

BSC4933/ISC5224: Introduction to Bioinformatics

Laboratory Section: Wednesdays from 2:30 to 5:00 PM in Dirac 152.

Protein Sequence Attributes

Lab Nine, Wednesday, March 4, 2009

Author and Instructor: Steven M. Thompson

Estimating protein secondary structure and physical attributes:

The various methods, their usefulness, and their limitations are all covered. This includes proteolytic digestion mapping, molecular weight and amino acid composition determination, isoelectric point estimation, hydrophobicity and hydrophobic moment determinations, surface probability and antigenicity mapping, and secondary structure prediction, with an emphasis on homology based inference methods (e.g. PredictProtein).

Steve Thompson
BioInfo 4U
2538 Winnwood Circle
Valdosta, GA, USA 31601-7953
stevet@bio.fsu.edu
229-249-9751

[†]GCG[®] is the Genetics Computer Group, a product of Accelrys Inc.,
producer of the Wisconsin Package[®] for sequence analysis.
© 2009 BioInfo 4U

Introduction — protein secondary structure and physical attributes —

What can I learn about my protein's physical properties and secondary structure from its sequence?

Grateful acknowledgement of Susan J. Johns (University of California, San Francisco) for contributing much to this tutorial from our days together at Washington State University (1990–1998).

The determination of protein secondary structure has been an intriguing puzzle. When Linus Pauling first predicted that proteins would be composed of alpha helix and beta sheet units in 1948, no protein structures had yet been experimentally determined. His prediction was based solely on the idea that the potential hydrogen bonding possible in such structures would increase their stability and make them more probable.

Initial studies on homo-polypeptides provided the first experimental evidence for secondary structure in proteins. Improvements in x-ray diffraction techniques made it possible to solve complete protein three-dimensional structures, and Pauling's predicted subunits were found to be present. As more protein structures were solved it appeared that the conformation of residues in proteins was similar to their homo-polymeric form. This correlation is far from perfect, however.

Perceptible folding patterns were recognized in soluble, globular proteins, as more and more structures were experimentally determined. According to Chothia (1984), as paraphrased by von Heijne (1987), "The principle underlying the structure of helices, sheets, and turns is the simultaneous formation of hydrogen bonds by buried peptide groups and the retention of single residue conformations close to those of minimum energy. The shape of the helix and sheet structures make these structural elements pack together in a small number of relative orientations. The links between secondary structures tend to be right-handed and short, and do not form knots." As a result soluble, globular proteins usually fold into set particular patterns, which can be roughly grouped into four arbitrary classes: all alpha, all beta, mixed alpha/beta formed from beta-alpha-beta units, and alpha + beta where the helix and sheet units are segregated.

As the number of determined structures rose in PDB, several research groups undertook statistical studies to determine the preferences for different individual amino acid to exist in given secondary structures. These efforts resulted in the classic empirical prediction schemes of Chou-Fasman (e.g. 1974) and Garnier-Robson (1978). The Chou-Fasman method is a group of rules applied to a given sequence. It was an ambiguous method that proved difficult to automate. The Garnier-Robson method is based on the consistent application of information theory with auxiliary information from circular dichroism (CD) used to bias its prediction. This method was unambiguous and easy to automate. Both methods are incorporated into the GCG Wisconsin Sequence Analysis Package, though much better methods are now available.

Dichroism measures the difference in polarized light transmitted through a sample. In circular dichroism, the light is not only polarized, but also caused to move in both the right and left directions in a circular manner. Chiral molecules, those that are not super-imposable on their mirror images, cause circularly polarized light to rotate differently in these two directions. CD devices measure this difference over a range of wavelengths for a given sample and output the results as a spectrum. CD studies can be used to determine experimental

secondary structure estimates by interpreting the spectra produced. They do not specify where those secondary structures lie, only in the percentage of each type of structure in the specimen.

Hydrophobicity and the hydrophobic moment

Hydrophobicity is a measure of how much a molecule hates water (hydro=water, phobia=fear). Each amino acid can be designated with a hydrophobicity value. This has been done by many researchers, hence the abundance of different hydrophobicity scales. In all hydrophobicity scales the more positive the number, the more hydrophobic the residue, i.e. the less polar it is; the converse holds in hydrophilicity (philos=love) scales. Hydrophobic, that is apolar, residues tend to lie buried in the interior of a protein, while hydrophilic residues tend toward a surface. Correspondingly, in membrane-associated proteins, those residues in contact with the lipid bilayer tend toward strong hydrophobicity.

The pattern of hydrophobic and -philic residues in a protein can often reveal aspects of protein structure. The most common structures hypothesized in this manner are membrane-spanning alpha helices. To search for this type of helix, window sizes of nineteen to twenty one should be used, since about twenty amino acids are required to span the membrane in a typical alpha helix.

A powerful approach looks at the hydrophobic periodicity in regular secondary structures. Such information can often be seen best with helical wheel diagrams where the view down the helical axis shows groupings of similar kinds of amino acids. The regular appearance of apolar residues spaced three or four residues apart, with a seven residue periodicity, could be a pattern indicative of alpha helices, while sheets might show uniformly apolar sections, if completely buried within a protein, or alternating polar and apolar residues, if on the surface. Many proteins have been shown to display these patterns. Such studies have resulted in the prediction scheme of Lim (1974), and in Eisenberg's (1984) hydrophobic moment technique that quantifies this phenomenon using vector mathematics.

T-cell antigenicity as a function of amphiphilicity

Amphiphilic secondary structures have one face hydrophobic and one face hydrophilic. These structures, therefore, have high hydrophobic moments. T-cell antigenicity correlates very highly with amphiphilic alpha helices, especially those present after partial cleavage and/or unfolding, so these regions are also known as amphipathic. Margalit et al. (1987) used this correlation to predict the location of possible T-cell antigenic sites by looking for potential amphipathic helices with a high hydrophobic moment.

Something to remember in all hydrophobic moment analyses is, in general, the methods do not predict the presence or absence of a given structural element. Rather, they attempt to answer the question: If this sequence region of this protein happens to be folded into this particular conformation in nature, either alpha helix or beta sheet, then how are the hydrophobicities of its constituent residues organized? Are they randomly distributed about the structure or do they segregate about it in an organized fashion?

Higher order approaches

Others have looked at all the possible structural conformations for various sequence sections that exist in known structures, and have tried to form prediction schemes based on their findings. The thought is that a similar sequence will have similar secondary structures wherever it is found. To do this, a measure of similarity must be established between the studied sequences and the possible conformations weighted to form a final prediction. The early algorithms of Nishikawa and Ooi (1986), Levin et al. (1986), and Sweet (1986) are all based on this theme. The differences result from the comparison choices made and the scoring systems used. Later refinements of this type of approach have led to current threading techniques (see e.g. Threader by Jones et al., 1992) for the prediction of supersecondary structure.

The formation of a peptide sequence into a helix, a sheet, or a turn primarily depends on the preferred conformations of the constituent residues and the packing quality of the surface formed, though long-range interactions also play a significant role and are almost impossible to model into any prediction algorithm. And chaperonins complicate matters further. Regardless, dozens of prediction schemes have been devised over the years based on only local or semi-local sequence patterns, though all have had only limited success. In spite of all this research, and all of these advances in our knowledge of protein structure, once past these generalities, the detailed mechanisms of folding is only vaguely understood, though modern computational techniques including neural-net and artificial intelligence approaches have considerably increased the reliability of these types of predictions.

Even as the body of determined protein structures grows, questions remain as to what the relationship is between solved crystal structures and proteins in solution in life. What roles do chaperonins play *in vivo*? What effect do ionic conditions have on secondary structure? What effect does protein concentration have? Do crystals with different space groups produce the same or similar protein structures? Do X-ray and NMR structure determinations on the same protein agree with one another? When they don't, why not?

Prediction Reliability: *don't believe everything your computer tells you!*

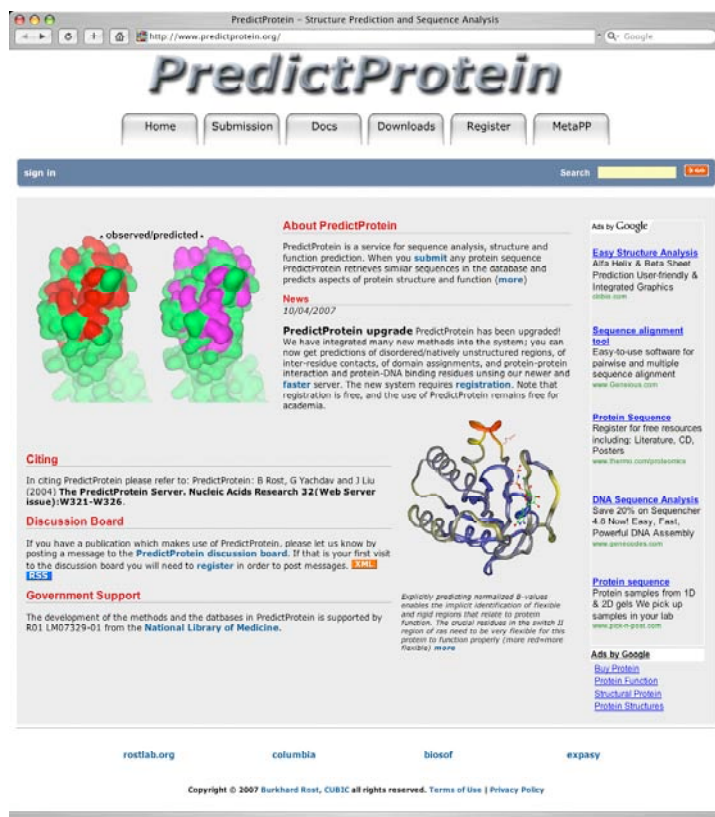
As the previous discussion should make apparent, protein secondary structure inference is fraught with difficulties. Its reliability is incredibly disheartening. Depending on whether three or four secondary structural elements are specified, random chance would result in either a 33% or a 25% chance of a prediction being correct. Most of the different approaches touched on here only improve those chances to between 45% and 55% of the prediction being correct. Reported higher percentages are often the result of a biased data set, not an actual improvement in technique. However, advances in the mid-90's, such as PredictProtein (Rost and Sander, 1993 and 1994) combine neural net technology with the strength of multiple sequence analysis to improve reliability up to and beyond 70% in many situations.

One of the more important things to realize is many of the algorithms are based on soluble, globular proteins. Using the same parameters with all types of proteins is not appropriate; therefore, when dealing with other

types of proteins you must alter parameters and interpret the results in this light. Since defaults are often based on the soluble type guidelines, you must be especially careful when working with membrane-associated or membrane-spanning proteins. The simplest parameter to change is often the window size. It should be set to approximately the size of the feature being analyzed (e.g., use a window size of about 21 when trying to find membrane-spanning alpha helices).

Using comparative multiple sequence approaches is by far the most reliable strategy. In my opinion, the best predictor of secondary structure around, available on the World Wide Web at <http://www.predictprotein.org/>, uses multiple sequence alignment profile techniques along with neural net technology. PredictProtein was originally developed by the Protein Design Group at the European Molecular Biology Laboratory, Heidelberg, Germany. A multiple sequence alignment is created with the MaxHom weighted dynamic programming method (Schneider, 1991) and a secondary structure prediction is produced by the profile network method (PHD). PHD is rated at an expected 70.2% average accuracy for the three states helix, strand, and loop (Rost and Sander, 1993 and 1994).

Their Web page provides default, advanced, and expert submission forms. One powerful advanced and expert option is to submit your own multiple sequence alignment. You'll be doing that in today's tutorial. Their automated search and alignment procedure is very good, but if you've been working for months on a multiple alignment, and you know it is the best it can be, you may want to force PredictProtein to use that information, rather than its own automated alignment. The welcome page presents a wealth of informational links and detailed instructions. It's shown here to the right:



Users of structure prediction schemes must be cautious in the application and interpretation of their results. It is best to use these predictions only in cases where other types of potentially confirming, experimental evidence is available, such as the presence of antibody producing epitopes, or estimates derived from physical data. In all cases the computer must be thought of as a tool only; experimental evidence should always be used to corroborate.

Reviewing data generated by various techniques

Often comparing all of the methods used simultaneously can be a big help. You also should have realized this with gene finding techniques. Annotating an RSF file in SeqLab, or merely creating a text-based file, can facilitate that. This way you can readily see where various methods agree or disagree by looking up and down the columns, just like in a multiple sequence alignment. As in the gene finding tutorial, the more data that you can supply, the more easily the problem is resolved.

X-ray PDB data can be interpreted in many different ways. The secondary structure assignments made by the author of the structure may not agree with assignments made via programs using the same coordinate data as input. Even the assignments made by computer software will vary. Actual X-ray data is a guide to, not a final confirmation for, the secondary structural elements of any given protein. The actual starting and ending points of these structural units are often subject to conjecture and may be somewhat subjective.

A comparison of four insulin-like growth factor II (IGF2) sequences and several secondary structure analysis methods follows below. Headings used in this example include data from the following sources:

- GCG PeptideStructure provides secondary structure estimates by both the Chou-Fasman and the and GCG PepPlot Garnier-Robson methods. Additional information can be obtained from the antigenic index (A.I.) values to determine possible surface conformations.
- Amphi estimates T-cell antigenicity through the prediction of amphiphilic helices.
- GCG HelicalWheel allows subjective testing of amphiphilic regions.
- PDB data provides structural determinations made by the author of the PDB structure.

The PDB entry is a model of the human mature form only, the Swiss-Protein sequence is the human precursor protein, the GenBank entry is a translation based on the reference CDS information from human entry HumIGF2g, and the Profile consensus is based on the conserved portion of a multiple sequence alignment of several unique IGF2 protein entries. Following each block of the sequence alignment are the predicted and modeled secondary structural elements of the protein using H's to represent helices, B's for beta sheets, T's for turns, and x's for symbolizing the presence of A.I. peaks (upper case versus lower case is used in Chou-Fasman's scheme for strong and weak predictions respectively). Other prediction methods should also be incorporated into such a comparison to increase its power. The SeqLab editor could be used to assemble the same sort of comparison. As in most forms of computational molecular biology analysis, the more data that you can synthesize together, the more accurate will be the interpretation.

Insulin-Like Growth Factor II: secondary structure comparisons

```
PDB                                AYRPSETLCGGELVDTLQFVCGDRGFYF...SRPAS  33
SwissPro MGIPMGKSMVLVLLTFLAFASCCIAAYRPSETLCGGELVDTLQFVCGDRGFYF...SRPAS  57
GenBank  MGIPMGKSMVLVLLTFLAFASCCIAAYRPSETLCGGELVDTLQFVCGDRGFYF...SRPAS  57
Profile                                AYRPSETLCGGELVDTLQFVCGDRGFYFRLPSRPSS  36
```


window, and always close X Windows when you are through with them to conserve system memory. Furthermore, to activate X items, just <click> on them, rather than holding your mouse button down. Also, X buttons are turned on when they are pushed in and shaded. Finally, don't close X Windows with the X-server software's close icon in the upper right- or left-hand window corner, rather, always, if available, use the window's own "File" menu "Exit" choice, or "Close," or "Cancel," or "OK" button.

A 'Real-Life' Project Oriented Approach: Protein Sequence Attributes

Activate and log on to the computing workstation you are sitting at and then log onto HPC with an X-tunneled ssh session. When using an X-aware terminal window on Mac OSX or UNIX/Linux issue the following command (the X has to be capitalized and replace "user" with your account name):

```
$ ssh -X user@mendel.scs.fsu.edu
```

Preliminary preparations

Change your directory from 'home' to last week's subdirectory. List that directory and check out the files left over from last week's tutorial. Look through them and remove any that you don't want to save. Next, change directory back to your home directory, create a subdirectory for this week's tutorial data, and then change directory into it.

After you've taken care of these file maintenance chores launch SeqLab with the standard command:

```
$ seqlab &
```

Next, it will again be helpful to change your SeqLab working directory to your present location so that everything you do today will automatically be saved in your new directory rather than last week's directory. Do this as before with SeqLab's "Options" "Preferences. . ." "Working Dir. . ." button.

Now verify that you are in SeqLab's "Main List" "Mode:" and start a new list to contain this week's data. Therefore, select "New List. . ." from the "File" menu and give your new list an appropriate name. It's not essential to use the file name extension ".list" but it's a good idea. Check "OK."

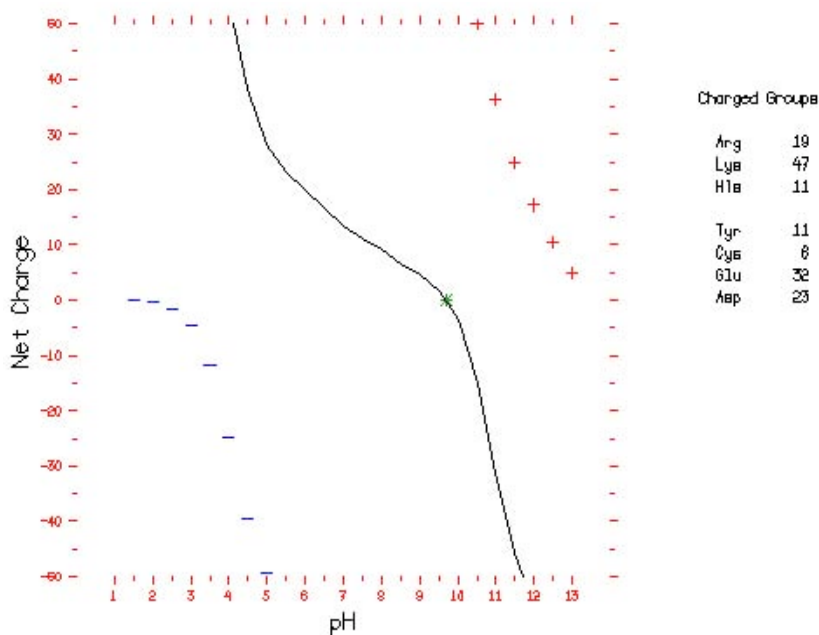
You should now be in List Mode with an empty window. Go to the "File" menu and select "Add Sequences From" "Sequence Files. . ." Use the "Directories" column to move from your present directory over to Lab Seven's subdirectory, and then replace the text in the "Filter" text box with the name or a wildcard specification that will identify your final protein dataset RSF file from that tutorial. This should be an aligned dataset containing about fifty or so of your selected project molecular system protein sequences, annotated with database and program-generated feature descriptions. Press the "Filter" button and select the correct entry. Press the "Add" button to add it into your new empty list file and then "Close" the "Add Sequences" window. Use the "File" menu to "Save" your new list. Select the RSF file and switch "Mode:" to "Editor." Select one of the sequences in your alignment that has not had its three-dimensional structure solved (as described in its database annotation, and as visible by secondary structure annotation when displayed in

graphic mode) and that is one of the least similar sequences in the dataset to your original query sequence. Press the “**COPY**” button and “**PASTE**” a copy of the sequence at the bottom of your alignment. Be sure that just the new copy of your sequence is selected and then go to the “**Edit**” menu and choose “**Remove Gaps** . . .” “**All gaps**.” You’ll need to work on this ungapped sequence for most of this tutorial. This is because some of the protein analysis programs will not work with gaps in their input. Do not get rid of the rest of the alignment though. We’ll need it later on today as well.

Physical characteristics/protein mapping

The GCG programs PeptideMap, PeptideSort, and Isoelectric enable you to generate protease digestion mapping data, molecular weight and amino acid composition information, and HPLC retention and isoelectric point values, as well as the molar extinction coefficient at 280 nm. All results can be experimentally verified and often may assist in experimental design. All three no longer run under CentOS 5. EMBOSS provides some alternatives that I’ve written into the SeqLab “**Extensions**” “**EMBOSS programs**” “**Protein analysis**” menu: **Digest**, **PepStats**, and **IEP**. Read about these and all the subsequent EMBOSS programs used today through <http://emboss.sourceforge.net/apps/release/5.0/emboss/apps/index.html>. Run through these three programs using default parameters in SeqLab on your ungapped selected sequence from above. They are very fast and easy to use, and may prove useful in your own labs. Rerun them a few times with various combinations of parameters to get a better feel for how they operate.

Be aware that Isoelectric cannot take into consideration the folded shape of the protein, and any electrostatic interactions within the protein caused by that shape, so it’s calculations should be considered appropriate for the denatured, not native, protein. The GCG Isoelectric program plot on EF-1 α from Brine Shrimp produced the adjacent plot:



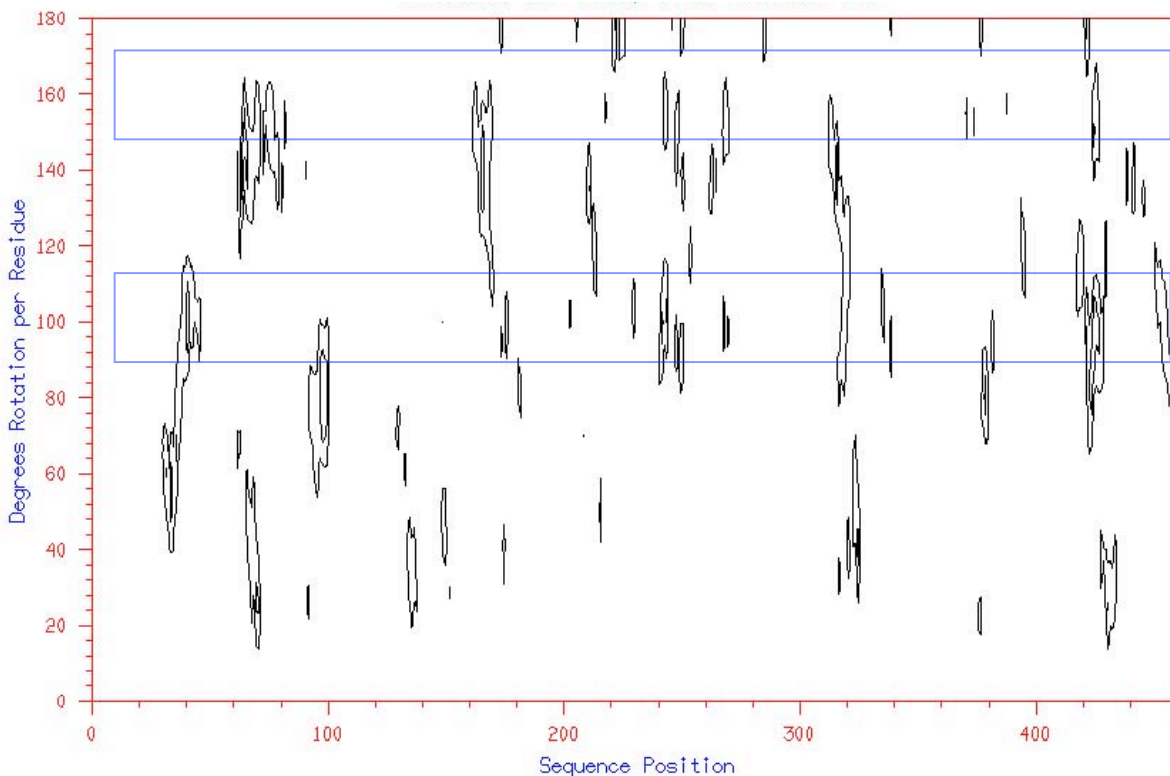
Hydrophobic moment

The helical hydrophobic moment, as described by David Eisenberg, quantitatively shows how asymmetrically distributed residue hydrophobicities are, by using vector mathematics. This value, calculated with an appropriate window size, can often help you identify ‘amphiphilic’ structures. As described in the Introduction,

these are alpha helices or beta sheets with one polar and one apolar face. This type of structure is often found in membrane channels, with several amphiphilic secondary structural elements clustered together, their hydrophilic faces toward the middle aqueous channel and their hydrophobic surfaces in contact with the membrane. Amphiphilic structures are also commonly found on the surface of globular protein domains. These have their hydrophilic face exposed to the solvent and their hydrophobic face interacting with the rest of the protein. Membrane associated proteins may also possess amphiphilic structures, with their polar face interacting with the mass of the protein and their apolar face in tight association with the lipid membrane, though their moment peaks will likely not be nearly as striking as the previous two types.

GCG plots the hydrophobic moment of a protein with the Moment program. It's another broken one under CentOS 5, so we'll use EMBOSS's HMoment instead. (Also see entirely different approaches with EMBOSS's TMap and Octanol programs.) Run "**HMoment**" on your sequence from the "**Extensions**" "**EMBOSS programs**" "**Protein Analysis**" menu. Run plots at typical helical and sheet conformations.

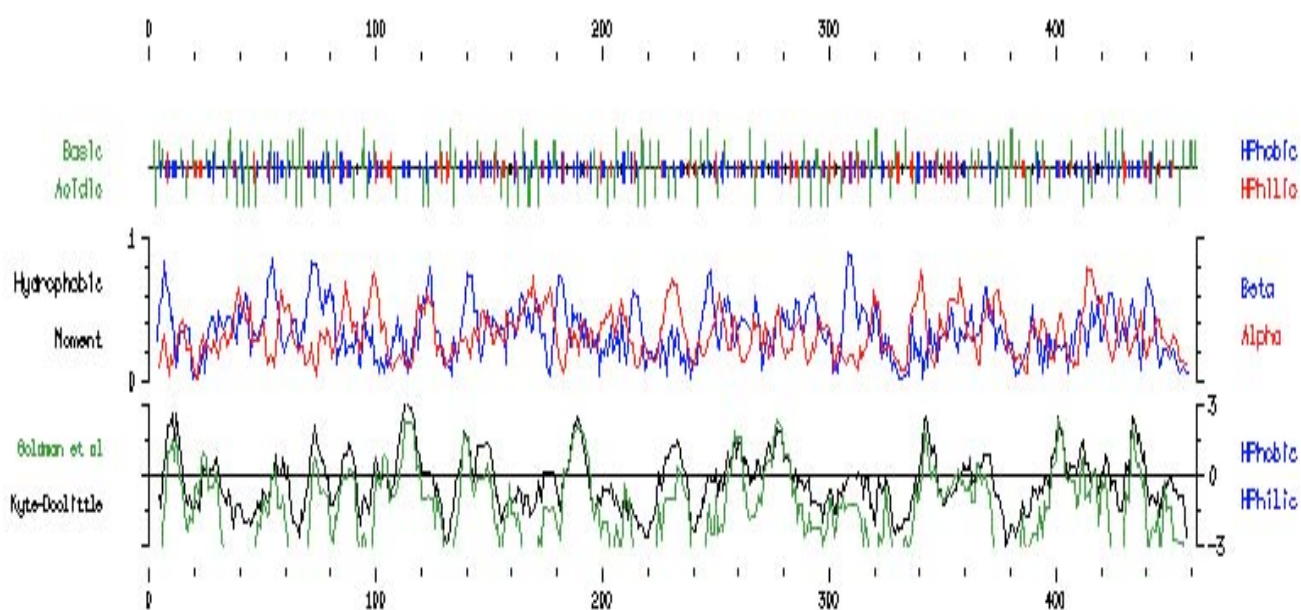
The GCG 3D contour plot is confusing with all angles of rotation drawn on the same plot, just remember that residues are offset from each other in a typical alpha helix by 100° and in a typical beta strand by 160° . The EMBOSS 2D plots are a bit easier to understand, with separate 100° and 160° plots. Take notes of the sequence location of any particularly striking moment peaks in your selected protein molecule. My example GCG Moment plot on the Brine Shrimp EF-1 α molecule follows below. I've placed blue open boxes along the two appropriate angles of rotation. Looks like if there are alpha helices around positions 40, 100, 180, 240, 320, or 420, there's a strong possibility they're quite amphiphilic:



Secondary structure prediction programs — combination approaches

GCG's "PepPlot" can produce up to nine different graphical panels displaying various secondary structure and physical attributes, including hydrophobicity, but it unfortunately cannot create RSF output, so it is unable to add feature annotation to an RSF file. It's also broken under CentOS 5 so we won't be running it anyway.

Hydrophobicity-based analyses are very worthwhile, nonetheless. Be sure you understand the difference between this type of analysis and the hydrophobic moment described previously. Window settings are very important in all hydrophobicity analyses. Since all of our project molecules are relatively small, globular proteins, a window size of seven or nine would be good to look for interior versus exterior parts of the molecule. A sample GCG PepPlot graphic using Brine Shrimp EF-1 α is illustrated below:



Run the comparable Kyte-Doolittle hydrophathy plot in the EMBOSS system with their "**PepWindow**" program. Find this under the SeqLab "**Extensions**" "**EMBOSS programs**" "**Protein Analysis**" menu.

GCG's PeptideStructure/PlotStructure program pair is another combination approach. It makes secondary structure, B-cell antigenicity, flexibility, and surface probability predictions, as well as a hydrophilicity determination, and graphically displays these predictions. The program is optimized for soluble, globular proteins, and it can create RSF output with this program — this can be very helpful for adding annotation to your dataset. However, it's another broken one under CentOS 5. Therefore, we'll use two different EMBOSS programs to see a couple of the attributes that GCG's PeptideStructure/PlotStructure display — antigenicity and Garnier-Robson secondary structure predictions. Run EMBOSS's "**Antigenic**" and "**Garnier**," both off the "**Extensions**" "**EMBOSS programs**" "**Protein Analysis**" menu. Accept the program defaults with both.

Garnier-Robson four-state secondary structure predictions are listed in the output from Garnier. However, remember from the Introduction that Garnier-Robson predictions use a very old and somewhat unreliable algorithm, so don't put a whole lot of faith in it! We'll see how well it compares to PHD later.

EMBOSS's Antigenic predicts epitopes according to the method of Kolaskar and Tongaonkar (1990). This antigenic index, as opposed to the amphipathicity scale based on hydrophobicity distribution (hydrophobic moment) discussed previously, is based on the surface prediction of the hydrophobic residues Cys, Leu and Val, rather than the prediction of amphipathic helices. In other words, it looks for loose, 'floppy' portions of the molecule sticking out from the surface of the protein. As such, it attempts to predict all major immunogenic determining sites, especially those associated with B-cell humoral response epitopes, rather than T-cell. And, even though part of its calculation relies on unreliable surface predictions, overall it does a pretty decent job of estimating where the antigenic regions lay in your sequence.

Secondary structural information in PDB files

All of the project molecules for this laboratory course have had at least one, and in most cases many, homologues' structures experimentally determined; therefore, PDB access codes are available for them. You may know this access code from preliminary database perusals in Lab Two. You may have even explored visualizing your project molecule's structure with Entrez's Cn3D or the Protein Data Bank's KING viewer in that lab. PDB files contain a lot of information beyond just the atom's three-dimensional coordinates. You can read this annotation within Entrez by using the "Structure Info" button. And, if you have the PDB file, you can just directly read it from the file. The structure of these files is a bit cryptic though. Every line is identified with a code word that describes what type of information is on that particular line. Having some familiarity with those code words can be quite helpful, if you'll ever be using them directly.

PDB file subject codes

A listing of common PDB file subject area codes helpful to sequence and secondary structure analysis follows. Knowing these codes makes finding relevant information in a PDB file much easier.

HEADER	type of the material studied
COMPND	name of the material studied
SOURCE	source of the material used for the crystal
AUTHOR	who did the work
REMARK	comments on the crystallization or refinement process, references, or data changes
SEQRES	the amino acid sequence of the material studied in three-letter code
HET	the names of nonpeptide units in the structure other than water
HELIX	helical assignments within the structure and their type
SHEET	sheet assignments within the structure and their type
TURN	turn assignments within the structure and their type

SSBOND the location of disulfide linkages in the structure

Independent of PDB, the UniProt database also lists secondary structure annotation for those molecules whose structures have been solved, so it can, therefore, also be used to easily find the known secondary structure assignments of a protein. Find a UniProt entry in your RSF file that has a solved structure. It should be an entry with red zigzag and green box graphical annotation. GCG uses red zigzags to annotate helix and green boxes for strand secondary structures in UniProt entries. Double-click the entry's name (or click the "INFO" icon) to read the author's secondary structure assignment. Compare the secondary structure information in it to what you discovered with its homologue using all the above predictive analyses. Take notes of your findings, they'll form the basis of the Report Form. A warning needs to be heeded though — the aligned dataset has gaps in it, this produces a numbering discrepancy between it and the sequence you analyzed. Therefore, if using the cursor to locate elements onscreen, pay attention to the "Position:" not the "Column:" indicator in SeqLab for the location of the secondary elements in the solved structure.

Graphical verification of amphiphilic helices

After other analyses have finished, GCG's HelicalWheel (also broken under CentOS 5) or EMBOSS's PepWheel can be used on those areas of the molecule that actually are, or show potential of being, a helix or a sheet. These programs draw a segment of your protein in a 'wheel' fashion, like you were looking down the axis of a helix. Examining the results of helical wheel analysis on those areas of a predicted or of a known secondary structure element enables you to evaluate any asymmetrical ordering of the hydrophobicity pattern of the residues within that element, i.e. its hydrophobic moment, and hence, its amphiphilicity, and it can often contribute to understanding potential packing patterns of the protein, both with itself, and with any ligands and/or cell membranes that protein may interact with. It is by far the easiest way to visualize this phenomenon.

In the previous section you should have been able to infer just where the potential helical regions of your protein most likely lay by using the actual secondary assignments of one of its homologue's solved structure. Be sure that you took any potential numbering discrepancy between the two sequences into consideration. Repeat with the sheet specifications if you want, or if there are no alpha helices in your protein.

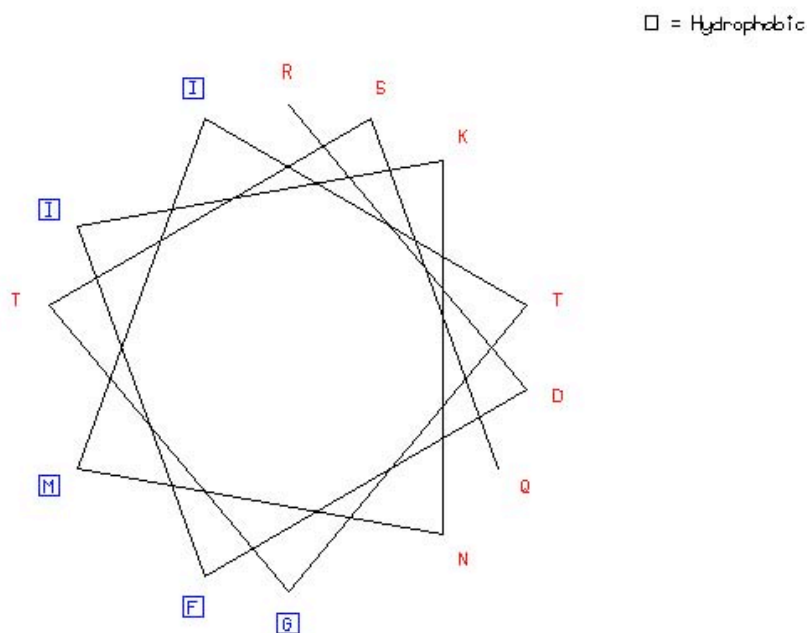
Run "PepWheel" on each alpha helical region identified, and note any ordering of the helical surfaces shown, i.e. the clustering of polar or nonpolar amino acids on one side of the wheel. Launch "PepWheel" off the "Extensions" "EMBOSS programs" "Protein Analysis" menu. Specify each individual region within the PepWheel main program window (the sbegin1 and send1 parameters) and run it for each region separately. Record your findings for the various helices tested.

If you want to test your sheet regions, or if it looks like your molecule doesn't have any helices at all because none of its homologues with solved structures do, perform the analyses with changed "Steps" and "Turns" parameters according to the table on the following page:

<u>Helix</u>	<u>phi</u>	<u>psi</u>	<u>omega</u>	<u>res/turn</u>	<u>transl</u>	<u>turns</u>	<u>steps</u>
Alpha	-57	-47	180	3.6	1.50	5	18
3-10	-49	-26	180	3.0	2.00	1	3
pi	-57	-70	180	4.4	1.15	5	22
PP I	-83	158	0	3.33	1.9	3	10
PP II	-78	149	180	3.0	3.12	1	3
PG II	-80	150	180	3.0	3.1	1	3
anti Beta	-139	135	-178	2.0	3.4	4	9
para Beta	-119	113	180	2.0	3.2	4	9

Note: PP is polyproline and PG II is polyglycine.

The best amphiphilic helix in my EF-1 α example is shown below. It ran from residue position 95 through 107 and corresponds to hydrophobic moment peaks in both the GCG Moment and GCG PepPlot programs:



Specialized protein analysis programs in SeqLab

Several other GCG programs in the “**Protein Analysis**” section of the “**Functions**” menu have very specific purposes. “**TransMem**” predicts transmembrane alpha helices based on a hidden Markov Model of what transmembrane alpha helices and loop regions ‘feel’ like (Sonnhammer et al., 1998). It works remarkably well. None of our project molecules are membrane bound, but run yours through this program nonetheless, just to see what happens. Note that a RSF file output option is available. Remember that this is how you can build up the annotation of your RSF data automatically. “**HTHScan**” uses a log-odds PSSM (position specific scoring matrix) profile of three different helix-turn-helix (H-T-H) motifs, to look for these elements in your sequence. The three motifs, the araC and lysR families of H-T-H, and the homeobox domain, are all indicative of gene regulatory DNA binding structures. HTHScan can also produce RSF output. See if your sequence has any predicted H-T-H domains. “**SPScan**” scans proteins for the presence of secretory signal peptide sequences using an enhanced PSSM approach. SPScan uses a combination of the von Heijne

(1987) and McGeoch (1985) methods. Be sure to use the appropriate “**Eukaryote**” weight matrix. It can also produce RSF output. The final specialized purpose GCG protein analysis program is “**CoilScan**.” It looks for coiled-coil regions in protein sequences, again using a PSSM method (Lupas, 1996), and can also produce RSF output as well as a plot. The prediction is only valid for solvent exposed coiled-coils, particularly for parallel and anti-parallel two-stranded coiled-coils, and for parallel three-stranded coiled-coils. Even though none of our project molecules contain coiled-coils, go ahead and run your sequence through CoilScan.

You may have been surprised to have found some of these structures predicted to occur in your sequence, even though by now you should be familiar enough with your project molecule system to know what should and should not be in it. This only reinforces the notion that false positives are still a major problem in sequence analysis. It’s not easy!

Exit SeqLab with the “**File**” menu “**Exit**” choice and save your RSF file and any changes in your list with appropriate responses. Accept the suggested changes and designate names that make sense to you; SeqLab will close. Log out of your current UNIX session on HPC and on your workstation.

Internet secondary structure predictions

The PredictProtein Internet service by the Protein Design Group at the European Molecular Biology Laboratory, Heidelberg, Germany was mentioned in the Introduction. A multiple sequence alignment is performed by a weighted dynamic programming method (MaxHom, Schneider, 1991) and a secondary structure prediction is produced by the profile network method (PHD). The PHD secondary structure prediction method is rated at an expected 70%+ average accuracy for the three states helix, strand, and loop (Rost and Sander, 1993 and 1994).

NNPredict is a service of the San Francisco campus of the University of California that uses neural net technology to predict protein secondary structure (Kneller, et al., 1990). The basis of the prediction is a two-layer, feed-forward neural network. By adding neural network units that detect periodicities in the input sequence, they have modestly increased the secondary structure prediction accuracy. The use of predetermined tertiary structural classification causes a marked increase in accuracy. The best-case prediction was 79% for the class of all-alpha proteins.

Both servers are available over the World Wide Web, at:

<http://www.predictprotein.org/> and
<http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html> respectively.

Several other protein secondary and even tertiary structure prediction Web servers are available; many are bookmarked through the Protein Analysis servers link (compiled by Susan J. Johns, UCSF):

<http://bio.fsu.edu/~stevet/ProteinAnalysis.html> and Baylor has a great Web portal:
<http://searchlauncher.bcm.tmc.edu/seq-search/struc-predict.html>.

Homework assignment

For the Report Form figure out how to submit your entire alignment to PredictProtein, and report your success. You'll need to use its advanced mode, not its default, and you'll have to convert your RSF file to an alternative format for them to accept it. You'll probably want to 'scp' your file from HPC to another computer too. Other than these hints you're on your own. Describe how good of a job it does with discovering the correct secondary structure elements of your dataset. If you can't get PredictProtein to accept your alignment, just submit the single sequence that you used for the rest of the tutorial. The Report Form will also ask you to evaluate many of the GCG programs you ran in the tutorial today.

Conclusion: caveats and considerations

You've now gone all the way from probe design, through contig assembly, database similarity searching, multiple sequence alignment, and phylogenetic inference, up to secondary structure prediction, all based on molecular sequence data. As you can see the further we get into theoretical realms, the more loosely we have to entertain the results — reality and predictions don't always quite match. Oftentimes the resultant predictive data derived from sequence analysis will directly conflict with the known structural data, but methods also sometimes agree. Newly sequenced genes often have no structural information available; we must try and use whatever is available, always keeping in mind the reliability of the methods. If your protein is very similar to another protein, as identified by searching algorithms, and belongs to a distinct family, then many parallels may be drawn. In fact, even three-dimensional modeling is sometimes possible. This is known as homology modeling and is the topic of next week's lab tutorial.

References

- Chothia, C. (1984) Principles that Determine the Structure of Proteins. *Annual Review of Biochemistry* **53**, 537–572.
- Chow, P.Y. and Fasman, G.D. (1974) Prediction of Protein Conformation. *Biochemistry* **13**, 222–245.
- Eisenberg, D.M., Weiss, R.M., and Terwilliger, T.C. (1984) The Hydrophobic Moment Detects Periodicity in Protein Hydrophobicity. *Proceeds of the National Academy of the Sciences, U.S.A.* **81**, 140–144.
- Garnier, J., Osguthorpe, D.J., and Robson, B. (1978) Analysis of the Accuracy and Implications of Simple Methods for Predicting the Secondary Structure of Globular Proteins. *Journal of Molecular Biology* **120**, 97–120.
- Genetics Computer Group (GCG®), (Copyright 1982-2006) *Program Manual for the Wisconsin Package®*, version 11.0, <http://www.accelrys.com/products/gcg/> Accelrys Inc., San Diego, California, U.S.A.
- Goldman, Engleman, and Steitz (GES) (reviewed in *Ann. Rev. Biophys. Biophys. Chem.* **15**, 321–353 [1986])

- Jones, D.T., Taylor, W.R., and Thornton, J.M. (1992) A New Approach to Protein Fold Recognition. *Nature* **358**, 86-89.
- Kolaskar, A.S. and Tongaonkar, P.C. (1990). A Semi-Empirical Method for Prediction of Antigenic Determinants on Protein Antigens. *FEBS Letters* **276**, 172–174.
- Kneller, D.G., Cohen, F.E., and Langridge, R. (1990) Improvements in Protein Secondary Structure Prediction by an Enhanced Neural Network. *Journal of Molecular Biology* **214**, 171–182.
- Kyte, J. and Doolittle, R.F. (1982) A Simple Method for Displaying the Hydrophobic Character of a Protein. *Journal of Molecular Biology* **157**, 105–132.
- Levin, J.M., Robson, B., and Garnier, J. (1986) An Algorithm for Secondary Structure Determination in Proteins Based on Sequence Similarity. *FEBS Letters* **205**, 303–308.
- Lim, V.I. (1974) Algorithms for the Prediction of α -Helical and β -Structural Regions in Globular Proteins. *Journal of Molecular Biology* **88**, 873–894.
- Lupas, A. (1996). Prediction and Analysis of Coiled-Coil Structures. In *Methods in Enzymology*, (R.F. Doolittle, ed.), **266**, pp 513–525, Academic Press, San Diego, California, USA.
- Margalit, H., Spouge, J.L., Cornette, J.L., Cease, K.B., Delisi, C., and Berzofsky, J.A. (1987) Prediction of Immunodominant Helper T Cell Antigenic Sites from Primary Sequence. *Journal of Immunology* **138**, 2213–2229.
- McGeoch, D. (1985). On the Predictive Recognition of Signal Peptide Sequences. *Virus Research* **3**, 271–286.
- Nishikawa, K. and Ooi, T. (1986) Amino Acid Sequence Homology Applied to the Prediction of Protein Secondary Structures, and Joint Prediction with Existing Methods. *Biochimica Biophysica Acta* **871**, 45–54.
- Rost, B. and Sander, C. (1993) Prediction of Protein Secondary Structure at Better than 70% Accuracy. *Journal of Molecular Biology* **232**, 584–599.
- Rost, B. and Sander, C. (1994) Combining Evolutionary Information and Neural Networks to Predict Protein Secondary Structure. *Proteins* **19**, 55–77.
- Sander, C. and Schneider, R. (1991) Database of Homology-Derived Structures and the Structural Meaning of Sequence Alignment. *Proteins* **9**, 56–68.
- Sonnhammer, E.L., von Heijne, G., and Krogh, A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology* **6**, 175–182.

Sweet, R.M. (1986) Evolutionary Similarity Among Peptide Segments is a Basis for Prediction of Protein Folding. *Biopolymers* **25**, 1565–1577.

von Heijne, G. (1987) *Sequence Analysis in Molecular Biology; Treasure Trove or Trivial Pursuit*. Academic Press, Inc., San Diego, California, U.S.A.

Structure-related databases at EMBL

3D_Ali: Found in /pub/databases/3d_ali. A documentation file is included, 3d_ali.doc.

Pascarella, S. and Argos, P. (1992) A data bank merging related protein structures and sequences. *Protein Engineering* **5**, 121–137.

FSSP: The directory /pub/databases/fssp contains the database of families of structurally similar proteins for each of 154 representative protein chains (below 30 % sequence identity).

Holm, L., Ouzounis, C., Sander, C., Tuparev, G., and Vriend, G. (1992) A database of protein structure families with common folding motifs. *Protein Science* **1**, 1691–1698.

HSSP: The directory /pub/databases/hssp contains the database of homology-derived protein structures (HSSP). There is one HSSP file for each PDB protein, as well as utility programs.

Sander, C. and Schneider, R. (1991) Database of homology-derived protein structures. *Proteins* **9**, 56–68.