UCSC BME 205 assignment: protein information

Due Wed 20 Nov 2013
(Last Update: 10:04 PST 7 November 2013)

Find out information about a new protein

The purpose of this exercise is to guide you through the process of finding information and making predictions about the structure and function of a protein given a gene for it.

What you turn in should be a stand-alone paper that a biologist or bioinformatician can read without having any prior knowledge of this class or of the protein. Remember that biologists like to look at figures. If you can show alignments, structure predictions, repeat structure, domain structure, or anything else pictorially, it will probably make your paper more attractive to a biologist. Biologists are much more likely to read and understand a report if there are pictures illustrating the key points!

Be sure to provide proper citations for all papers and web sites that you get information from. You should cite a paper for each tool you use (they generally tell you what to cite). A bare URL is not an adequate citation for a web site—you need to provide enough information that someone can find it with google if it has moved without being changed—title, URL, and date of publication or date of access is minimal, and author or corporate author should be provided whenever possible. A typical number of citations in the past has been around 20.

This paper really should look like a report on the protein, not like a homework exercise. I have given some suggestions below to help you get started, but these are not questions to answer sequentially, nor are they necessarily the most productive directions for your search.

Don't just print out the results of web searches, but interpret the results to see what (if anything) they say about the target protein. Please be precise in your descriptions of what you did: Don't just say "blast" but give what version of blast searching what database.

To make the project interesting, I selected a protein that is medically interesting and that has some literature, but for which a lot is still unknown. Rather than give you a protein sequence, I'll give you a DNA sequence of a PCR product that contains the gene for the protein (gene.fasta). Your first task will be to find the long ORF that codes for the protein and translate it. Remember that it may be any of the six coding frames (three on each strand). You may write your own code to do this, or use any of the many web tools, graphical tools, or command line tools for this common task.

This particular PCR product was extremely difficult to sequence, as it is longer than can be sequenced with Sanger sequencing and has long internal repeats that make finding unique primers almost impossible. It was eventually sequenced with PacBio long reads, and the current assembly is believed to be correct.

PacBio sequencing has a high error rate for single-base indel (insert-delete) errors, though, which would cause frameshifts in the coding region. You should look for possible frameshifts.

Once you have found the protein, you should find out everything you can about it, either in the literature or by bioinformatic means.

Find the sequence

One of the first things to do with a new sequence is to find out whether it is already known and named. A common tool for doing such a search is NCBI blast. We have blast installed on the School of Engineering machines (as /projects/compbio/bin/x86_64/blastall) and we do weekly updates of nr to /projects/combio/data/nrp/nr If you run it, you may want to set up a .ncbirc file containing

```
[NCBI]

Data=/projects/compbio/programs/blast2/data
```

And then run

```
blastall -p blastp -d /projects/compbio/data/nrp/nr -i protein.fasta
```

Alternatively, you can do what biologists around the world do and use the NCBI website: http://blast.ncbi.nlm.nih.gov/Blast.cgi

Use the blastp program or website to do protein-protein search of the nr (non-redundant protein) database. (Note: you can also use blastx to search protein databases with the original DNA sequence or tblastx to search DNA databases with the DNA sequence, looking for similar proteins.) Because the nr database has gotten large (over 33 million sequences), this may take quite a while—the NCBI site is likely to be faster than running the search on our machines, except in the middle of the afternoon.

Get the name(s) and organism of the closest matches for the sequence. Remember to italicize species names. If you abbreviate the genus name (like *E. coli*), remember to use an unbreakable space after the period, to avoid unfortunate line breaks.

Literature search

First you should look at what has already been determined (or predicted) by previous researchers. Use resources like Swissprot http://us.expasy.org/, the human genome browser http://genome.ucsc.edu, the archeal and prokaryotic browser http://microbes.ucsc.edu, and organism-specific databases (SGD for yeast http://www.yeastgenome.org/, flybase for Drosophila http://flybase.org/), ...) to find information about the sequences you found with BLAST.

The NCBI blast search conveniently provides links to Unigene, Entrez Gene, Medline, and even PubChem BioAssay databases, which makes the web search much easier. Remember that PubMed is a medical database, so will tend to have more articles about human proteins and pathogen proteins than about similar proteins from other organisms (though popular model organisms may have quite a few articles).

Do google searches using the protein name and its accession number or database identifier(s) to try to find web pages about the protein.

Use PUBMED and other databases at Entrez (now mysteriously renamed GQuery)http://www.ncbi.nlm.nih.gov/gquery to find papers that talk about the protein.

For some proteins, you may want to use BIOSIS from the library website http://library.ucsc.edu/ to see if there are articles there. (BIOSIS is better at plant biology and non-pathogenic microbiology, for example, than PUBMED is.)

Remember that reference list should contain all and only those papers cited in the main body of your paper. Don't pad your reference list with papers that you didn't actually cite. (LaTeX and BibTeX take care of this for you automatically, and I've heard that EndNote, Zotero, and Mendeley also work.) If you do use BibTeX, remember that \cite can take a comma-separated list of citations, and that this is the right way to do multiple citations at a single location.

Find out what else you can get from the protein sequence

After you've done a literature search, you should find out what else you can about the protein by bioinformatic means. This could include such things as looking for homologs, looking for internal repeats, splitting up into domains, looking for transmembrane helices or other special features, doing protein-structure prediction, and so forth.

The blast site has several other programs (psi-blast, for more remote protein homology; rpsblast, for conserved domains; ...). Find the probable homologs of the sequence, find out what can be expected about it based on the homologs. Explore and summarize what you can find.

Remember that "hypothetical protein" as an annotation does not tell you anything about how "real" a protein is, just that there was no direct experimental evidence for the protein at the time of the annotation. Annotators are encouraged to be rather cautious in putting functional identification of proteins into the database, since false positives are much more damaging than false negatives. Since the annotation is rarely updated, even proteins that have now had extensive experimental work may still be labeled as "hypothetical" in some databases.

One popular thing to do is to check for known protein domains, using tools like Pfam (available on-line at http://pfam.janelia.org/) and SUPERFAMILY (available on-line at http://supfam.cs.bris.ac.uk/SUPERFAMILY/). Prosite http://prosite.expasy.org/ can also be useful, though you have to be aware for the high probability of false positives.

If you find some good hits to domains or prosite motifs, do some literature search on them also, so that you know roughly what they do and what they tell you about the structure or function of the protein. Summarize your findings.

Another popular thing to do is to check for transmembrane helices and secretion signals. There is a good suite of tools at the Technical University of Denmark: http://www.cbs.dtu.dk/services/ and I've found TMHMM and SignalP to be particularly useful. You should be aware that TMHMM does a good job of identifying transmembrane helices, but is not much better than random at deciding what is inside and what is outside the cell. I believe that Phobius at http://phobius.sbc.su.se/ gets the inside/outside prediction somewhat better, but it believes that TM helices near the beginning of the sequence are all signal peptides, which is a different sort of error.

Finding homologs

It is often useful to get a large number of putative homologs to your target sequence—both to find annotation about the function and to make multiple alignments for looking for conservation signals. You can get a quick list with BLAST, but this will only provide sequences that are rather similar, and you can get some confusion with multiple-domain proteins that only match on one or two of the domains.

Your best bet (usually) is to break the protein up into domains, and do searches for homologs on each domain separately. If you restrict yourself to domains that do not contain transmembrane helices or transmembrane beta barrels, then you can try submitting the domains to structure prediction servers also.

One of my favorite ways of finding homologs for a protein (or protein domain) is to use the SAM-T08 server (at http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html), which not only finds the probably homologs and aligns them, but produces sequence logos, secondary structure predictions, and tertiary structure predictions. It is a bit slow, though, particularly on long proteins, so you might want to also try the more popular PSI-BLAST method (at the BLAST website: http://blast.ncbi.nlm.nih.gov/Blast.cgi).

You might want to do a fast HMM-HMM search like HHPred (or use the other tools there, like CS-BLAST and HHBlits).

If you do manage to get SAM-T08 to run, take a look at the sequence logos for secondary-structure prediction also. Do the proteins have strongly or weakly predicted secondary structure? Are the conserved residues in areas of strongly predicted secondary structure? Does the SAM-T08 software make a strong prediction for tertiary structure? What E-value does it give to the best template?

You might also want to give the sequence to a metaserver such as http://pcons.net/.

Multiple alignments

Once you have a collection of sequences, it is useful to make a multiple alignment of them. There are many methods for doing this (indeed, psi-blast, CS-blast, HHBlits, and SAM-T08 provide multiple alignments that may be all you need to work with). Multiple alignments tend to be more useful if there is a moderate diversity of sequences: many almost identical sequences tell you little when aligned, and a few very different sequences may be difficult to align accurately.

If you are given a set of sequences without a multiple alignment, or if you do not quite believe the multiple alignment you got from psi-blast or SAM-T08, you may wish to realign the sequences with a different tool.

One very popular (though no longer considered very good) tool is CLUSTALW. This is a progressive method of multiple alignment. It will do all-pairs scoring on a sequence set, then build a guide tree with the sequences on the leaves. Sequences with a high similarity score are assigned to nodes with a common parent on this tree. The alignment is built from the bottom of the tree by merging sibling sequences into pairwise alignments, and then progressively merging the most similar pairwise alignments into multiple alignments.

Since ClustalW is rather ancient code, with poor performance relative to newer tools. I recommend using Clustal Omega instead. You can try it out at the EBI web server at http://www.ebi.ac.uk/Tools/msa/clustalo/. Clustal Omega can handle huge numbers of sequences, but the web server may limit you.

If you generate a Clustal Omega, compare it to alignments found by other methods (BLAST, PSI-BLAST, SAM, MUSCLE, ...). Where the alignments differ, which one looks more reasonable to you? What positions contain highly conserved residues? Do the sequence logos from the SAM site suggest any conservation to look for that you did not expect from having just looked at the multiple alignments? (The SAM site may be too slow for a long protein. If you have a multiple alignment in A2M format, you can produce the sequence logos with /projects/compbio/bin/makelogo (online documentation). All aligners seem to use different output formats, many of which do not support the notion of insertions between alignment columns, so you might have some difficulty getting an alignment into A2M format

Another good multiple alignment program is MUSCLE (see http://www.drive5.com/muscle/). You can use MUSCLE to align your sequences and see how it differs from Clustal Omega or psi-blast.

Viewing with Rasmol

If you got any strongly predicted protein structures, try to look at them with rasmol, pymol, vmd, jmol, or some other structure-viewing tool.

If you are on a School of Engineering machine, you can download a protein from PDB with

```
/programs/compbio/bin/pdb-get 1foo
```

where *1foo* should be replaced by the proper pdb identifier. This program returns the name of the file that has been downloaded, so you can use

```
rasmol `pdb-get 1foo`
```

to look at proteins, assuming that your paths are correctly set up.

If you need to download Rasmol for your home computer, there are several sources, including http://www.bernstein-plus-sons.com/software/rasmol/. Rasmol is a command-based viewer, and you will have to use "help" a lot while learning to use it. The download site listed above also has pointers to the web-based Rasmol manual.

Note: there are many other protein viewers on the web (DeepView=Swiss-pdbviewer, molmol, chime, protein explorer, molscript, vmd, jmol, firstview, cn3d, pymol, kinemage, ...). If you wish, you may substitute some other viewer for rasmol.

Look at the protein in various ways (as cartoons, as ball-and-stick models, as a backbone trace, ...). For example, in rasmol, with the protein in cartoon view, use "Select hetero and not HOH and not MSE" to select ligands (if there are any), and view them in space-filling mode.

Where are there insertions or deletions in the target relative to the template you chose? Are these in sensible places?

Microarray data

It it seems appropriate, look for microarray data on expression patterns for the gene associated with this protein. What information (if any) can you glean from the databases? I don't know which microarray databases are the easiest to use or the most informative, as I have rarely used them. I have found that the SGD database for yeast has good links to an expression database that does some useful clustering, but I have not found a really good clustering site that uses the public databases.

Note: there is a strong possibility that this protein is not closely related to proteins from any of the model organisms---or that it is related to lots of proteins which don't all share the same function. Discuss the difficulties as well as the successes!

What can you conclude?

Based on your study of the protein, what can you conclude about it? What experiments would you want someone to do to get more information about the structure and function of the protein? If you wanted to change the function in some way, which residues would you consider mutating?

Things learned after assignment

SoE home

Kevin Karplus's home page

Biomolecular Engineering Department

BME 205 home page

UCSC Bioinformatics research

Questions about page content should be directed to Kevin Karplus
Biomolecular Engineering
University of California, Santa Cruz
Santa Cruz, CA 95064
USA
karplus@soe.ucsc.edu
1-831-459-4250
318 Physical Sciences Building