

Sequence Characterization [Protein]

February 27, 2002

Sequence Analysis & Consulting Service

Seminar Series

Susan Johns

(johns@cgl.ucsf.edu)

Log into your socrates account by clicking on the **socrates** icon in the small window on the upper left hand side of the screen. Respond with your account name and password to the prompts.

First, copy over to your account the necessary data. Enter the following commands. [The first one does the actual copying. And the second one moves you into the working directory for today's tasks.]

```
protein_setup<rtm>  
cd protein_analysis<rtm>
```

As a result of these commands, you are now in a new sub-directory called **protein_analysis** in your account where the necessary files are located. Activate the GCG software and set your graphics device.

```
gcg<rtm>
```

Run **setplot** and select the **TEK4105** option.

In sequence analysis there are some very basic assumptions made. On the protein side of things these assumptions are:

sequences start at the N terminus and end at the C

standard one letter amino acid codes are used

Before starting to determine the nature of the unknown protein sequences, look at the sequences to get an idea of their composition.

```
socr<x>: cat unknown*<rtm>
```

Collect basic information on the protein sequences you are working on.

Running the GCG program PEPTIDESORT provides basic information.

- molecular weight
- isoelectric point
- extinction coefficient
- amino acid breakdown

After looking at the screen trace page, you can either use the information there to do a

PEPTIDESORT run in your socrates window on either unknown1 or unknown2 or run that process on the web.

Use the cat command to display the results of your PEPTIDESORT runs on socrates.

```
socr<x>: cat unknown1.pepsort<rtm>
```

or

```
socr<x>: cat unknown2.pepsort<rtm>
```

Do your results show any unusually high amounts [10 percent or higher] for any one amino acid?

The next thing to check is to see if the proteins contain any established functional patterns or known profiles of well studied protein families.

Functional patterns or motifs are checked for by the GCG program MOTIFS, while profiles are looked for by the PROFILESCAN program.

The idea behind these two programs is similar, the approach taken differs. Both start out by aligning a number of members from a given protein family.

To **generate a motif**, a section of the alignment that is known to correspond to a function is zeroed in on. A string pattern is developed for that region and tested to see just how specific it is.

Once a motif is established, it is added to a listing of such patterns that the program checks against to see if a protein contains an established motif.

Run your protein through the motifs program either in your socrates window using the screen trace as a guide or on the SACS **basic** or **advanced** pages.

Motif output takes some explanation, especially the patterns that were used to find the functional motifs.

The results show the name of the found motif, the pattern being searched for and what the match was in the given protein.

```
Ck2_Phospho_Site      (S,T)x2(D,E)  
                      (T)x{2}(D)  
  
                    50: AAVAD      TASD      AAAAA
```

The sought for pattern is either a S or a T in position 1, followed by any two amino acids, followed by either a D or an E. The actual pattern found in the protein was T followed by any two amino acids followed by a D.

The number 50 is the starting position of the found pattern in the protein being searched. Five flanking

amino acids on each side of the found pattern are displayed.

A motif can be found more than once in a sequence. This results in multiple lines starting with the position number of the found pattern.

A ~ before a series of characters within a set of parenthesis means that any character is accepted in that position except the ones given within the parenthesis.

C~(C,P,W,H,F)~(C,P,W,R)CH~(C,F,Y,W)

Numbers within curly brackets indicate a region that can vary in size from the starting number up to and including the ending one.

```
RITEAxPDPxAKAxPAAx{2,12}TASDAAAAAxAxTAxAxAAAAxAxTAAxAxAAAAxAxTx{0,26}ARG
```

Check your own motifs output by using the more command if you used socrates or review the generated web pages.

Look for anything in your output that mentions membrane or a well known protein family.

The profile system uses aligned sequences to generate a table of all the comparison information of the group of aligned sequences. The table contains as many rows as there are positions in the aligned sequences. Each row contains scores for the alignment of the corresponding position with each possible residue.

There is a much smaller number of established profiles than there are motifs.

Run your own proteins through the PROFILESCAN program and then use cat to display your results.

Check the web for additional information on your sequences. The two sites listed below are picky about the sequence input format. You will need to remove all blank lines and numbers from the data you paste in prior to doing a run.

There is a web site that allows the checking of profiles generated by the PROSITE folks. Run your proteins through this site. Just use the default settings and paste in your sequences, one per run.

Another web resource [pfam](#) allows for the searching of known protein families with hidden Markov model techniques. Recently GCG included pfam searching in their package as well as software to generate profiles on your own sequences.

When a sequence starts off with a **M** (methionine) chances are that there may be a signal peptide included in the beginning of the sequence. To determine if this is the case or not, run any such sequence through the SPSCAN program.

Run your sequence through the web version of this program. Does it have a signal region?

Another program to use to gather information on a protein sequence is called PEPTIDESTRUCTURE. This program is easy to use and requires careful interpretation to assess the results.

The running of the PEPTIDESTRUCTURE program produces a file with a p2s extension. The results in this file requires some explaining.

The output contains rows of data with values in nine different columns.

Pos	AA	GlycoS	HyPhil	SurfPr	FlexPr	CF-Pred	GORPred	AI-Ind
1	M	.	-1.675	0.414	1.000	h	H	-0.450
2	A	.	-2.100	0.267	1.000	h	H	-0.450
3	L	.	-2.217	0.181	1.000	h	H	-0.450
4	S	.	-1.800	0.264	1.000	h	H	-0.450
5	L	.	-2.129	0.194	0.938	h	H	-0.600

The first column is the position number, the second the residue name in that position.

Column three shows if that position is a glycolysation site. Membrane proteins usually have glycolysation sites.

Column four is the hydrophilicity value at that position. Philic values are positive, phobic are negative. Membrane proteins have regions of phobic values for about 18 positions (or more) in a row.

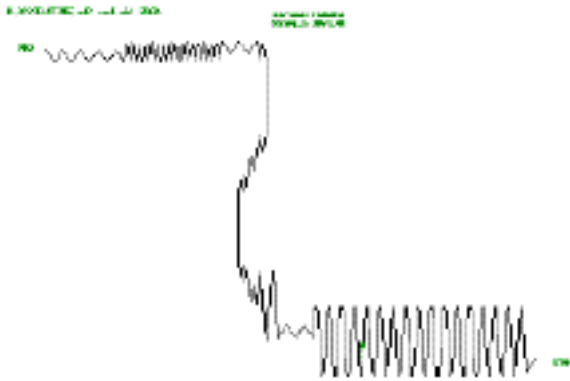
Surface probabilities and flexibility are columns five and six respectively.

Column seven are Chou-Fasman secondary structure values. This technique uses upper case characters for strong predictions and lower case ones for weak.

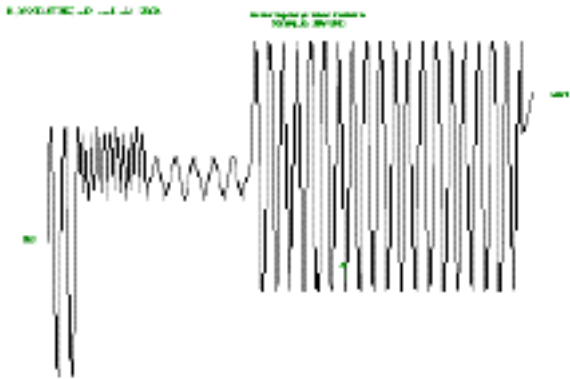
Garnier-Osguthorpe-Robson secondary structure prediction values are in column 8. Only upper case characters are used in this technique.

In the ninth column are antigenicity index values. Values here greater than 1 are significant and show possible surface area(s) capable of generating an antibody.

The PEPTIDESTRUCTURE p2s output file can also be displayed graphically to show secondary prediction values. Given on the next page is the results of running the ref.p2s file through the PLOTSTRUCTURE program.



Only the strong Chou-Fasman predictions are shown here.



GOR prediction results

Run your proteins through the web **basic** version of the program. The plot is automatically displayed, instead of needing a second step to see the results. However, in the basic version, it is a panel plot that is shown and not the secondary structure predictions as given above.

To see the secondary structure predictions, run your proteins through the **advanced** web version of the program. Select the non-default plot option to see the predictions. You can elect to display either the Chou-Fasman or Garnier-

Robson results.

As you can see the amount of data is growing and still you have just begun to explore what information is available. There are times when a serious protein analysis project requires the assembling of the bits and pieces of information into a single place.

One suggestion for doing this is to start with a copy of the sequence being explored and then add data as you work along.

original sequence data:

```
1  MALSLFTVGQ LIFLFWTLRI TEANPDPAK AAPAAVADPA AAAAAAVDT
51 ASDAAAAAAA TAAAAAKAAA DTAAAAAKAA ADTAAAAAEA AAATARG
```

This data after running motifs, profilescan, spscan and peptidestructure.

motifs: - only phosphorylation sites found
 profile scan:- none found
 spscan:- possible signal from 1-23
 glycos: - none found

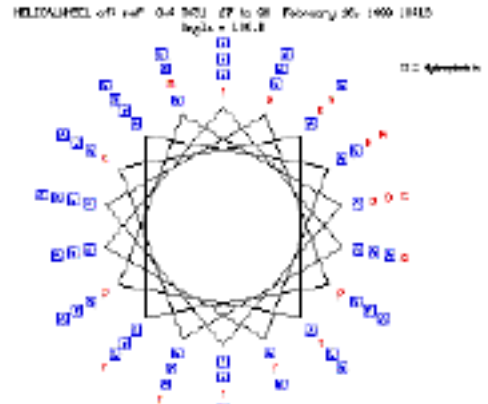
```
AI:                                     xxx
GOR: HHHHHHHHBB BBBB                 HHHH HHHHHHHHHH HHHHHHHHHH
CF: hhhhhhhh B BBBBHHHHHH           TTHHHH HH   HHHH HHHHHHHHHH
signal: xxxxxxxxxxx xxxxxxxxxxx xxx
1  MALSLFTVGQ LIFLFWTLRI TEANPDPAK AAPAAVADPA AAAAAAVDT
```

```
AI:
GOR: HHHHHHHHHH HHHHHHHHHH HHHHHHHHHH HHHHHHHHHH HHHHHH
CF: HHHHHHHHHH HHHHHHHHHH HHHHHHHHHH HHHHHHHHHH HHHHHH
signal:
51 ASDAAAAAAA TAAAAAKAAA DTAAAAAKAA ADTAAAAAEA AAATARG
```

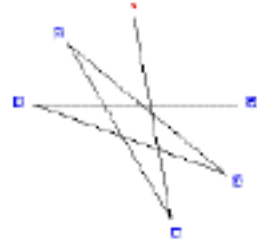
As more data is collected it is added to the information file.

Sequences with secondary structure predictions can be checked to see if the helical or sheet elements are organized. That is, are they aligned on the surface of the predicted feature.

The program to check for this is HELICALWHEEL. [This is a graphics program.]



helical section of ref (27-96)
 Running this program with the **-beta** switch allows you to determine the sidedness of sheets.



sheet section of ref (10-15)

Sidedness may give clues as to where in a structural

conformation a given element is (surface or internal core).

Select a possible helical element from your peptidestructure results and run it through the helicalwheel web page.

With this amount of basic information, it is time to see if the unknown sequences are similar to any proteins that have had their structures solved. The **NRL_3D** database just contains the primary sequence data from proteins with solved structures.

One program to check for this is GCGFASTA. The NRL_3D database is rather small and this can be done interactively.

[In the screen trace for ref only the mature form of the protein was used to increase the chances of finding a hit. X-ray structures are only done on mature proteins.]

For our purposes, any hit with a fasta z score of less than .1 will be considered to be significant. After all, we are just attempting to find out what well studied protein(s) our unknowns are similar to.

Using the screen trace as a guide, run your own unknown through the program. When the results reach the screen, use the scroll bar to look back to determine how many of the hits to show alignments for.

Needless to say, not all unknown proteins will find a hit in the NRL_3D database. The number of protein families represented there is small in comparison to the total known.

If your search didn't produce any real high quality hits, repeat your GCGFASTA run, this time with a larger database. The following is a pecking order to try. The databases on the list go from small to large and from well documented to less well documented. You could also do a BLAST runs with your sequence.

- nrl_3d - x-ray structures (~23,000)
- swissprotein - excellent documentation (~115,000)
- pir - 4 divisions varying quality (~680,000)
- genpept - very little documentation (~897,000)

The amount of time needed to complete any such GCGFASTA or BLAST run increases with the size of the database being searched.

Once a quality hit has been found, the question is, have you really found a family member or not.

- Do they share motifs?
run MOTIFS and/or PROFILESCAN on hit
- Are their alignments statistically significant?
determine real z-scores (GAP or BESTFIT -random=100)
- Do they give similar results in PEPTIDESORT or PEPTIDESTRUCTURE?

What happens next depends on what you found.

modelling attempt:

If you have reasonable confidence that you have found a possible member of a protein family with solved structure, then make an attempt at getting a possible structure off the web.

The site to do this (Swiss-Model) is given below.
<http://www.expasy.ch/swissmod/SWISS-MODEL.html>

If a model can be generated, you will be sent back e-mail containing a PDB formatted file with the coordinates of your structure.

other tests:

There are other programs to run to determine different possible protein features.

COILSCAN

This program attempts to detect coiled-coil segments in proteins. Coiled coils are bundles of two or more alpha helices that are supercoiled together. The method accurately predicts parallel and antiparallel two-stranded coiled coils and parallel three-stranded coiled coils.

HTHSCAN

The program scans protein sequences for the presence of helix-turn-helix motifs, indicative of sequence-specific DNA-binding structures often associated with gene regulation.

Very few transmembrane proteins have had their structures solved. If you have reasonable confidence that you have found a possible transmembrane protein, then make an attempt to firm up that by confirming it from the web.

The sites to do this are:

DAS - Prediction of transmembrane regions using the Dense Alignment Surface method (Stockholm University)

- input format: sequence only
- output format: html page (data displayed as graph)

HMMTOP - Prediction of transmembrane helices and topology of proteins (Hungarian Academy of Sciences)

- input format: fasta, pir/nbrf, swissprot id or accession number has options to select
- output: html page

MEMSAT2 - Prediction of transmembrane regions (University College London)

- input format: sequence only enter server, select MEMSAT2 from list of prediction methods
- output format: email message

MEMSAT (original - local) - Prediction of transmembrane regions

- input format: sequence, GI or accession number
- output format: text with TOPO image of prediction

PHDhtm - Advanced page for selecting only transmembrane helices prediction (Columbia University -PredictProtein)

- input format: sequence only
- email address required
- select the PHDhtm option from the type of prediction menu
- output: email message

PRED - TMR - Method based on novel statistical analysis of SwissProtein database. (University of Athens, Greece)

- input format: plain text, swissprot, fasta
- output: html page with predicted regions shown

Sosui - Classification and Secondary Structure Prediction of Membrane Proteins. (Tokyo University of Agriculture and Technology, Japan)

- input format: sequence only
- output: html page with predicted regions shown as well as hydropathy profile and helical wheels

Split - Membrane Protein Transmembrane Secondary Structure Prediction Server (University of Split, Croatia)

- input format: sequence only or browse SwissProtein
- output: html page with predicted regions graph, can look at numerical results

TMAP single- Prediction of transmembrane regions

- input format: sequence only
- output format: html page of predicted locations plus downloadable ps file of results

TMHMM - Prediction of transmembrane helices in proteins (CBS: Denmark)

- input format: fasta
- output: html page with image of predicted locations

TMpred - Prediction of transmembrane regions and protein orientation (EMBnet-CH)

- input format: sequence, swissprot id or accession number, GenBank gi
- output: html or ascii

TSEG Prediction using clustering and transmembrane segment discriminant analysis. By Daisuke Kihara, Doctoral Student, Institute for Chemical Research, Kyoto University (Japan)

- input format: sequence only
- output: form implies email, but results show up on html page.

Run any possible transmembrane protein through these sites to see if that prediction holds.

TMpred results:

The sequence positions in brackets denominate the core region. Only scores above 500 are considered significant.

Inside to outside helices : 3 found

from	to	score	center
20 (25)	43 (43)	2183	35
53 (53)	71 (71)	1856	63
98 (101)	119 (117)	2021	109

Outside to inside helices : 3 found

from	to	score	center
20 (20)	40 (38)	2164	30
53 (53)	72 (70)	1746	62
98 (102)	121 (119)	1716	110

TopPred2 results:-

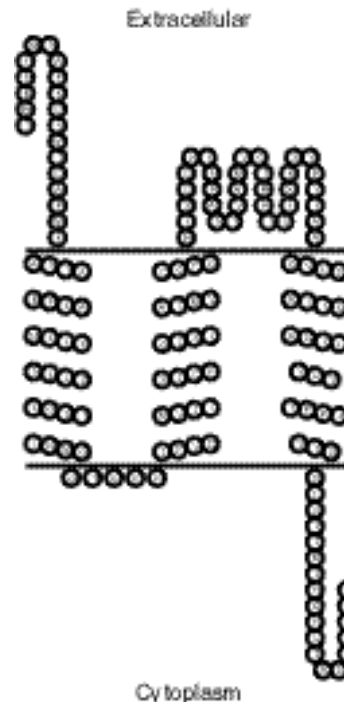
Candidate membrane-spanning segments:

Helix	Begin	End	Score	Certainty
1	21	41	1.895	Certain
2	52	72	1.874	Certain
3	105	125	1.643	Certain

There appears to be three transmembrane segments

predicted from web	predicted from p2s
20 - 43	22 - 39
49 - 72	51 - 68
105 - 126	96 - 131

topo image



URL's used in this seminar

local:

coilscan: <https://www.sacs.ucsf.edu/secure/cgi-bin/coilscan.pl>

hthscan: <https://www.sacs.ucsf.edu/secure/cgi-bin/hthscan.pl>

motifs (basic): <https://www.sacs.ucsf.edu/secure/cgi-bin/analyze.pl?anatype=motifs>

motifs (advanced): <https://www.sacs.ucsf.edu/secure/cgi-bin/motifsadv.pl>

peptidesort: <https://www.sacs.ucsf.edu/secure/cgi-bin/analyze.pl?anatype=peptidesort>

peptidestructure (basic): <https://www.sacs.ucsf.edu/secure/cgi-bin/analyze.pl?anatype=peptidestructure>

peptidestructure (advanced): <https://www.sacs.ucsf.edu/secure/cgi-bin/pepstructadv.pl>

spscan: <https://www.sacs.ucsf.edu/secure/cgi-bin/spscan.pl>

remote:

profilescan server (prosite profiles): <http://hits.isb-sib.ch/cgi-bin/PFSCAN>

pfam site: <http://pfam.wustl.edu/hmmsearch.shtml>

swissmodel: <http://www.expasy.ch/swissmod/SWISS-MODEL.html>

protein tools: <http://us.expasy.org/tools/>

transmembrane prediction sites:

DAS: <http://www.sbc.su.se/~miklos/DAS/>

HMMTOP: <http://www.enzim.hu/hmmtop/>

MEMSAT2: <http://bioinf.ucl.ac.uk/psipred/>

MEMSAT(original - local): <https://www.sacs.ucsf.edu/secure/cgi-bin/memsat.pl>

PREDICTPROTEIN: http://dodo.cpmc.columbia.edu/predictprotein/submit_adv.html#top

PRED - TMR: <http://o2.db.uoa.gr/PRED-TMR/>

Sosui: <http://sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html>

Split: <http://pref.etfos.hr/split/>

TMAP single: <http://www.mbb.ki.se/tmap/single.html>

TMHMM: <http://www.cbs.dtu.dk/services/TMHMM-1.0/>

TMPRED: http://www.ch.embnet.org/software/TMPRED_form.html

TSEG <http://www.genome.ad.jp/SIT/tseg.html>

display sites:

TOPO (new version) <http://www.sacs.ucsf.edu/secure/cgi-bin/topo2.pl>

TOPO (old open to the general public) <http://www.sacs.ucsf.edu//TOPO/>