

# **BSC4933/5936: Introduction to Bioinformatics**

**Laboratory Section: Tuesdays from 3:45 to 5:45 PM.**

## **Protein Sequence Attributes**

**Week Nine, Tuesday, October 21, 2003**

**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, particularly using methods based on homology inference (e.g. PredictProtein).

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

\*GCG<sup>®</sup> is the Genetics Computer Group, part of Accelrys Inc., a subsidiary of Pharmacoepia Inc.,  
producer of the Wisconsin Package<sup>®</sup> for sequence analysis.  
□ 2003 BioInfo 4U

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

### Standard disclaimer

I write these tutorials from a 'lowest-common-denominator' biologist's perspective. That is, I only assume that you have fundamental molecular biology knowledge, but are relatively inexperienced regarding computers. As a consequence of this they are written quite explicitly. Therefore, if you do exactly what is written, it will work. However, this requires two things: 1) you must read very carefully and not skim over vital steps, and 2) you mustn't take offense if you already know what I'm discussing. I'm not insulting your intelligence. This also makes the tutorials longer than otherwise necessary. Sorry.

I use three writing conventions in the tutorials, besides my casual style. I use **bold** type for those commands and keystrokes that you are to type in at your keyboard or for buttons or menus that you are to click in a GUI. I also use bold type for **section headings**. Screen traces are shown in a 'typewriter' style Courier font and "//////////" indicates abridged data. The arrow symbol (>) indicates the system prompt and should not be typed as a part of commands. Really important statements may be underlined.

As you've learned, specialized X-server graphics communications software is required to use GCG's SeqLab. I'll remind you of a few user hints while using X: X Windows are only active when the mouse cursor is in that 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.

### Introduction — protein secondary structure and physical attributes

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).

Protein secondary structure determination is 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 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.

Experimental indications of the possible secondary structures of proteins first came from initial studies on polypeptides. 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.

As more and more structures were determined, the beginnings of possible folding patterns were observed. Soluble globular proteins have started to be understood in some general terms. 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, globular proteins usually fold into a few common patterns, though many argue with such arbitrary distinctions. These proteins can roughly be grouped into four 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.

After a number of structures had been determined, various research groups attempted statistical studies to determine preferences of different individual amino acid to have given secondary structure types. 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 is an ambiguous method that has proven 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 is unambiguous and easy to automate. Both methods have been incorporated into the GCG Wisconsin Sequence Analysis Package, though I seldom trust their output.

Dichroism measures the difference in 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.

### **Hydrophobicity and amphiphilicity**

Hydrophobicity is a measure of how much a molecule hates water. 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; the converse holds in hydrophilicity scales. Hydrophobic 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 using hydrophobicity is to look for 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 Eisenberg's (1984) hydrophobic moment technique.

### **T-cell antigenicity as a function of amphiphilicity: Amphi**

Amphi (Margalit et al., 1987) can help confirm the location of possible surface helices. T-cell antigenicity has been found to correlate very highly with amphiphilic (also known as amphipathic because of this correlation) alpha helices, especially those present after partial cleavage and/or unfolding. This program attempts to determine if potential amphipathic helices exist in a sequence through hydrophobic moment analysis. Its main function, therefore, is to locate possible T-cell antigenic sites that correlate with those amphipathic helices. Given sufficient interest we will port Amphi onto Mendel. Let me know.

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 weighed 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. Dozens of prediction schemes have been devised over the years based on only local or semi-local sequence patterns but all have only relatively 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? If not, why not?

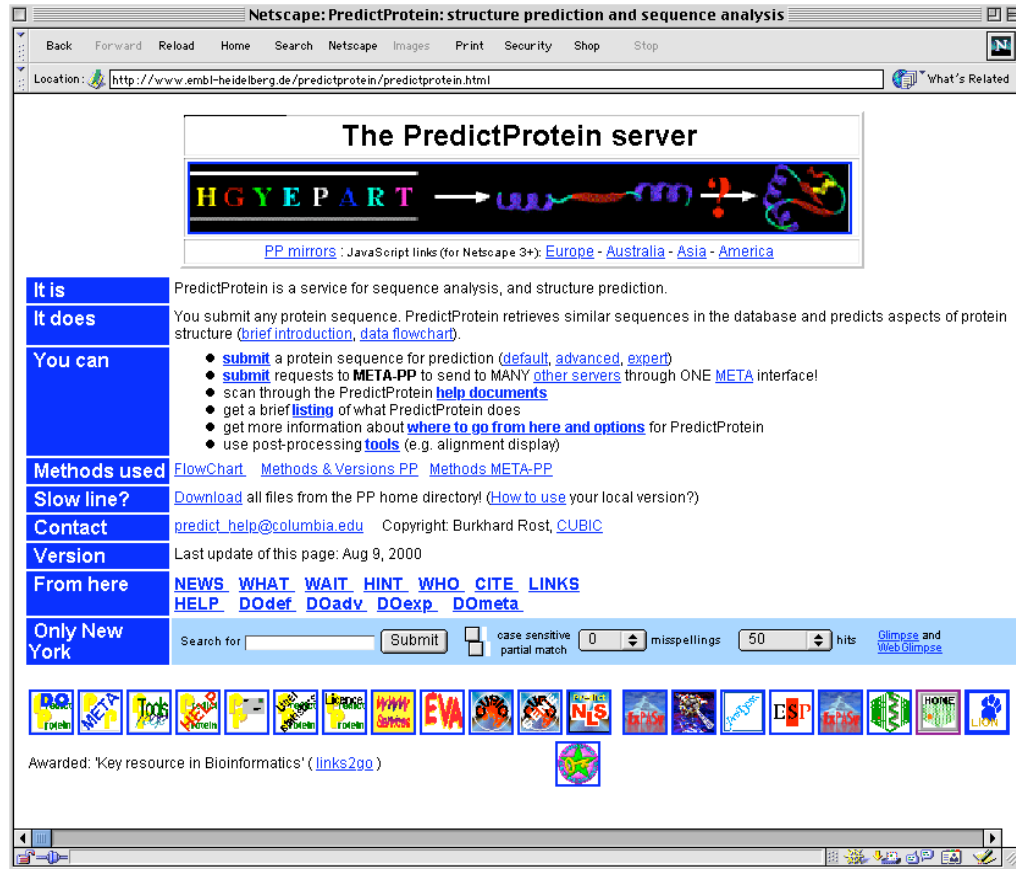
**Prediction Reliability: *don't believe everything your computer tells you!***

Structural inference is fraught with difficulties. Protein secondary structure prediction reliability is disheartening. Depending on whether three or four secondary structural elements are specified, random chance would result in either a 25% or a 33% 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, recent advances, such as PredictProtein 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; therefore, when dealing with other types of proteins you must alter parameters and interpret the results in this light. Using the same parameters with all types of proteins would not be appropriate. Since defaults are often based on the soluble type guidelines, one must be especially careful when working with membrane-associated or membrane-spanning proteins. The applicability of these parameters is vital and one must tailor them appropriately. The simplest parameter to change is often the window size. It should be set approximately to 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.embl-heidelberg.de/predictprotein/predictprotein.html> in Europe and <http://cubic.bioc.columbia.edu/predictprotein/> in North America, uses multiple sequence alignment profile techniques along with neural net technology. PredictProtein was 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. It's shown below:



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 potential 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 seen this with the gene finding techniques. Annotating an RSF file 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 structural 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 my example include:

- PeptideStructure provides secondary structure estimates by both the Chou-Fasman and the and 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.
- 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 SwissProtein 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. 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

PDB secondary structure data:
                                HHHHHHHHHHH
                                TTTTT      TTTTT
GCG CF:      TtHtHHHHHHHHHHHtBBBBB Tt tTT BBBBTTTBBB...BBttt
GCG GOR:      HHHHHHHHHHHHHHHH TTTTTTTT BBBBTTTBBBBTTTTTTT
GCG AI:      xxx x          AAAAAA          xxxx      xx
Amphi:
HelicalWheel:      HYDROPHOBIC          amphiphilic

PDB      RVSRRSR.....GIVEECCFRSCDLALLETYCATPAKSE 67
SwissPr  RVSRRSR.....GIVEECCFRSCDLALLETYCATPAKSERDVSTPPT 99
GenBank  RVSRRSRGIVEECCFRKQHSSTMPGIVEECCFRSCDLALLETYCATPAKSERDVSTPPT 117
Profile  RVNRRSR.....GIVEECCFRSCDLALLETYCATPAKSE 70

PDB secondary structure data:
                                HHHHHHHH  HHHHHH
                                TTTTT
GCG CF:      ttttthhhhhhhhhhhT TThhhhhhhhtttHHHHHH hhhhhhhhh tt

```

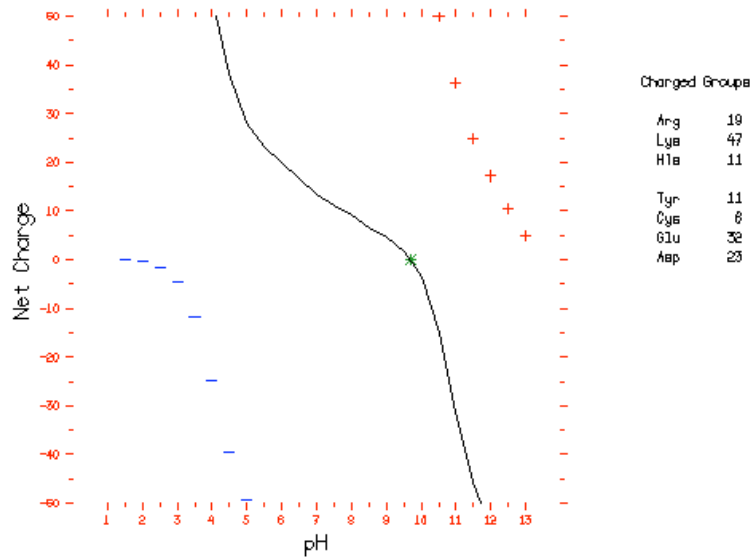


You should now be in List Mode with an empty window. Go to the **“File”** menu and select **“Add Sequences From” “Sequence Files. . .”** Use the **“Directories”** column to move from your present directory over to Week Eight’s subdirectory and then replace the text in the **“Filter”** text box with the name or a wildcard specification that will identify your RSF file used in that tutorial, created and refined in Week Seven. This should be an aligned dataset containing about twenty of your selected Project Molecular system 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. 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). 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 un-gapped sequence for most of this tutorial. This is because many of the protein analysis programs will not work with gaps in their input.

### **Physical characteristics/protein mapping: PeptideMap, PeptideSort, and Isoelectric**

These three GCG programs 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. Run through these three programs in SeqLab on your un-gapped selected sequence from above. They are very fast and easy to use, and may prove useful in your own labs. Launch them from the **“Functions” “Protein Analysis”** menu. Use any combination of parameters that you would like. Especially note that you can specify all, any, or none of the proteolytic enzymes listed in the **“Enzyme Chooser”** for the two mapping programs. 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. Isoelectric on EF-1□ from Brine Shrimp produced the following plot:

ISOELECTRIC of: input\_37.ref(Cefixime\_a1) Ck: 8540 1 to 461 March 2, 2003 17:33  
 \* = Isoelectric point = 9.72

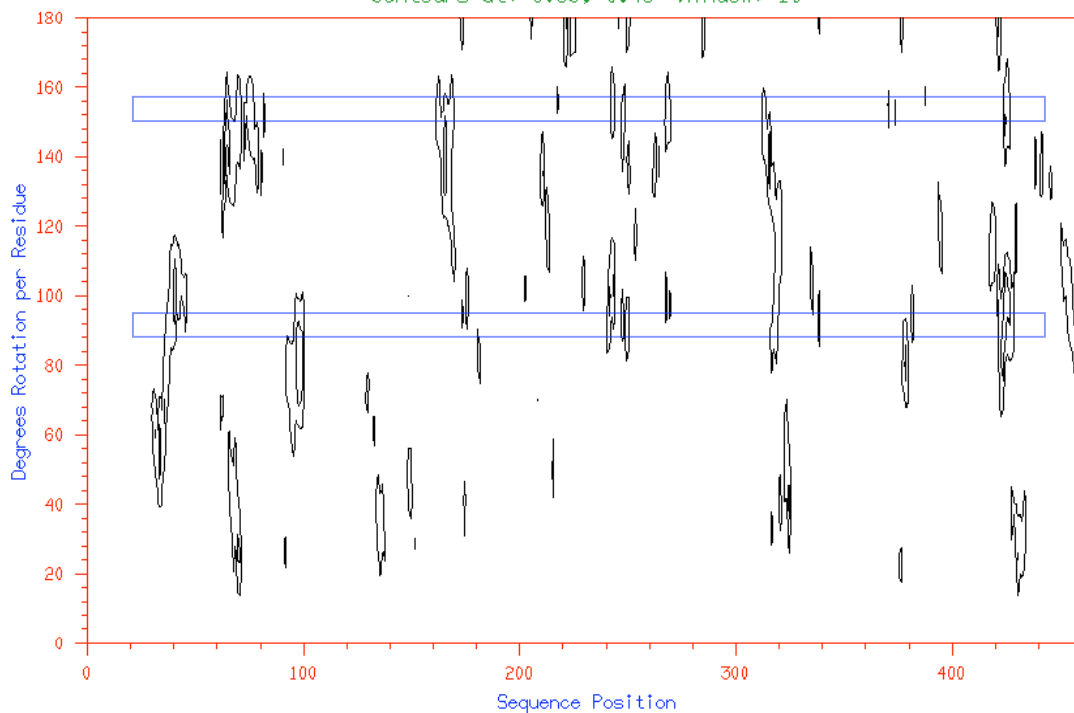


### Hydrophobic moment: Moment

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. These are alpha-helices or beta-sheets with one polar and one apolar face. These type of structures are 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. Run Moment on your sequence from the "Functions" "Protein Analysis" menu. The 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°. Take notes of the sequence location of any particularly striking moment peaks in your selected protein molecule. My example Moment alpha plot on the Brine Shrimp EF-1 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:

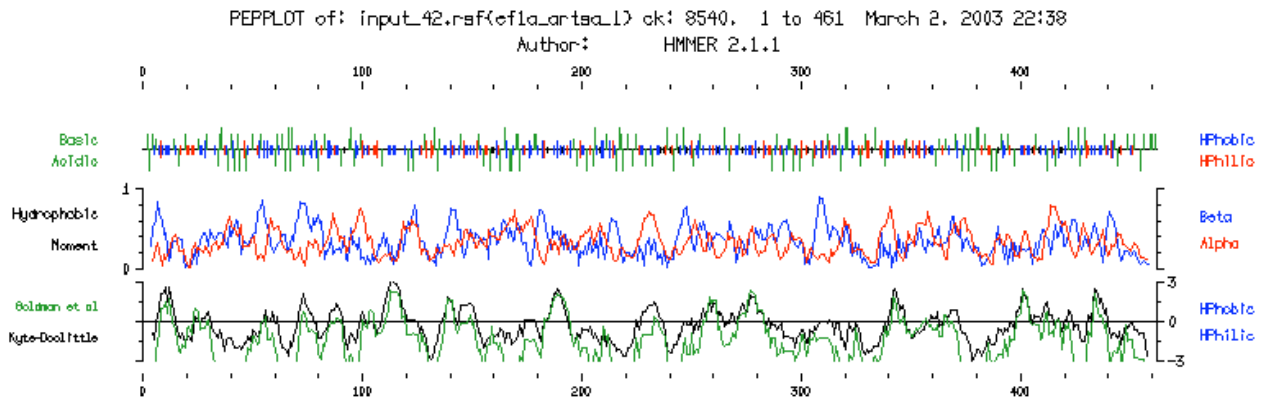
MOMENT of: input\_41.rsfc(ef1a\_antsa\_1) Ck: 8540, 1 to 461 March 2, 2003 21:01  
Contours at: 0.35, 0.45 Window: 10



## Secondary structure prediction programs — combination approaches

### PepPlot analysis

The first combination program to investigate is GCG's "**PepPlot**." It can produce up to nine different graphical panels displaying various secondary structure and physical attributes, but it cannot create RSF output. Running the program with only a few of the panel displays activated can be much more effective than showing them all. I especially like to turn off the secondary structure predictors since they are all old, unreliable algorithms. The hydrophobicity based analyses are very worthwhile, however. Two important command line options related to hydrophobicity to note are the `-HWindow` and `-GESWindow` settings. These appear as "**Window size for Kyte-Doolittle hydrophathy calculation**" and "**Window size for GES hydrophobicity calculation**" respectively in the "**Options**" window. They are set at two different values by default; this will not yield nicely congruous hydrophobicity plots. Therefore, decide on an appropriate window size for your circumstance and assign it to both window sizes. Since all of our Project Molecules are relatively small, globular proteins, a window size of seven or nine would be good. My sample PepPlot graphic using Brine Shrimp EF-1 is illustrated below:



### PeptideStructure/PlotStructure analysis

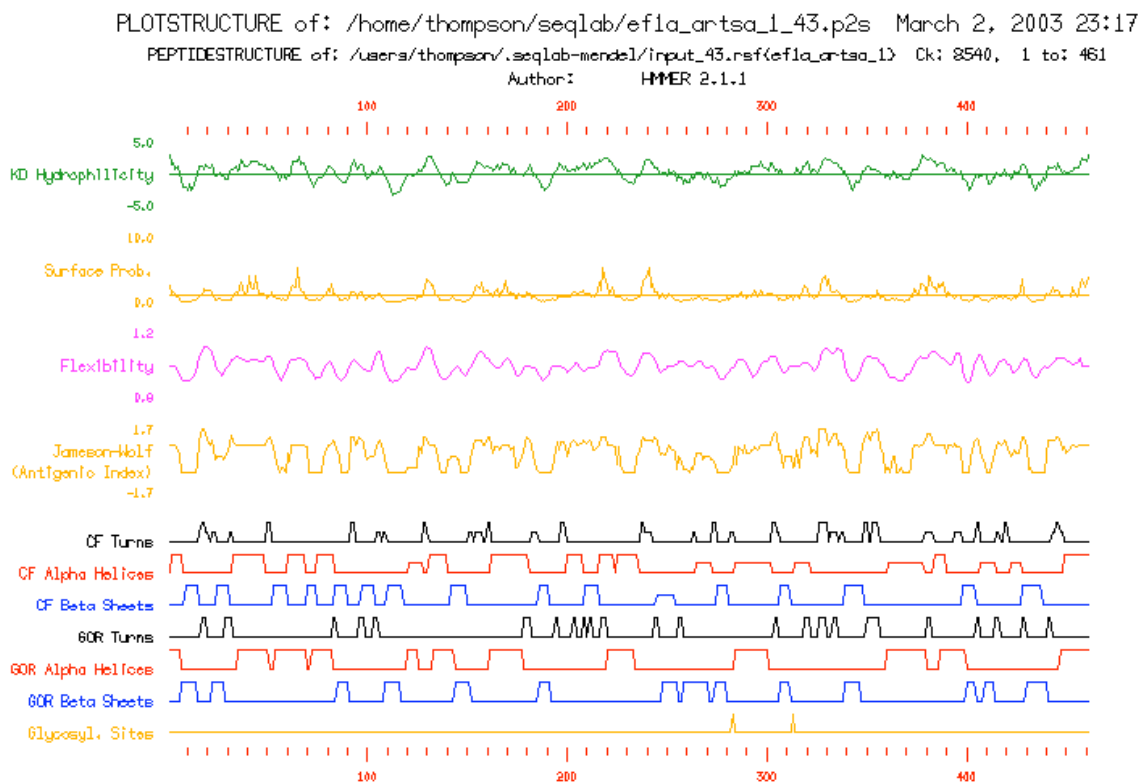
PeptideStructure makes secondary structure predictions, including B-cell antigenicity, flexibility, and surface probability, as well as a hydrophilicity determination; PlotStructure graphically displays these predictions. PeptideStructure must be run first and the output is automatically passed on to PlotStructure by SeqLab. The program is optimized for soluble, globular proteins; therefore, the window size should be changed for anything other than those sort of proteins. Run “**PeptideStructure**” on your sequence to see GCG’s second combination approach secondary structure prediction program. Notice that you can create RSF output with this program — this can be very helpful for adding annotation to your dataset. However, you cannot turn off the secondary structure predictors in this program. The output will contain all the attributes whether you want them or not. The -Broadening command line option can be helpful; accept the default Kyte-Doolittle hydrophilicity scale and window size of seven residues.

The output text file from the run, something.p2s, contains the structural and antigenic index predictions. Data is displayed in this file in columns under given headings. The Chou-Fasman predictions are under the CF-Pred heading, the Garnier-Robson are under the GORPred heading and the antigenic index information under the AI-Ind heading. Chou-Fasman uses both upper and lower case letters in their prediction scheme. Upper case letters denote strongly predicted structures while lower case indicate weakly predicted ones. Garnier-Robson uses only upper case letters in their prediction scheme, not bothering with weak predictions. But remember that both Chou-Fasman and Garnier-Robson predictions use very old and unreliable algorithms, so don’t put much faith in them!

Consider an antigenic index value of 1.0 or greater to be a probable antigenic site. This antigenic index, as opposed to that discussed previously regarding Amphi, is based on the amount of predicted surface exposure, flexibility, hydrophilicity values, and secondary predictions all combined, rather than the predicted existence 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, not T-cell. And, even though part of its

calculation relies on unreliable secondary structure predictions, overall it does a pretty decent job of estimating where the antigenic regions lay in your sequence.

“**PlotStructure**” automatically displays the data in the .p2s file graphically. You can specify a panel or a “squiggly” graph by changing the default before the run. The panel graph is much more informative than the squiggly plot and much easier to interpret. The EF-1□ example is shown below:



## Secondary structural information in PDB files

All of the Project Molecules for this class have had either their structures determined or close homologues' structures determined; all have PDB access codes. You should know this access code from earlier searching efforts. In those efforts you may have explored visualizing their structure with Entrez's Cn3D and/or RasMol. PDB files contain a lot of information beyond just the atom's three-dimensional coordinates. You can read this other annotation within Entrez by using the “Structure Info” button. Naturally, if you have the PDB file, you can also just scroll it to the screen. The structure of these files is such that 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 helpful.

## PDB file subject codes

A listing of common PDB file subject areas helpful to sequence and secondary structure analysis follow. The use of these codes makes the location of certain types of desired data from 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                                |

Since the NRL\_3D database corresponds to all of the sequences from PDB, and also contains secondary structure annotation, we can also use its entry as an easy way to see the secondary assignments found in the PDB file. NRL\_3D uses a sequence naming convention based on the corresponding PDB entry. For instance, the NRL\_3D entry that contains the sequence of PDB entry 1GF2, the insulin-like growth factor II that I used as an example for the comparison file shown previously, is 1GF2. In cases where the PDB file contains multiple chains, NRL\_3D adds numbers or letters to the end of the name to differentiate chains. You should know the variation of the PDB code that NRL\_3D uses for your Project Molecule. The Swiss-Prot database also lists secondary structure annotation for those molecules whose structures have been solved, so it can, therefore, also be used to easily read the known structural information on a protein.

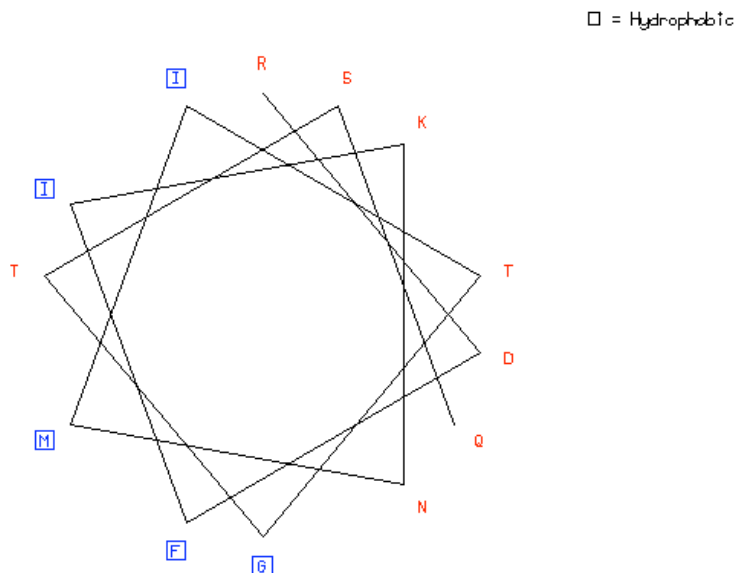
Find the entry in your RSF file that has a solved structure. If you haven't included an NRL\_3D sequence, then it's likely to be a Swiss-Prot entry with red diamond graphical annotation. GCG uses red diamonds to annotate secondary structure. Click the **"INFO"** icon to read the author's secondary structure assignment. Compare the secondary information in it to what you discovered in its homologue using the above predictive analyses. Take notes of your findings. A few warnings about PDB data need to be heeded though. First, PDB data is always on the mature protein whereas standard sequence database entries are usually precursor molecules. This yields a numbering discrepancy between PDB/NRL and PIR/Swiss-Prot. Furthermore, the aligned dataset has gaps in it. This produces another numbering discrepancy. Therefore, be sure to pay attention to the **"Position:"** not the "Column:" indicator in SeqLab. Second, PDB data entries are often multiple chains, especially when the molecule exists in nature as a complex. Both issues create confusion.

### **HelicalWheel – verification of amphiphilic helices**

After other analyses have finished, HelicalWheel can be used on those areas of the molecule that are, or show, potential of being an amphiphilic helix or sheet. Examining the results of helical wheel analysis on those areas of a predicted or known secondary element can verify whether any asymmetrical ordering of the hydrophobicity pattern of the residues within that element is present and can often give information on potential packing patterns. It is by far the easiest way to visualize this phenomenon.

Determine where the actual helical regions of your protein are by using the PDB secondary assignments found in the above section. Be sure to take the numbering discrepancy between the two database entries and between the gapped entry in your alignment and the un-gapped entry that we've been analyzing into consideration. Repeat with the sheet specifications if you would like.

With the locations of the alpha helical regions specified, run "**HelicalWheel**" on each of them 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. Record your findings for the various helices tested. If you want to test your sheet regions, repeat the analyses with the beta option. The best amphiphilic helix in my EF-1 example is shown below. It ran from residue position 95 through 107 and corresponds to peaks in both the Moment and PepPlot programs:



### Specialized protein analysis programs in SeqLab

Several programs in the "**Protein Analysis**" section of the "**Functions**" menu have very specific purposes. As you launch the following programs press the "**Help**" button to read about the algorithms. "**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 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. SPSscan 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!

### **Internet secondary structure predictions**

PredictProtein is an electronic mail service by the Protein Design Group at the European Molecular Biology Laboratory, Heidelberg, Germany. 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). PHD is a new secondary structure prediction method rated at an expected 70.2% 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://cubic.bioc.columbia.edu/predictprotein/> 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>.

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 Mendel and exit the X software on the workstation that you are sitting at.

## Homework assignment

Figure out how to submit your 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. 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.

## Conclusion: caveats and considerations

You've now gone all the way from probe design, through contig assembly, database searching, multiple sequence alignment, and phylogeny reconstruction, 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 discovered genes usually 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.

## 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.

## 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<sup>®</sup>), (Copyright 1982-2003) *Program Manual for the Wisconsin Package<sup>®</sup>*, version 10.3, [http://www.accelrys.com/products/gcg\\_wisconsin\\_package/index.html](http://www.accelrys.com/products/gcg_wisconsin_package/index.html) Accelrys, a wholly owned subsidiary of Pharmacia 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.
- 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  $\alpha$ -Helical and  $\beta$ -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.