-bound putative aminotransferase fro clostridium difficile 630
crystallized with magnesium forma
2. 4dq6(A)      2.3e-38    28.608    388  Crystal structure of plp-bound putative aminotransferase fro clostridium difficile 630
3. 3t32(A)      3.2e-31    29.248    359  Crystal structure of a putativE C-s lyase from bacillus anth
4. 3kax(A)      3.2e-31    29.248    359  Crystal structure of a putativE C-s lyase from bacillus anth
5. 3l8a(A)      7.4e-28    28.496    379  Crystal structure of metc from streptococcus mutans
```

Why the difference between these ProFunc results and the above BLAST searches?

Well, potentially (i) usage of BLAST for local alignment, versus FASTA (in this ProFunc output); (ii) the relative sizes of these sequence DBs at least partially accounts for the differences in *E*-values between the two sets of top-5 results.

Step 4: Scan against sequences pattern, domain, and family classification DBs

4.1 "*What is the corresponding UniProt accession for the protein (6 letter/number combination [e.g., Q9X0Y2])? (This number is sometimes also listed on the PDB page so if you have trouble you can try checking there.)* "

Q6NFZ9 (“primary [citable] accession number”)
Q6NFZ9_CORDI (“entry name”)

4.2 *“Follow the link of the UniProt accession. What info is contained on this site?”*

We land at http://www.uniprot.org/uniprot/Q6NFZ9, which contains much info:

- Taxonomic lineage and identifiers, classification, gene ontology (GO) terms, evidence for “protein existence” (“evidence at protein level”)
- Lots of info in terms of ontologies, molecular functions, and bound molecules, such as chloride, even have link-outs (small molecules link-out to PDB, sequences to EMBL), etc. etc. etc. There’s also a nice (simple) **Sequence** section, a (literature) **References** section, and a section of **many** cross-references to other DBs and online bioinformatic resources (phylogenomic DBs, organism-specific DBs, family and protein domain DBs, EvoTrace even!)— Indeed, this hub of resources may be one of the most useful features of the UniProt results page.

4.3: "*Along the top of the page click the header “Sequences” to jump down the page. Under the “tools” column there is a dropdown menu. Select ProtParam and click go. What information does ProtParam provide?* "

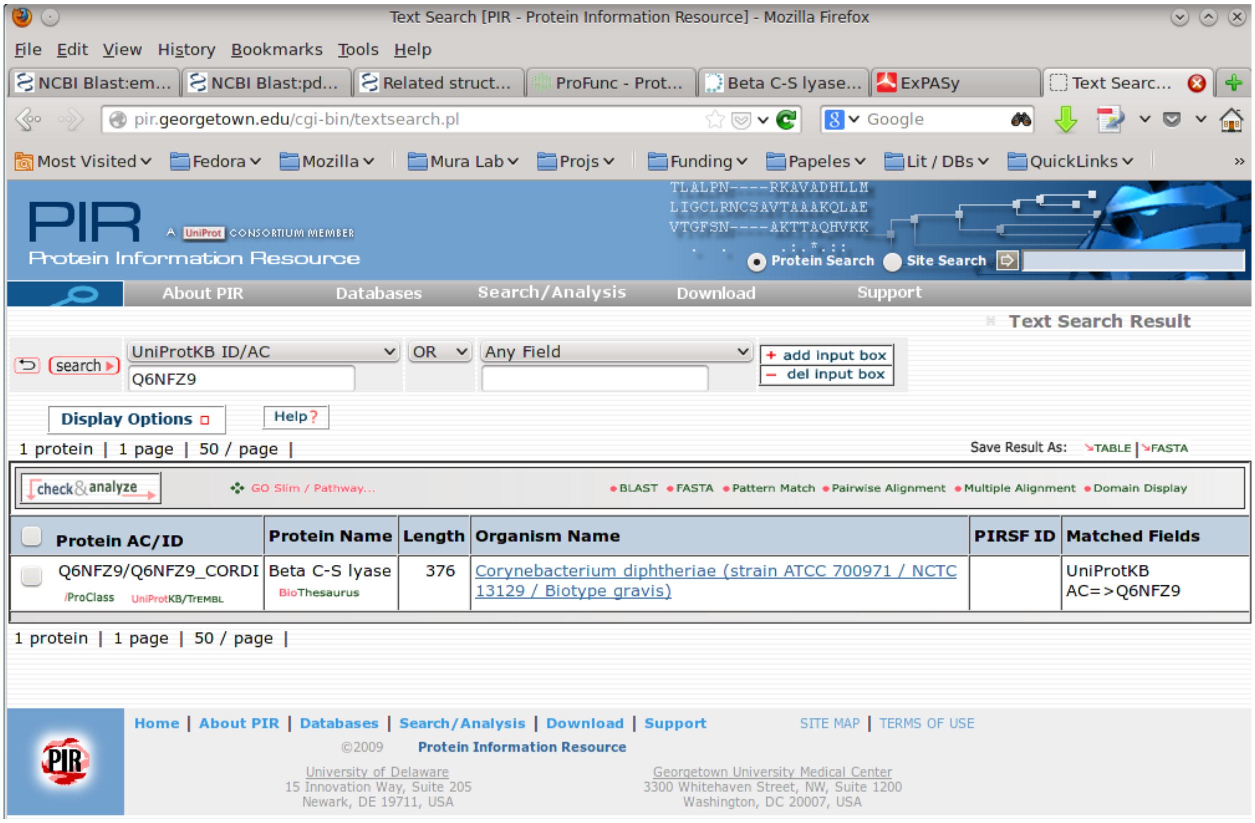
The ProtParam tool provides a thorough list of the ***physicochemical properties of the POI***, including properties that will be very useful for us to keep in mind even for experimental work (the protein MW, extinction coefficient @ 280nm, and so on). Here's the complete list (most useful bits highlighted):

```
Number of amino acids: 376
Molecular weight: 41722.3
Theoretical pI: 5.13

Amino acid composition:
Ala (A) 45 12.0% Arg (R) 17 4.5% Asn (N) 12 3.2% Asp (D) 23 6.1% Cys (C) 6 1.6% Gln (Q) 16 4.3% Glu (E) 21 5.6% Gly (G) 24
6.4% His (H) 11 2.9% Ile (I) 26 6.9% Leu (L) 29 7.7% Lys (K) 12 3.2% Met (M) 7 1.9% Phe (F) 21 5.6% Pro (P) 23 6.1% Ser (S) 19
5.1% Thr (T) 22 5.9% Trp (W) 8 2.1% Tyr (Y) 13 3.5% Val (V) 21 5.6% Pyl (O) 0 0.0% Sec (U) 0 0.0% (B) 0 0.0% (Z) 0 0.0% (X) 0
0.0%

Total number of negatively charged residues (Asp + Glu): 44
Total number of positively charged residues (Arg + Lys): 29

Atomic composition:
Carbon      C         1890
Hydrogen    H         2869
Nitrogen    N          497
Oxygen      O          547
Sulfur      S           13

Formula: C1890H2869N497O547S13
Total number of atoms: 5816

Extinction coefficients:

Extinction coefficients are in units of  M-1 cm-1, at 280 nm measured in water.

Ext. coefficient    63745
Abs 0.1% (=1 g/l)   1.528, assuming all pairs of Cys residues form cystines

Ext. coefficient    63370
Abs 0.1% (=1 g/l)   1.519, assuming all Cys residues are reduced

Estimated half-life:
The N-terminal of the sequence considered is M (Met).
The estimated half-life is: 30 hours (mammalian reticulocytes, in vitro).
                            >20 hours (yeast, in vivo).
                            >10 hours (Escherichia coli, in vivo).

Instability index:
The instability index (II) is computed to be 30.34
This classifies the protein as stable.

Aliphatic index: 85.21
Grand average of hydropathicity (GRAVY): -0.075
```

4.4: "*Does your protein have a PIR link?... …. … If yes, follow the PIR (Protein Information Resource) link under cross-references and access the iProClass of the protein. What information is contained on this site?* "

Unfortunately this particular protein ('Q6NFZ9' in UniProt-speak) does **not** have a PIR link in the Cross-references section of the protein page. So, we go directly to the PIR website…

…in so doing, we get a page that looks like this:



● Protein Search ○ Site Search

Text Search Result

UniProtKB ID/AC: Q6NFZ9 | OR | Any Field:

1 protein | 1 page | 50 / page |

Save Result As: TABLE | FASTA

□ Protein AC/ID	Protein Name	Length	Organism Name	PIRSF ID	Matched Fields
□ Q6NFZ9/Q6NFZ9_CORDI iProClass UniProtKB/TrEMBL	Beta C-S lyase BioThesaurus	376	Corynebacterium diphtheriae (strain ATCC 700971 / NCTC 13129 / Biotype gravis)		UniProtKB AC=>Q6NFZ9

1 protein | 1 page | 50 / page |

©2009 Protein Information Resource

University of Delaware
15 Innovation Way, Suite 205
Newark, DE 19711, USA

Georgetown University Medical Center
3300 Whitehaven Street, NW, Suite 1200
Washington, DC 20007, USA

Click on the small *i*ProClass logo (lower left, just under the Protein AC/ID field), and this leads us to another useful page which contains all sorts of (thus-far-new) info, such as enzyme/function annotation, pathway (KEGG), UniRef and Pfam domain links (protein families), InterPro (which is an umbrella of sorts for all of this—an aggregator of sorts for much of the bioinformatic data), and so on……. This is another treasure trove of information about our POI.

4.5: "*What Pfam and COG is the protein a member of?....*"

Pfam: PF00155: Aminotransferase class I and II (23-368)

COG: Uh-oh, is this no longer being maintained? Looks like it—"*The previously existing COG pages are retired.*" (as of 02oct2013, at http://www.ncbi.nlm.nih.gov/COG). Nevertheless, I was able to find the search tool http://oxytricha.princeton.edu/cgi-bin/BlastO/blastO.cgi; the links to NCBI COG pages are broken…; but able to see this (for what it's worth…):

First, *change the pull-down menu to 'COG' and not 'KOG'* (the latter is restricted to Eukarya)… Then:

Hit #1: Ortholog group COG1168 Taxa: 19 Genes: 21 [E]Bifunctional PLP-dependent enzyme with beta-cystathionase and maltose regulon repressor activities. Weighted score: 108
Hit #2 is Ortholog group COG0436 … … ...

4.6 "*Returning to the iProClass page, what is the PIRSF (PIR Super Family) id and classification of the protein? This number has the format PIRSF######.*"

Can use http://pir.georgetown.edu/pirwww/search/pirsfscan.shtml to scan for PIRSF number if it's not listed anyplace else (note that the PIRSFscan tool on the PIR website states "A UniProt consortium member", so it makes sense that the above strategy can work). For our particular protein, unfortunately we obtain the following: "*Your query sequence does not match to any of the fully curated PIRSFs*." Ack!

4.7 "*Follow the link of the PIRSF ID. What information is contained on this site?*"

N/A, unfortunately, because of the lack of a PIRSF ID for this POI (see answer to 4.6).
To try to make lemonade from this lemon, we notice something while sitting at http://www.uniprot.org/uniprot/Q6NFZ9. Specifically, if we click on the "EMBL CAE50265.1" link (brown box near the top, near the "Protein names" section), we are led to a useful page http://www.ebi.ac.uk/ena/data/view/CAE50265 which provides suggestions like "Similar to Corynebacterium glutamicum beta C-S lyase AecD TR:Q46061 (EMBL:M89931) (325 aa) fasta scores: E(): 3e-66, 53.93% id in 330 aa" and provides links to other DBs.

4.8 "*Do the PIRSF, COGs, and Pfam results suggest a function? If so, what is it?*"

See above answer to 4.7.

4.9 "*Go to the Prosite page (http://prosite.expasy.org) and type in the UniProt ID of your POI. Does your protein have any Prosite hits? If so, select one and describe what the consensus pattern represents. Does this help you narrow down the function of the protein?*"

Well, this is too bad: When we search Prosite using either the PDB ID (3fdb) or the UniProt ID (Q6NFZ9), we obtain ***no hit***. That's if we use the default settings for the motif search, which includes the option to "*Exclude motifs with a high probability of occurrence from the scan*" (post-translational modifications (PTMs) such as phosphorylation or glycosylation; also this will retrieve compositionally biased/'low-complexity' regions, etc.). Now, if instead we disable that option, then we see some hits—but these are for phosphorylation and *N*-myristoylation sites in the case of our POI, therefore not so illuminating in terms of our assessing potential aminotransferase (AT) substrate specificities…

So……

Hmm…

Step 5: Search against structural family databases for structural classification

5.1 *"From the iProClass (step 4.4) scroll down to the SCOP Fold section within the Family Classification section."*

OK… do this.

5.2 "*How many SCOP superfamilies does your POI have?*"

There is no SCOP entry for this POI. But there is a CATH entry, and CATH is—like SCOP—a protein domain family classification scheme. So, we click on the CATH link (right next to SCOP link). And then…? Broken link ("Sorry, couldn't get CATH_WEBSERVER_HTTP_BASE from parameter file….")! But, do not despair, we can try to be resourceful & stubborn about it: Click 'Search' on the CATH menu bar, enter '`3fdb`' in the text search field. We get a hit for `3fdb`, but it's quite thin—no information! So, then what? Down near the bottom of the page, click on 'Sequence' tab. Then click on "BLAST against CATH" button. Then that leads to two hits – "Blast: Domain `1c7nA02`" and "Blast: Domain `1d2fA02`". Both belong to "Superfamily `3.40.640.10`: "Type I PLP-dependent aspartate aminotransferase-like (Major domain)" and with only ~29% sequence ID (so in the "twilight zone" of sequence similarity w.r.t. homology!)… So, maybe this could be useful…? – assessing that would require a further look to see if there's substrate specificity profiles for that superfamily 3.40.640.10.

NB: *SUPERFAMILY* (http://supfam.cs.bris.ac.uk) is also a valuable resource in this type of work.

5.3 "*Click the link next to family. What is provided? How can this help you figure out a likely function of your protein?*"

See above. It's too bad that we can't get to a SCOP entry for this POI…

Step 6: Identify functional residues (NB: This is a *very* crude and impartial treatment here… just a first pass relative to what you'll have to do overall, this semester and next.)

6.1 Go to PDBSum (the link is not in table so you will need to search for it) and search with your PDB id.

OK, doing that led to http://www.ebi.ac.uk/pdbsum/3FDB. (To be efficient in looking-up other PDB entries, note how the URL is constructed.) Also another note: PDBsum entry for this POI is linked from its UniProt entry (in the 'Cross-references' section at http://www.uniprot.org/uniprot/Q6NFZ9).

6.2 What information is available about the protein?

There is a wealth of structural info available on the PDBsum entry, ranging from structure validation (PROCHECK output to make Ramachandran plots), to very sophisticated structure-derived properties, such as a "tunnel analysis" of the holes/voids, connected tunnels, through the protein. What looks to be especially powerful with regards to substrate specificity is the PDBsum's emphasis on ligands, metals, cofactors, etc. in the PDB structure... i.e., not just sequence-derived properties but also 3D structural. There are also cross-references and links to all the other DBs such as UniProt and Pfam, as well as some databases that we haven't seen before—e.g. 'ArchSchema' (for Pfam domain analysis), a "*ligand cluster analysis*" tool for all PDB entries related to this one (often this will be same exact protein but different crystallization conditions, maybe co-crystallized with ligands or substrate, etc.). Over on the LHS there is a very useful hierarchical menu that links to protein, ligands, metals, etc. for this PDB file.

6.3 There are tabs across the top of the page (protein, ligands, prot-prot, clefts, and links). Go to ligands. In a few weeks, you will look at the PDB to try to figure out which ligands are biologically relevant, so bookmark this page. In the meantime, ignore inorganic salts and focus on small organic molecules. Click on the ligands and find out what they are; explore and report what you can find about your POI.

In this case we click on the "ligands" link mentioned above and find that there is not too much of interest—just 9 molecules of 'EDO' in the asymmetric unit. 'EDO' is the residue code for 1,2-Ethanediol (ethylene glycol), which was probably used as a 'cryo-protectant' to prevent formation of ice when flash-cooling the crystals for x-ray diffraction measurements at 100 K.

6.4 Click on the LIGPLOT (below the LIGPLOT link) to view. Save the image and list the residues that directly interact with cofactors.
Note: This is only for ligands that co-crystallized with the protein. Substrates and co-factors may not be present and you should compare with other known structures and biochemical studies of similar proteins that may have cofactors bound to identify functional residues in your protein. ← This is the iterative, looping back-n-forth nature of any bioinformatics-informed study – now the next thing you would do is visit "structural neighbors" (see below and above) of your POI, and then see what the PDBsum entry for these looks like—in particular, are there ligands/cofactors/substrates (or substrate analogs) bound in any of those that might have some implications for your POI?

We do that, and we see that the ethylene glycol····POI interactions are quite sparse and fairly non-specific in nature: (i) a H-bond between the EDO hydroxyl and the hydroxyl of a SER side chain; (ii) H-bond between EDO hydroxyl and main-chain carbonyl oxygen from a proline residue (…so that's quite generic [not sequence-specific], and therefore unlikely to have anything to do with substrate specificity or any specific enzymatic function of this particular POI).

Step 7: Search for structural neighbors and perform structure-guided alignment

7.1 Go to VAST and search with your PDB id

OK, did so (http://structure.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml), landed at a page specific for `3fdb`. One new piece of potentially useful (later) info we find on this page is the **`MMDB ID: 68462`** for this POI.

7.2 Scroll down to the "Molecules and interactions" section. Click on "show annotation." Hover over the red/pink superfamily box and the gray COG box.

In terms of superfamily (pink and red boxes), we see annotation as "AAT_like" and "AAT_I" superfamilies, consistent with our earlier findings (Stage 4, above). Also "MalY" at the (grey box) COG level. Also, clicking the grey COG box leads us to http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=COG1168. This is interesting—the fact that we are pointed to COG**1168** is consistent with what we found earlier, via a more circuitous approach (see above, Stage 4.6).

7.3 Do the annotated superfamily and COG of VAST agree with your Prosite analysis (see 4.9)?

Yes, see above answer to 7.2

7.4 Scroll back up and click on the gray VAST button next to "Similar Structures." In the pop-up, select the appropriate alignment range (entire chain is the full protein; you may want to come back and select specifically for the conserved domain). Select the top 3 structures (**or a structure that you have literature on in terms of active site residues**) by clicking in the box on the left and then click "View 3D Alignment" up at the top (you need to have Cn3D installed). List the 3 structures you selected.

To be more explicit: In the VAST "Similar Structures" pop-up, click on the `Alignment range ("1-377")` hyperlink. Here are the three structures that I selected: **`2r5e_A`**, **`1c7n_a`**, **`1kmj_a`** (NB: many of these ATs seem to be dimeric, so not surprising that there are often 2 chains in the *asymmetric unit* of the PDB file, hence the '`_a`', '`_b`' suffixes that we often see accompanying these PDB IDs).

7.5 Go to DALI and search using your POI PDB id. What hits do you get that are structurally similar? (You can copy and paste results of 10 structurally similar hits. Be sure your hits are not redundant (e.g., chains A, B, C, and D of one protein).)

Visit http://ekhidna.biocenter.helsinki.fi/dali_server/start.

Top 10 structurally similar hits:

```
No:  Chain   Z    rmsd lali nres  %id PDB   Description
1:  3fdb-A   68.4  0.0  377   377  100 PDB   MOLECULE: PUTATIVE PLP-DEPENDENT BETA-CYSTATHIONASE;
2:  3l8a-A   49.8  1.7  369   382   28 PDB   MOLECULE: PUTATIVE AMINOTRANSFERASE, PROBABLE BETA-CYSTATHI
3:  3l8a-B   49.6  1.8  371   384   27 PDB   MOLECULE: PUTATIVE AMINOTRANSFERASE, PROBABLE BETA-CYSTATHI
4:  4dq6-B   48.7  1.9  372   388   27 PDB   MOLECULE: PUTATIVE PYRIDOXAL PHOSPHATE-DEPENDENT TRANSFERAS
5:  4dgt-A   48.7  1.9  372   387   27 PDB   MOLECULE: PUTATIVE PYRIDOXAL PHOSPHATE-DEPENDENT TRANSFERAS
6:  4dq6-A   48.6  1.9  372   388   27 PDB   MOLECULE: PUTATIVE PYRIDOXAL PHOSPHATE-DEPENDENT TRANSFERAS
7:  4dgt-B   48.6  1.9  372   387   27 PDB   MOLECULE: PUTATIVE PYRIDOXAL PHOSPHATE-DEPENDENT TRANSFERAS
8:  3b1d-A   48.4  1.8  372   388   29 PDB   MOLECULE: BETAC-S LYASE;
```

```
9:  3b1e-C   48.4  1.8  371   386   29 PDB    MOLECULE: BETAC-S LYASE;
10:  3b1d-B 48.4  1.8  371   388   30 PDB  MOLECULE: BETAC-S LYASE;
```

How is structural similarity assessed? Are the hits the same as the results you obtained with VAST?

> Structural similarity is assessed by the *Z* score and by RMSD, "lali" is the length of the alignment (which we want to be most of the protein). The DALI results page states "Similarities with a Z-score lower than 2 are spurious." A perhaps unexpected finding is that `1c7n` and `1kmj` (from VAST) do show-up in the DALI results, albeit much lower than the 10th hit, while the other VAST hit (`1r5e`) does not appear in the DALI output.

You could download the PDB for structure alignment from the DALI hits and explore in a few weeks with PyMOL. (You need to open the PDB you searched with and the downloaded hit in PyMOL one right after the other, and make sure not to move the molecule before opening both files.)

> Yes, note that the DALI results page also states "*Each neighbour has links to pairwise structural alignment with the query structure, to pre-computed structural neighbours in the Dali Database, and to the PDB format coordinate file where the neighbour is superimposed onto the query structure.*" This is the 'PDB' link in the tabular formatted results, and we could download some of those if we wished to.

Step 8: Phylogenetic trees and analyses

8.1 Go back to the front VAST page, scroll down to the "Molecule and interactions" box, click on "show annotation." Click on "Domain Families" (this is to the left of the sequence schematic). Mouse over the specific superfamily hit (red bar under gray query sequence bar). Read the pop-up and then scroll to the bottom of the box to find the cd ID; record that ID.

Mousing-over the "specific hit" gives a pop-up that contains the following info, including the cd ID:

> [Specific hit] **cd00609**, Aspartate aminotransferase family. This family belongs to pyridoxal phosphate (PLP)-dependent aspartate aminotransferase superfamily (fold I). Pyridoxal phosphate combines with an alpha-amino acid to form a compound called a Schiff base or aldimine intermediate, which depending on the reaction, is the substrate in four kinds of reactions (1) transamination (movement of amino groups), (2) racemization (redistribution of enantiomers), (3) decarboxylation (removing COOH groups), and (4) various side-chain reactions depending on the enzyme involved. Pyridoxal phosphate (PLP) dependent enzymes were previously classified into alpha, beta and gamma classes, based on the chemical characteristics (carbon atom involved) of the reaction they catalyzed. The availability of several structures allowed a comprehensive analysis of the evolutionary classification of PLP dependent enzymes, and it was found that the functional classification did not always agree with the evolutionary history of these enzymes. The major groups in this CD corresponds to Aspartate aminotransferase a, b and c, Tyrosine, Alanine, Aromatic-amino-acid, Glutamine phenylpyruvate, 1-Aminocyclopropane-1-carboxylate synthase, Histidinol-phosphate, gene products of malY and cobC, Valine-pyruvate aminotransferase and Rhizopine catabolism regulatory protein.

8.2 Click on the superfamily bar under the gray query sequence bar.

Mousing-over the "superfamilies" bar gives a pop-up that contains the following info:

> [Superfamily] **cl00321**, Aspartate aminotransferase (AAT) superfamily (fold type I) of pyridoxal…. . …. … .....

8.3 Select the Specific Domain Family in the Curated CD Hierarchy based on the cdID that you recorded in step 8.1.

Begin the navigation process by clicking on the red (specific hits) or the pink (superfamilies) bar—either one works fine (if you click on the pink then it's just a matter of locating the cd ID in the list and then clicking on it).

8.4 On the left hand side, click Interactive Display in the Hierarchy box. (You will need to have CDTree installed.)

NB troubleshooting: If you have Cn3D version >= v4.3 installed, then note that CDTree is bundled with it, rather than being provided as a standalone program. So, if you click on 'Interactive Display' and find that nothing occurs, then simply save the `.cn4` file to your local folder and manually open it in Cn3D.

8.5 In the sequence tree window, deselect Fit to Screen under the View pull-down.

There will be three windows that open with CDTree: (i) the *main CDTree 3.1 control window* (shows the CD Accession numbers), (ii) a *Taxonomy Tree* window, and (iii) a *Sequence Tree* window, which shows the actual sequence tree. Deselecting the Fit-to-Screen option is a good idea when there are many sequences displayed in the tree.

8.6 In the Taxonomy Tree window select your PDB id from the pull-down. If you don't find your protein find a closely related protein (one you identified in a previous step).

We don't see `3fdb`. BUT, all is not lost! – In the main CDTree control window, right click on `cd00609` entry (the only one there), then "Sequence List", then sort by "_Structure" (pull-down menu), and then we see our friend (from VAST and DALI [noted above])—the `1C7N` structure that keeps reappearing. So we can look at it on the tree. What's interesting about the relative position(/depth) of a particular POI on the tree? — Well, for one thing it could imply something about the degree of 'specialization' (substrate specificity) of the protein—if it's a particularly deep-branching ('basal') species, then it likely has a very generic function ("aromatic AATase"), whereas if it's a phylogenetically more recent POI then it may have specialized more in terms of substrate cofactors etc. ("a tyrosine AATase"). But, again, this type of information is a very weak signal, and what you have to do it correlated it with a lot of other information from (a) databases, (b) literature, and, ultimately, (c) experimental data.

8.7 Explore the tree and identify where in the tree the proteins you were using for the sequence and structure alignments are.

OK… (Do this.)

Step 9: Evidence-based assignment of biological function for your POI

Grand finale: Based on the data that you have collected, predict and describe the function of your POI as accurately as possible, and with as much precision as you think you can (again, given the data you've collected!). Also, explain what data supports what parts of your conclusion(s). Include cofactors, residues involved, other proteins that have the same function and data about them. **The more effort you put into this description, the easier your later assignments/work will be**. Keep track of your literature citations, bookmark useful sites and web-pages [1], tabulate your results from databases, and so on.

[1] NB: Many of these web-pages, especially for POI-specific results, are dynamically generated, on-the-fly, *via* CGI methods. CGI is the *Common Gateway Interface*. Such links often contain '`cgi-bin`' in the URL, so don't be frustrated if you try to visit such a page at a later time but find it blank (the URLs are often also dynamically generated, so they will probably expire). Therefore, it is of utmost importance that you carefully document the procedures you use (your *workflow*) to generate bioinformatic data—especially if you alter any default query/display/etc. settings on the tool's pages (e.g., in BLAST, in ProSite motif scans [see Stage 4.9 above], etc.).

This Step is left as an exercise to the reader ☺.