

# Practicum Bioinformatics - answer key

Prof. Kris Laukens - Adrem Data Lab  
Department of Computer Science, University of Antwerp

Academic year 2024-2025

## Contents

<b>1</b>	<b>Phylogenetic trees, protein structure &amp; gene ontology</b>	<b>2</b>
	Section 6: Phylogenetic trees . . . . .	2
	Exercise 6.1: Unknown organism and its position in the tree of life . . . . .	2
	Exercise 6.2: Orthologs and Paralogs: the evolution of hemoglobin . . . . .	4
	Section 7: Protein structure . . . . .	9
	Exercise 7.1: Primary sequence based information . . . . .	9
	Exercise 7.2: Secondary structure prediction . . . . .	24
	Exercise 7.3: Viewing 3D protein structures . . . . .	29
	Section 8: Gene ontology . . . . .	35
	Exercise 8.1: GO enrichment . . . . .	35
	Exercise 8.2: Searching the GO database (GO terms) . . . . .	36
	Exercise 8.3: Searching the GO database (protein) . . . . .	38

### Note about the answers

These are just example solutions. For many of the exercises there will be other correct solutions as well. The main goal of these practicals is to become acquainted with the different databases, techniques and principles of bioinformatics. Try to understand why we are using a specific technique, the different aspects of the outputs of the tools and how they work, rather than learning everything by heart. You will need to be able to interpret related questions to the ones you've solved here while using this text as a quick reference, but you will need to know the major sections and themes of the practicals if you want to be able to retrieve them in a timely fashion.

# 1 Phylogenetic trees, protein structure & gene ontology

## Section 6: Phylogenetic trees

### Learning goals

- know the terminology common in phylogenetic analysis
- understand the role and assumptions of different substitution models
- understand the principles and workings of a selected range of phylogenetic methods
- being able to interpret phylogenetic trees and their bootstrap values
- know practical research applications for phylogenetic methods

### Exercise 6.1: Unknown organism and its position in the tree of life

Imagine you are a bioinformatician and you are contacted by UZA for an urgent query. A patient of theirs was admitted to intensive care and suffers from multi organ failure. The situation is critical and they do not know what to do. UZA is considering a tropical pathogen since the patient had a recent trip to Nepal and started feeling sick 6 months after his return with unspecific symptoms like fever, diarrhea and weakness. His health rapidly declined over the following weeks and he had to be hospitalized leading eventually to the current situation. The only lead they currently have, is that they have detected an unknown parasite in the spleen (which was severely enlarged), of which they could amplify and sequence the 18S rRNA. Your job now is to find a drug with which they can treat the patient.

Of course you have a great idea! First, you will construct a phylogenetic tree with the 18S sequence of organisms that you know as well as the unknown organism. Using this tree, you will try to find an organism that is very related to this unknown pathogen and for which there are already approved drugs available.

Go to Blackboard and find the file "6-1-phylogeny.fasta". This file is a .fasta file and contains a set of 18S rRNA sequences of different organisms. The first sequence was sent to us by UZA and is labeled "unknown organism". The rest of the sequences are 18S rRNA genes from all over the tree of life, which you have collected using your awesome NCBI database skills from the first practical session.

- 18S rRNA is often used in species-level phylogeny, why?

This is a strongly conserved gene. Therefore, it is suited to make comparisons between rather distantly related species (as long as there is still a decent alignment possible). Note that for closely related species, this would actually be a bad choice, since you would not find many mutations/differences between the sequences to base your tree on. In conclusion, for closely related species or population-level phylogenies, you should use rapidly evolving genes instead.

We want to know more about its position in the tree of life and its closest neighbours. Therefore we will create a phylogenetic tree.

- Perform a multiple sequence alignment with T-COFFEE (practical 1; <https://www.ebi.ac.uk/Tools/msa/tcoffee/>). What kind of MSA algorithm was this again?

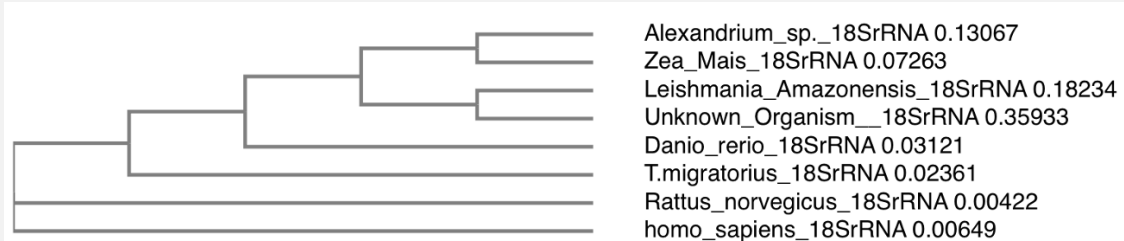
T-coffee is a consistency-based multiple sequence alignment program

- Go to the phylogeny tab and interpret the results. Note that we are looking at a *neighbour-joining tree*. What kind of tree building method is this? Does it make any assumptions you should be aware of? (A quick google search can help you to understand which species the Latin names stand for.)

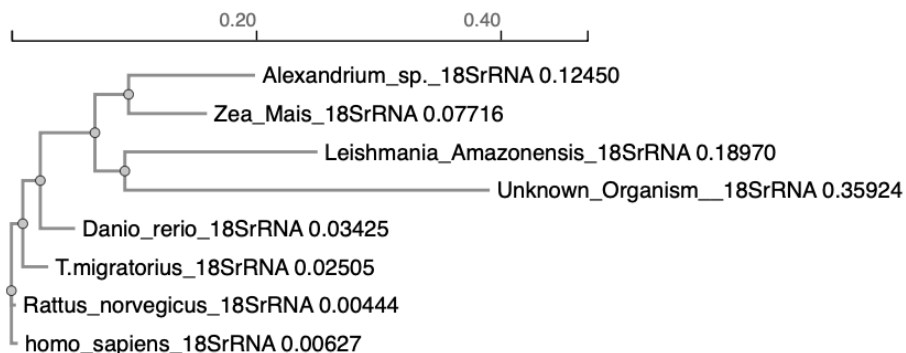
Neighbour-joining is a distance-based method, which starts from a distance matrix. The relationship between certain characters is therefore not taken into account (a lot of information captured within the sequences is discarded).  
For neighbour-joining we do not assume that the evolutionary rate is the same in all lineages.

- Explore the phylogram and its branch length, how is this different from a cladogram?

Cladogram:



Phylogram:



In a cladogram, the branch lengths do not mean anything (they are always equally long). In the phylogram, the branch length really represents the inferred evolutionary distance between the organisms.

- Look at the most closely related organisms to our sample, what does this tell you about the organism we are looking for?

The organism is apparently most related to *Leishmania amazonensis*.

We did not invent a 18s rRNA sequence, neither is it from an unknown species...

- Find out from which organism this rRNA was sequenced with a tool from last week.

Using BLAST, we can find out the species is likely *Leishmania donovani*.

- Were your suspicions correct?

### Exercise 6.2: Orthologs and Paralogs: the evolution of hemoglobin

Apart from looking at evolution between species, phylogenetic trees can also be used to study the evolution and duplication events within protein families.

In this exercise, we will take a closer look to the evolution of hemoglobin (and subunits) and myoglobin. These proteins are specialised in the binding of oxygen. Myoglobin is made up of 1 individual protein chain responsible for the transport and storage of oxygen in the muscles. Hemoglobin consists of 4 subunits (2 alpha globin chains and 2 beta globin chains) and is present in the blood to transport oxygen around the body.

- We prepared a multifasta file for you (6-2-globins.fasta), consisting of Hemoglobin subunit alpha, Hemoglobin subunit beta and myoglobin sequences from Human, mouse and zebrafish <sup>1</sup>
- Browse to <http://phylogeny.lirmm.fr/> (or [http://www.phylogeny.fr/simple\\_phylogeny.cgi](http://www.phylogeny.fr/simple_phylogeny.cgi) if the first link does not work)
- Choose the "one click" phylogeny analysis and upload the file. Find out which steps the program uses, and why!

1. Multiple sequence alignment with MUSCLE
2. Gblocks to eliminate poorly aligned positions and divergent regions
3. PhyML Maximum likelihood tree creation
4. Tree rendering (visualising)

- Look at the phylogenetic tree and interpret your results. What can you decide about the evolution of these globins? Hint: Look up the definition of orthologs and paralogs.

See below for tree and interpretation.

- Where are the common ancestors in a phylogenetic tree?

The common ancestors are the internal nodes (i.e. places where two branches connect).

- How does the tree show the direction of time?

---

<sup>1</sup>Of course, you can make it even more interesting and add other species/related proteins! You can go to UniProt and download sequences yourself (add to 'basket' and download all sequences at once in UniProt). For this exercise, we searched for example for HUMAN Hemoglobin subunit alpha, HUMAN Hemoglobin subunit beta, MOUSE Myoglobin, ZEBRAFISH Hemoglobin subunit beta, .... etc

Time flows from left to right in this tree, or from the root to the tips in general.

- Is the order of the taxa important at all?

No. Every binary split in the tree can be mirrored around the internal node, so the order does not mean much. Moreover, it is not true that the taxa on the top (or left) side of the tree are older or less advanced. All taxa we used as input are *equally evolved*.

- Is it true that the hemoglobin genes evolved from the myoglobin genes (i.e. myoglobin is an ancestor of hemoglobin)?

Not necessarily. We can only state that they share a common ancestor somewhere in the past. All present-day taxa should be considered evolutionary *cousins* of each other.

- What does the branch length stand for? Is this the case for all trees?

In this phylogram they represent the (estimated) evolutionary distance or genetic change. Not all branches are equal, since some taxa might have undergone faster evolution than others.

In cladograms only the tree topology (shape) is of importance, the branch lengths are meaningless and they will all be equally long).

- What are bootstrap values? How are they calculated and what is their importance? Can you find them on the generated figure?

Bootstrap values show you how reliable (robust) your tree topology is, i.e. how well supported are specific nodes or splits in the tree. They are calculated as follows:

Take random columns from your alignment with replacement, so a single column can be picked twice (permutation). Reconstruct the phylogenetic tree using this new changed (re-sampled) alignment. If you make 1000 trees in this manner and in 100 of them you find the same node, the support of the node (or bootstrap value) is 10. Usually we only accept nodes that have a bootstrap value of at least 95.

A good way to think about this is that we are trying to shake up the data by applying many small changes to it and asking ourselves: is the clade I found sensitive to these perturbations?

If the resulting clade no longer shows up after we tweaked the original data, this indicates that the clade was likely the result of a chance event. The only reason it showed up, was because of a few specific sites in our alignment that happened to be there. If we disregard these sites, the relationship disappears. In other words, the apparent signal (relationship between two taxa) was caused by noise in the data.

Conversely, if a specific pattern keeps showing up in the data, the overall signal in the data is all pointing to the same direction and small changes don't disturb the consensus conclusion.

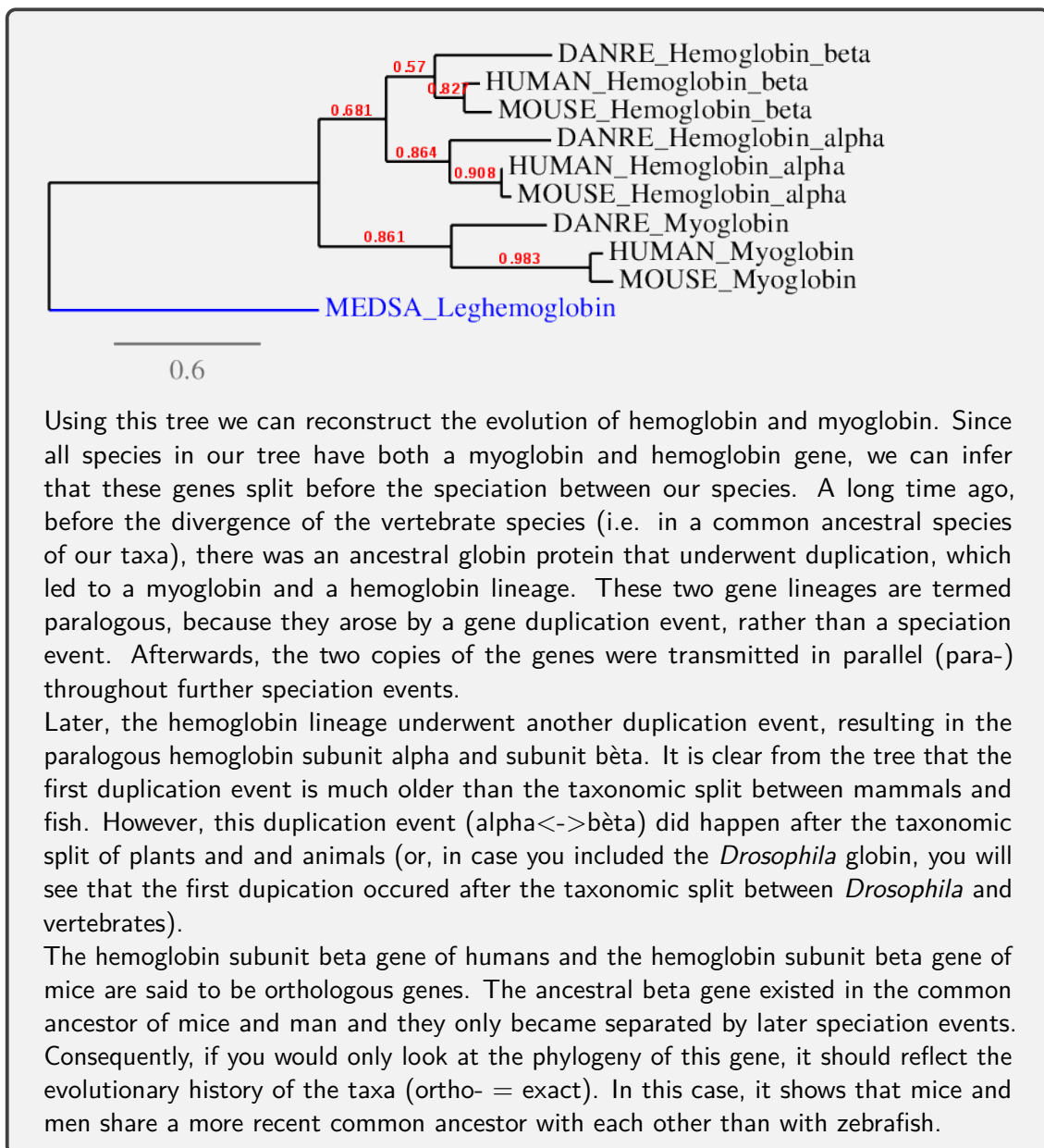
Note that these are NOT the same as statistical confidence intervals! They are a measure of confidence, but are calculated differently and under different statistical assumptions.

- Change the tree lay-out to "radial" and output as pdf. Do your conclusions stay the same?

The conclusions remain the same, but for some purposes, a radial tree could be more/less easy to interpret. However, this tool seems to have problems placing the labels in that case, unfortunately.

The final step is to root the tree by adding a less related globin (out-group). This will allow us to see at which point a protein splitted in a myoglobin and hemoglobin lineage.

- Search and add **Leghemoglobin** from *Medicago sativa*. Of course, you can add other globins as well.
- What does this tree tell you about the origin of the hemoglobin and myoglobin lineage? Do these genes track the speciation events? What are orthologs and paralogs again?



### Some notes on *tree-thinking*

To help you get better at interpreting evolutionary trees, go ahead and take a look at the following website ([click here](#)), which lists some of the more common misconceptions and pitfalls you can encounter when deciphering phylogenies.

- Try to describe evolutionary relatedness in your own words. What makes humans more closely related to chimps than chickens? Recall the meaning of monophyletic groups!
- Now go back to the phylogenetic tree of the different globins proteins. Is *Danio rerio* hemoglobin subunit beta 1 more closely related to *Danio rerio* myoglobin or to *Medicago* Leghemoglobin? Why? Contrast this with the branch lengths and their meaning!

- Which of the following ways should you use to decide on the relatedness between two species?
  1. Compare the distance between branch tips.
  2. Compare number of species in between the two species?
  3. Count the number of internal nodes in between the two species?
  4. Compare the time since the common ancestor?

Evolutionary relatedness should always be interpreted by looking at the most recent common ancestors. Humans and chimpanzees share a more recent common ancestor with each other than with humans and chickens. This is basically the definition of monophyletic groups.

The *Danio rerio* hemoglobin subunit beta 1 protein is more related to the *Danio rerio* myoglobin than the *Medicago* Leghemoglobin, since they share a more recent common ancestor.

**Never** measure the distance between branch tips to make conclusions about evolutionary relatedness. Different species/genes/proteins can evolve at different rates, so this can completely skew the picture!

Counting the number of species in between is also wrong, because you can rotate branches along their nodes without sharing the relation between the taxa.

Counting the number of nodes in between the two species does not work either, because it depends on how many closely related taxa you add to the tree. Imagine if we constructed the globin tree without any mouse genes. The relatedness of the human alpha unit would have fewer internal nodes in between it and any other protein!



## Section 7: Protein structure

### Learning goals

- understand what the primary, secondary and tertiary structure of a protein refers to and how these concepts are related
- acquire basic practical knowledge on how the secondary structure of a protein can be inferred from a sequence
- testing different structure prediction methods, and understand in broad terms which technique/algorithm you are using, and what the advantages, disadvantages and assumptions are
- exploring 3D protein structures and their use in a research setting

In this section, we will analyse the primary, secondary and tertiary structure of a globin and several other proteins.

ExPASy (Expert Protein Analysis System) hosts a collection of databases and tools related to biological data and data analysis, but focuses primarily on the analysis of protein data. You can find ExPASy at <http://www.expasy.org/>.

### Exercise 7.1: Primary sequence based information

**A)** The primary sequence of a protein can be used to deduce a number of biochemical properties. Although many of these properties are subject to complex and subtle interactions, these interactions are largely determined by the primary amino acid composition. Many tools exist that try to predict these properties by using approximate methods that exploit the information encoded in the primary protein sequence and the amino acid's physico-chemical properties.

Navigate to the ProtParam tool by finding it on the ExPASy homepage or by searching for it in the search bar. Based on the primary sequence composition, it will calculate a number of physico-chemical properties for your sequence. Which properties does ProtParam predict? Compare a randomly generated sequence with the hemoglobin subunit alpha:

```
>Randomly_generated_sequence
```

```
MGLYNIIWFDGLTQWAKAVESLHIHKCKLKGNISPFRLANINIIWLCPQLPQLVINETL
NVGSGRLPKNAYYFLRLSLYILLEVTPLVIFLSLALFTIMARIILYSDNYLVVVKYCLKN
SKLPYIINFYCNCNYMVYFLLFPIGFLFYTMPNERFPLEYVYDGFMTLINPALLVAGKG
IPVYLNRLQPSLFSLLVQR
```

```
>sp|P69905|HBA_HUMAN_Hemoglobin_subunit_alpha
```

```
MVLSPADKTNVKAAGKVGAGHAGEYGAEALERMFSLFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDDKFLASVSTVLTSKYR
```

Properties predicted by the ProtParam tool include:

- Number of amino acids
- Molecular weight
- Theoretical pI
- Amino acid composition
- Total number of positively/negatively charged residues
- Chemical formula
- Total number of atoms
- Extinction coefficients
- Estimated half-life
- Instability index
- Aliphatic index
- Grand average of hydropathicity

An explanation of these properties can be found here: <https://web.expasy.org/protparam/protparam-doc.html>

Some notable differences between the two protein sequences include:

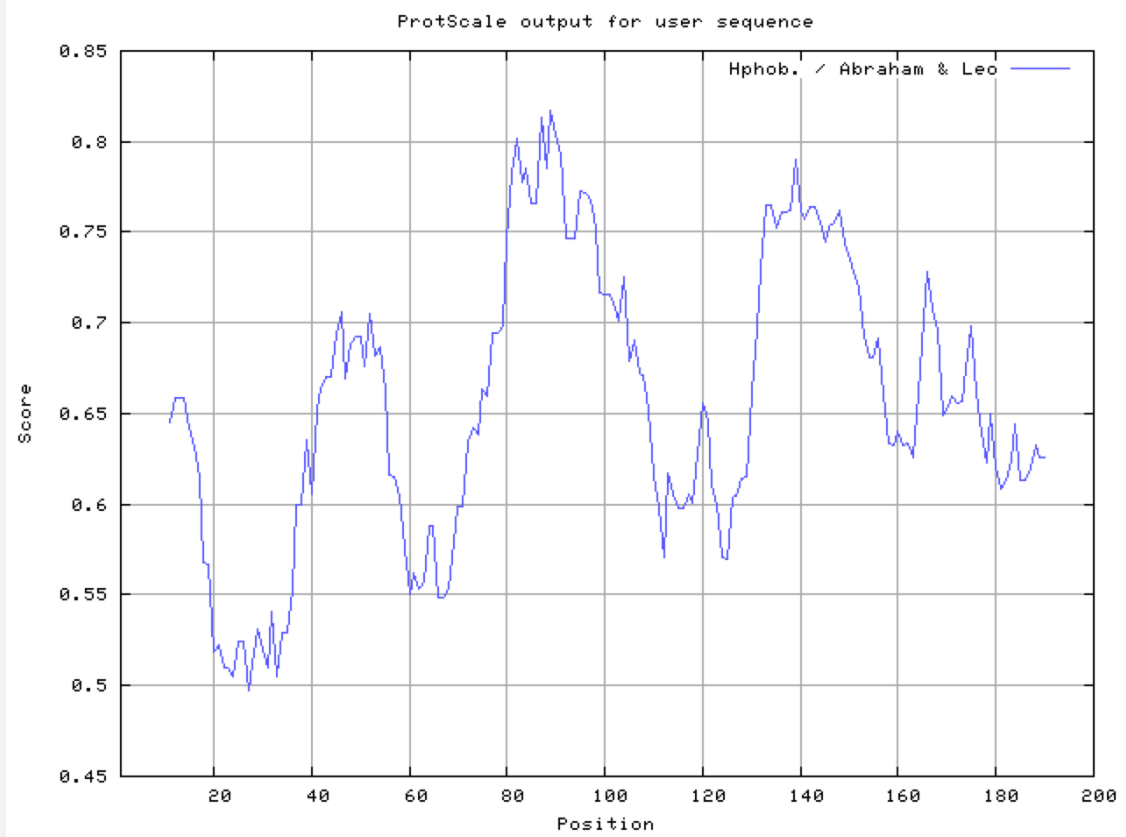
- the large number of leucines in the random sequence (17,5%)
- the instability of the random protein vs the stability of the hemoglobin subunit alpha protein

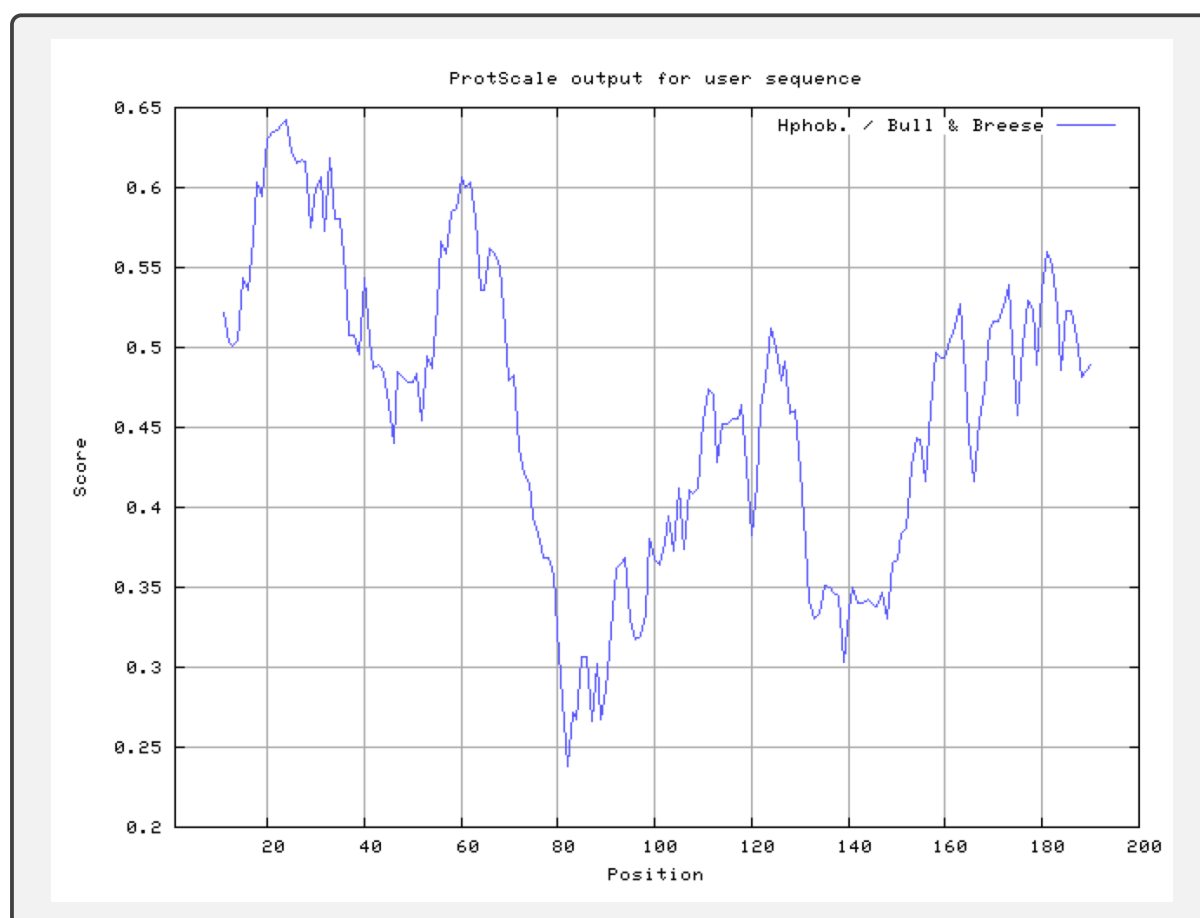
**B)** Find the ProtScale analysis tool on ExPASy. Similar to the ProtParam tool, ProtScale is capable of calculating different physico-chemical properties based on the primary protein sequence.

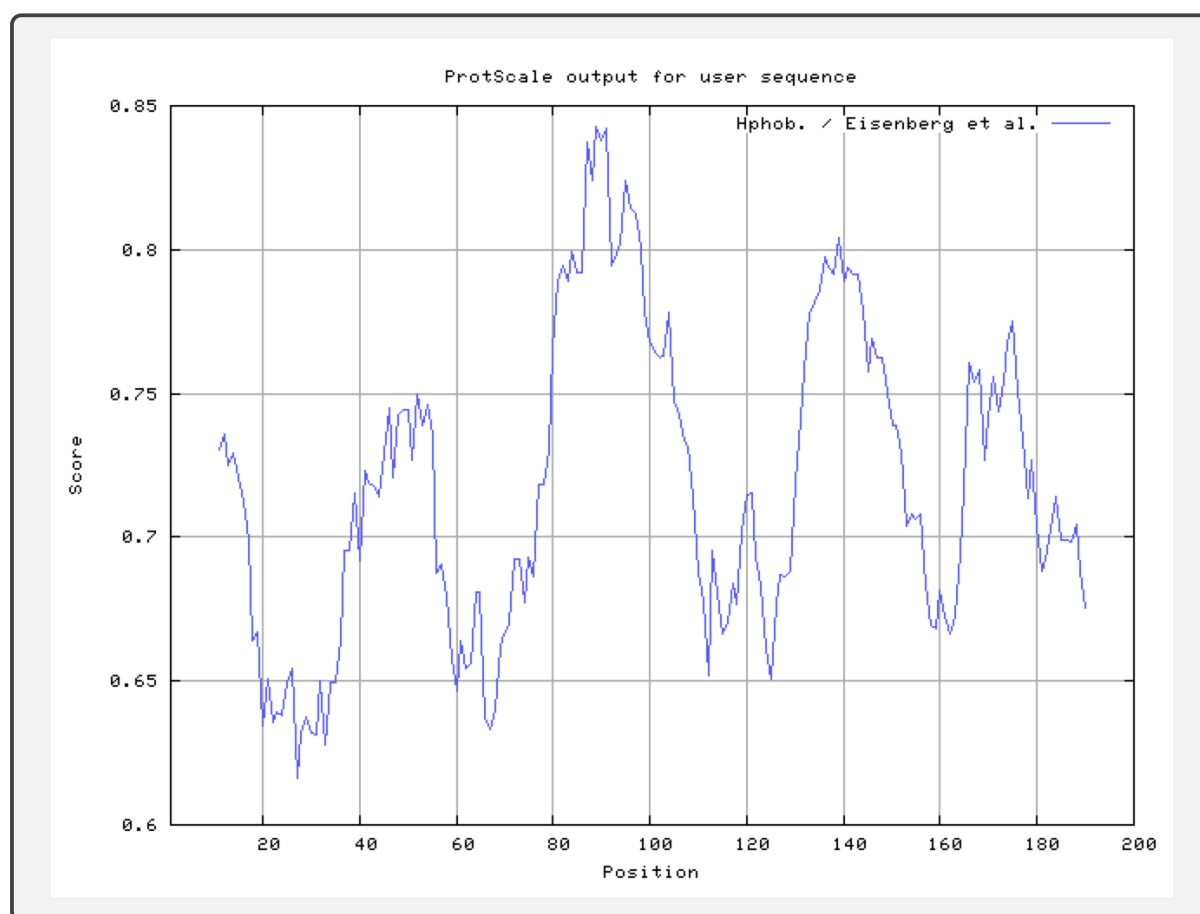
- Try at least 2 of the hydrophobicity prediction tools (denoted by Hphob. /). How well do the results match?
- Predict the transmembrane tendency of the protein and compare it to the results from the hydrophobicity prediction tools.
- Also predict the alpha-helices and beta-sheets with the Chou & Fasman method.

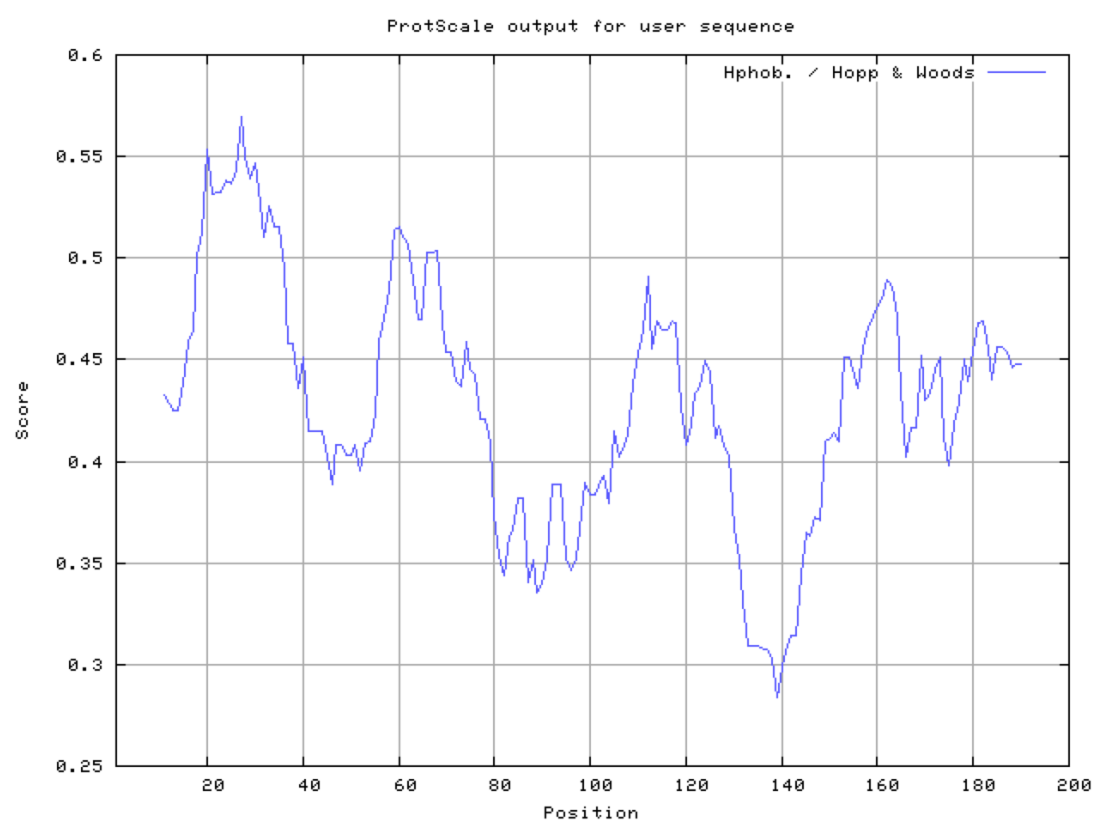
Keep all results for later comparison with other secondary structure prediction methods.

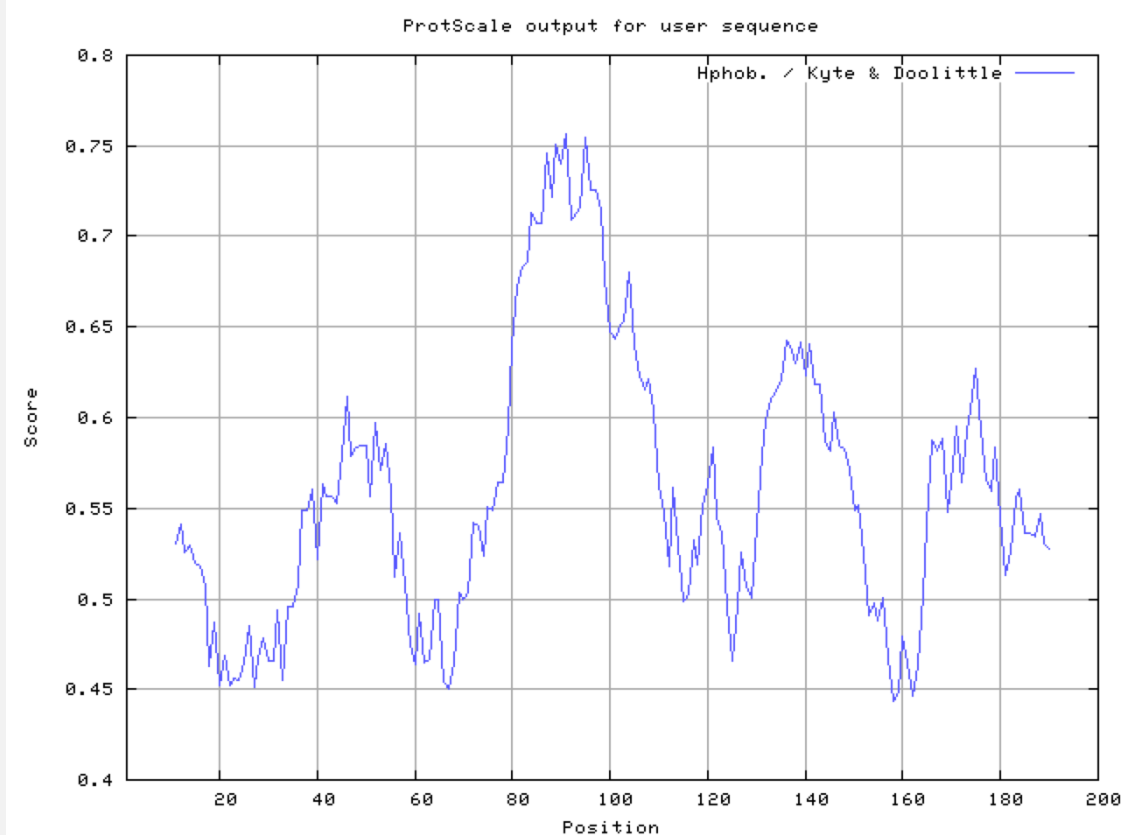
### Hydrophobicity plots - random protein sequence:







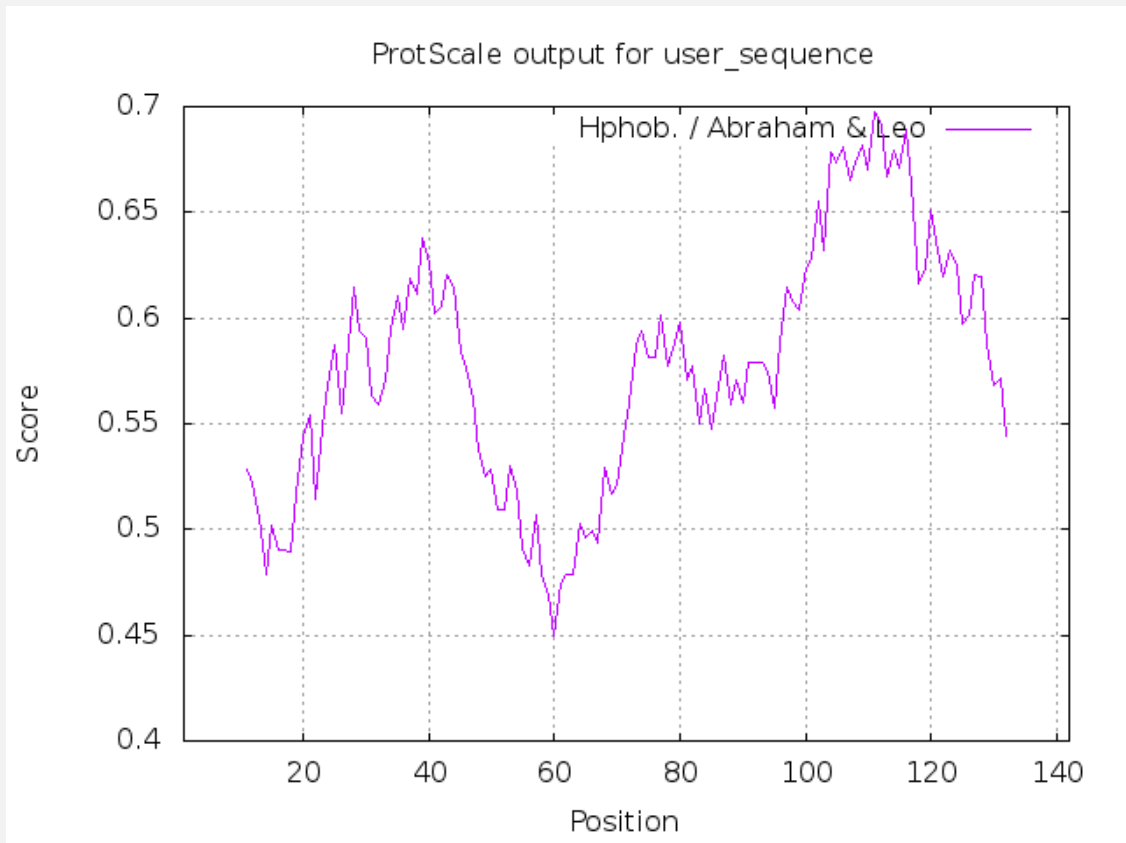




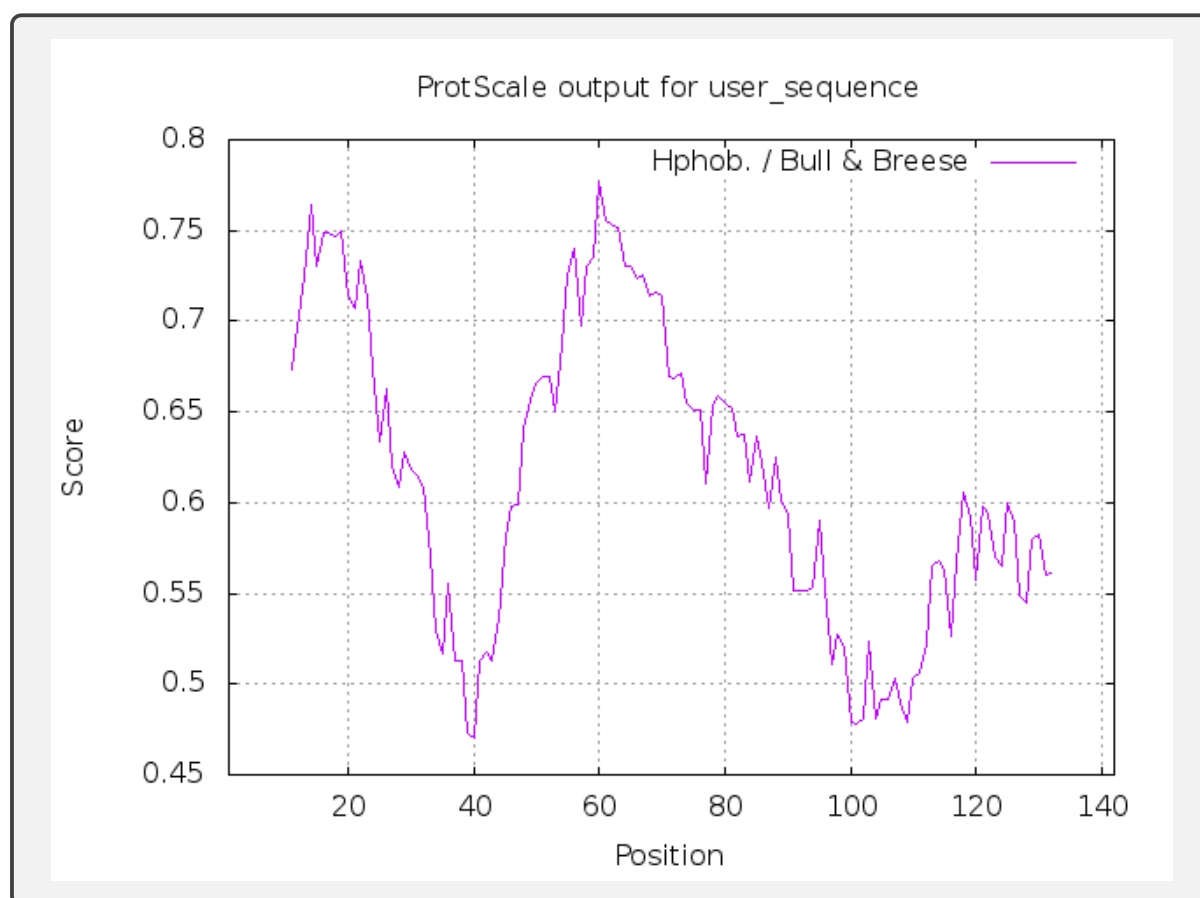
Based on the different predictions, it's not easy to assign hydrophobic regions to the protein sequence. Each sliding window method uses other amino acid hydrophobicity weights and thus returns a different result. To complicate matters, these predictions only give you a global profile, but do not provide a cut-off value. It is up to you to decide a threshold value, which can be rather hard to choose in a meaningful way. However, by comparing different methods, some trends can be seen. If a hydrophobic patch is predicted in multiple methods, the result becomes more robust. Since we are looking to compare different methods, it makes sense to rescale all the outputs between 0 and 1. It also makes sense to extend the sliding windows region to 19 or 21 amino acids. The range of the windows should be more or less equal to the length of the protein sequence expected to display the property we are predicting. Since hydrophobic patches are typically found in transmembrane segments of the protein, we expect them to have a reasonable length.

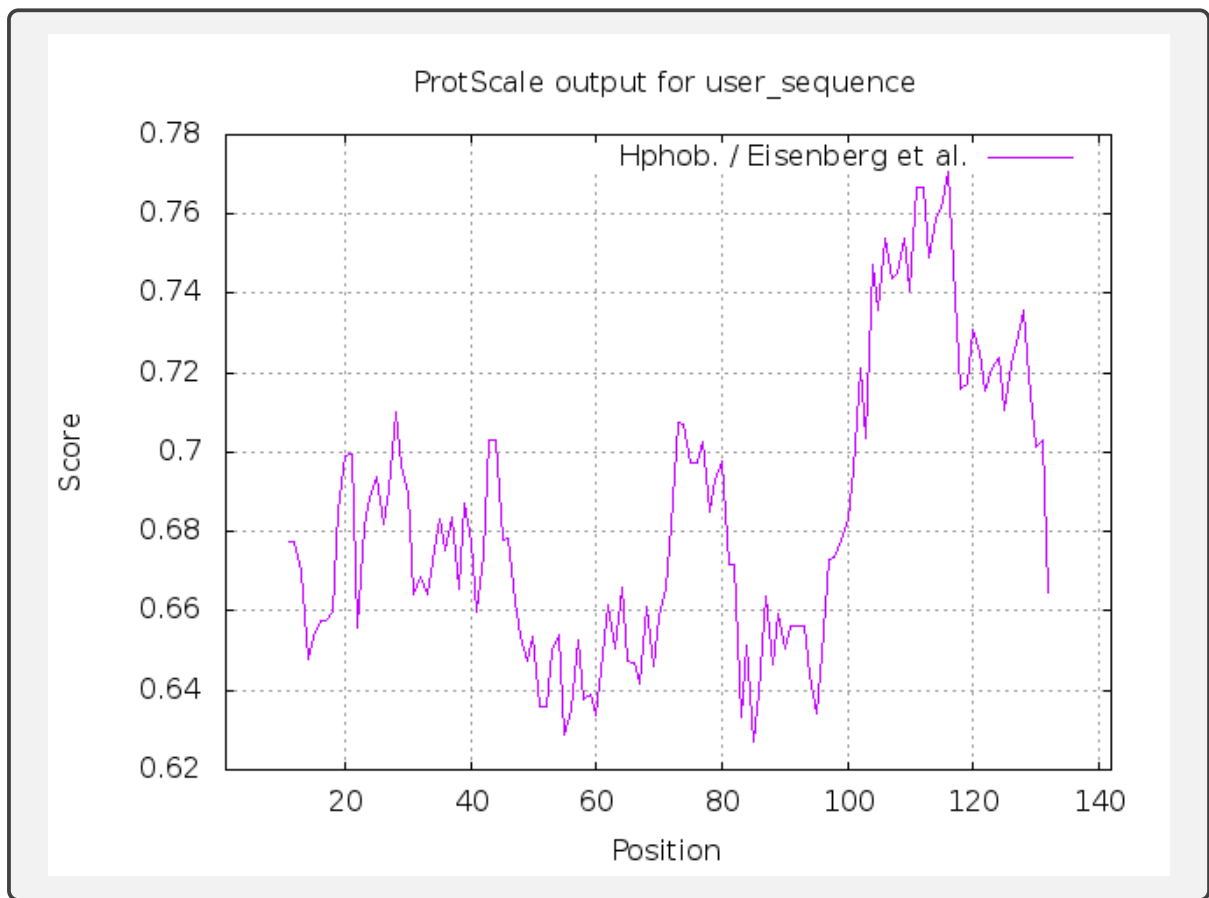
Overall, for the random protein sequence, a good guess would be to say that there are hydrophobic regions between amino acids 80-100 and 130-150 and there might be extra regions between 40-60 and 170-180. Note that the Hopp & Woods scale is inverted because it predicts hydrophilicity instead of -phobicity! It's a lot harder to find hydrophobic regions that are reproduced in a consistent way by most predictors, but a good guess would probably be: 80-100 and 130-150.

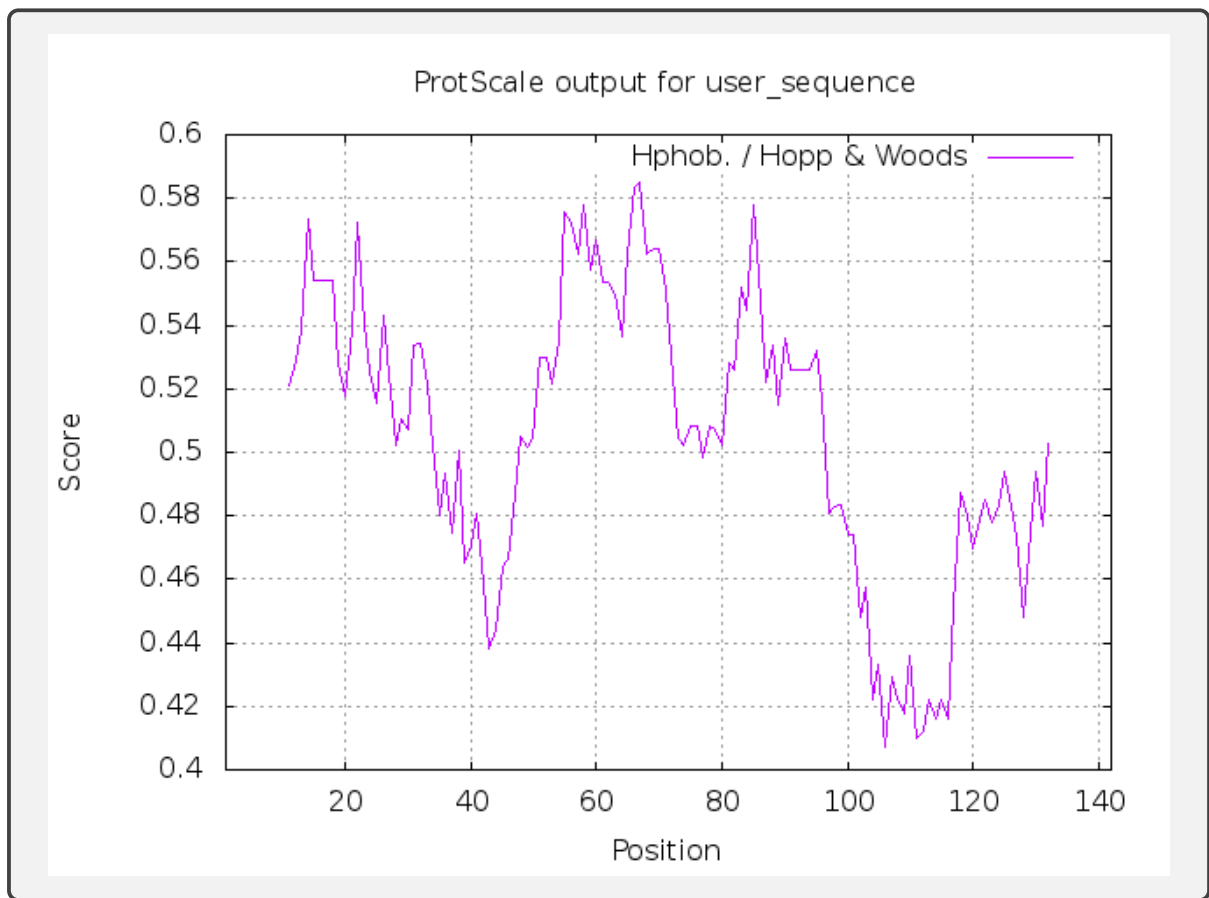
Hydrophobicity plots - Hemoglobin subunit alpha:

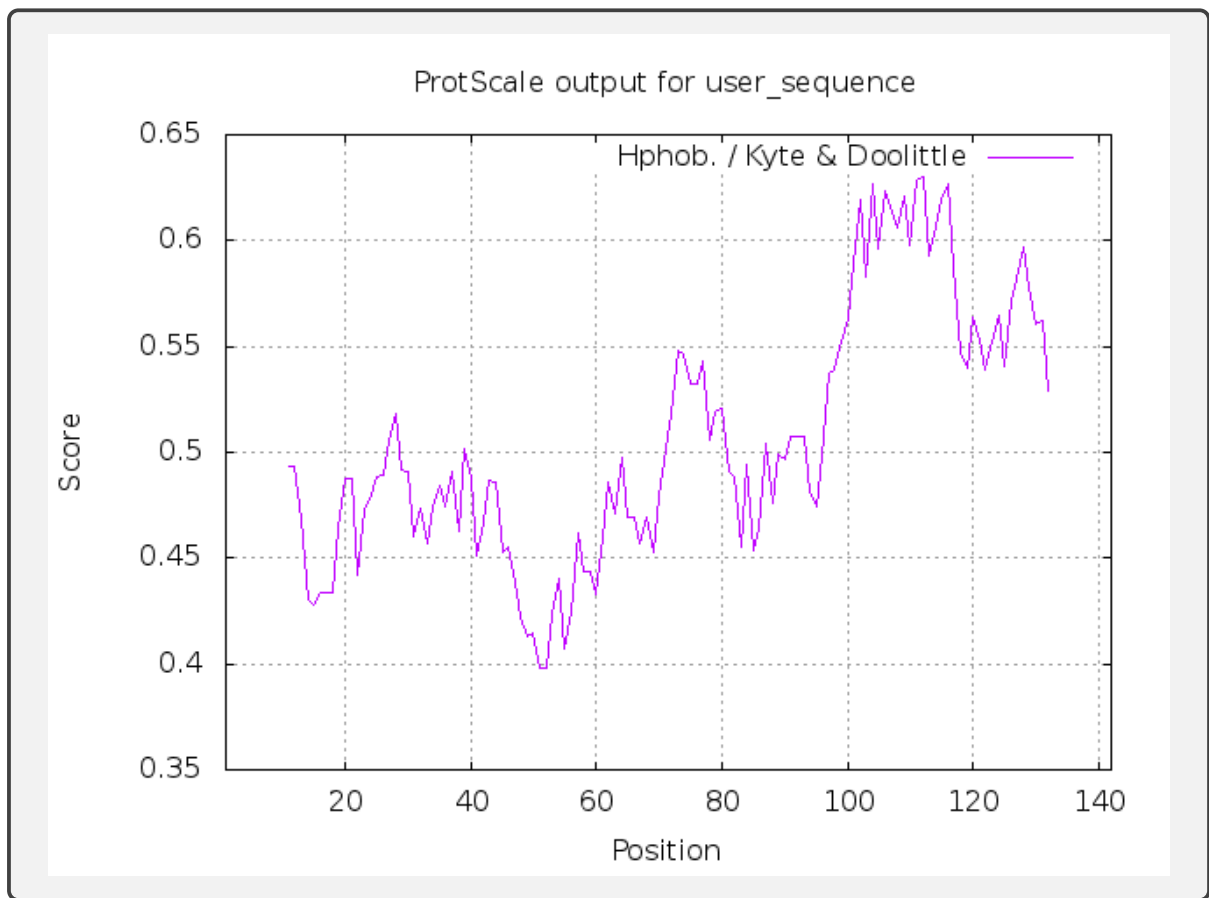




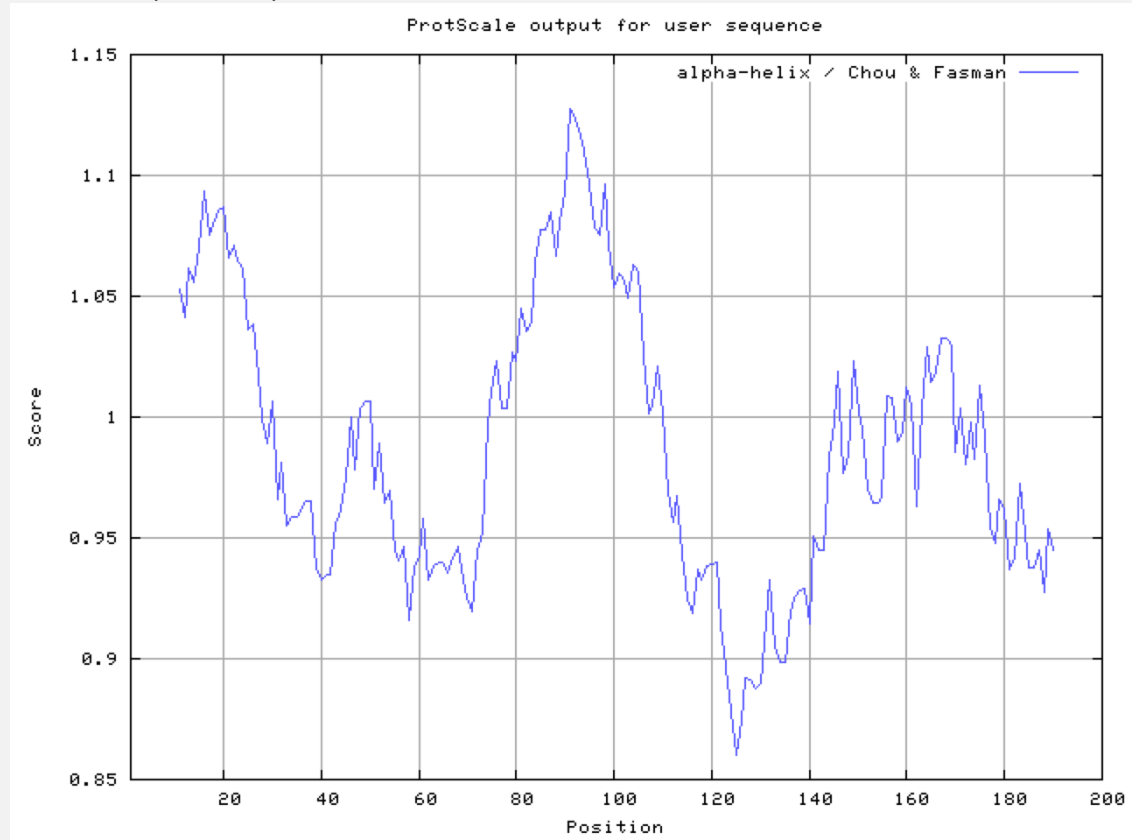


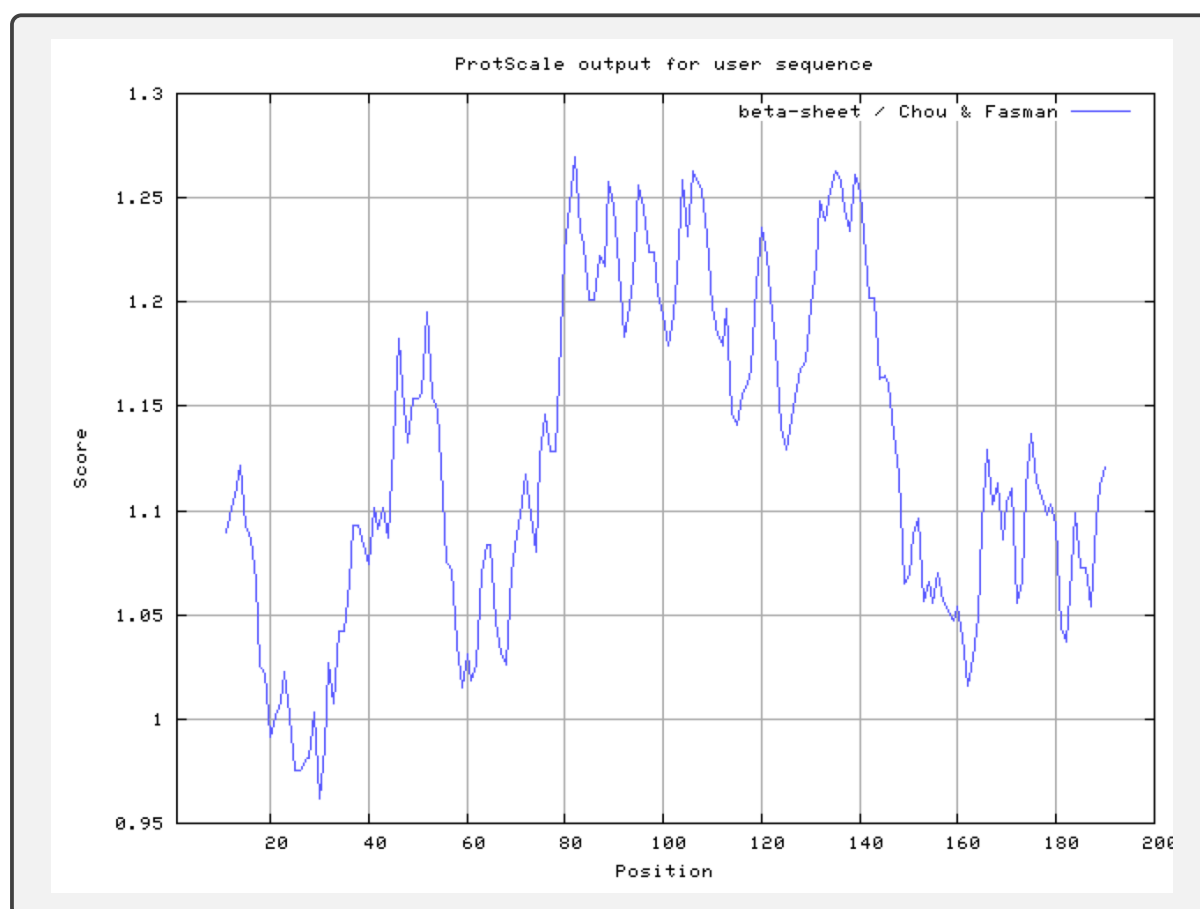




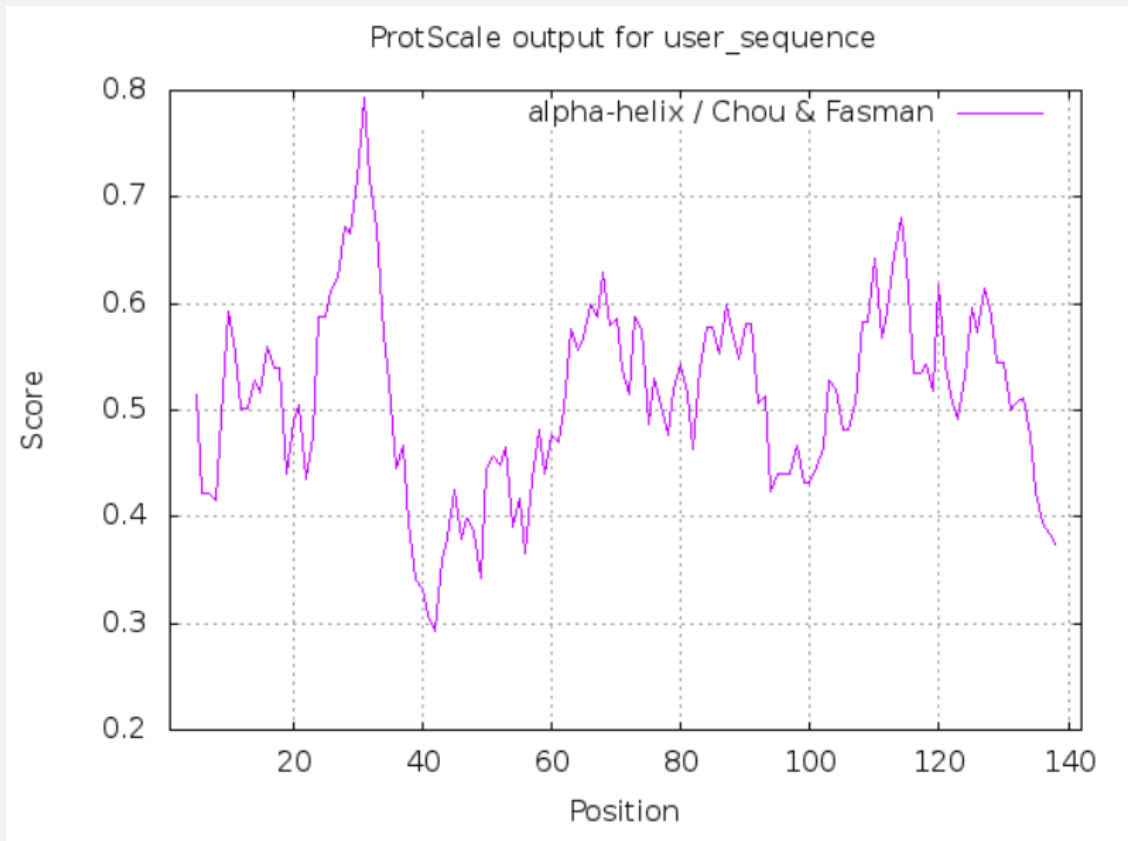


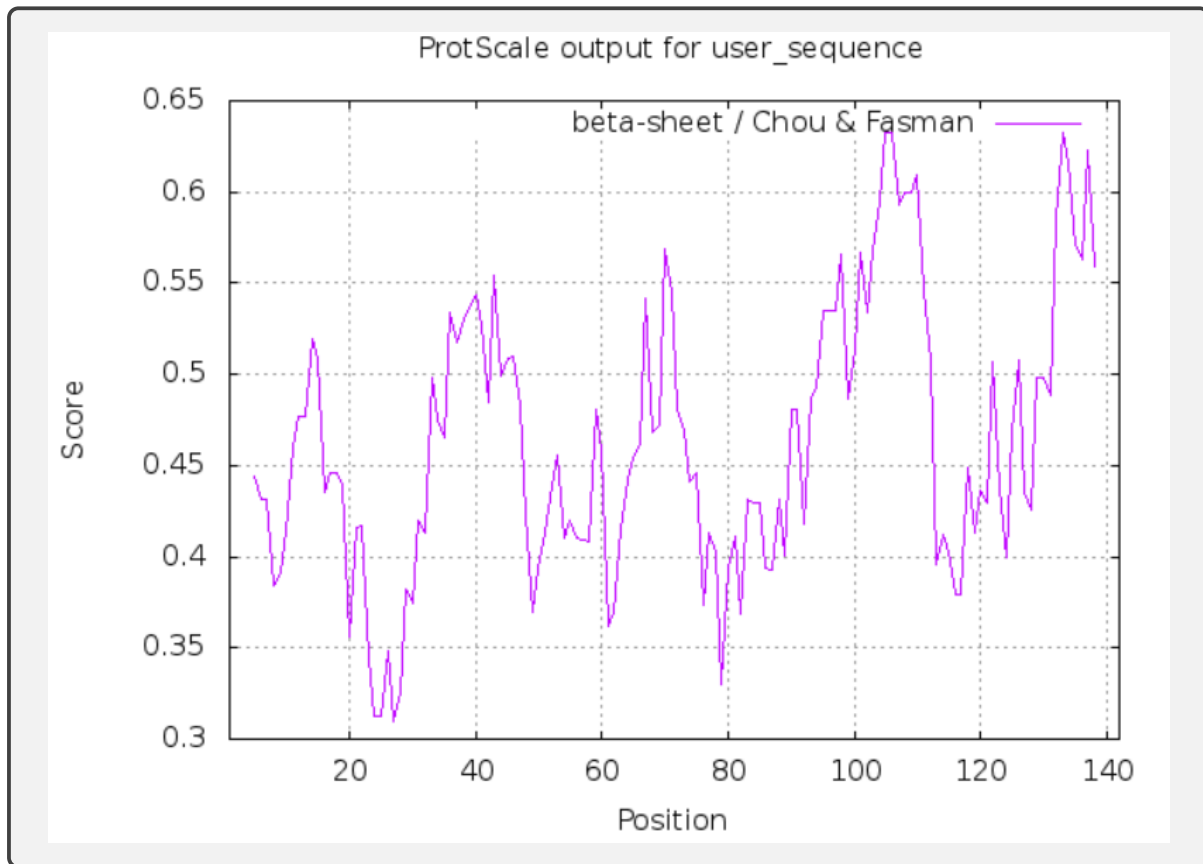
Random sequence: alpha helices - beta sheets:





Hemoglobin subunit alpha: alpha helices - beta sheets:



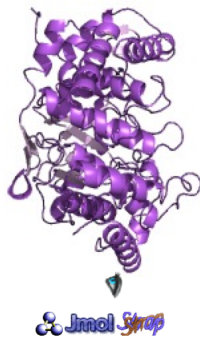


### Exercise 7.2: Secondary structure prediction

A) Although a lot of information can already be gained from the primary protein sequence, it is increasingly harder to predict higher order structural features. However, the prediction of secondary protein structure elements is still a feasible task. In exercise 7.1 we predicted some structural elements with a relatively simple method (Chou & Fasman). Jpred4 (<http://www.compbio.dundee.ac.uk/jpred/>) uses more sophisticated algorithms to predict the secondary structure of proteins. It consists of 2 different ways to predict the protein structure: a blast search to infer the structure from known related proteins, or a de novo prediction. We will predict the secondary structure elements for both sequences in both ways. Its use is very simple, just paste the sequence into the submission box and hit the "Make prediction" button! (Note: don't forget to uncheck the "Check to skip" box under "Advanced options". Otherwise you will not see the PDB results!)

Jpred will first use blast to try to find similar structures in the PDB database which may represent a more accurate assignment of secondary structure elements to the primary protein sequence than a new prediction might. Open the top hits for both sequences in a new tab, and explore the output.





Jmol

Contents

Protein chain

457 a.a.

Waters x463

PDB id: 1nth

Name: **Transferase**

Title: Crystal structure of the methanosarcina barkeri monomethylam methyltransferase (mtmb)

Structure: Monomethylamine methyltransferase mtmb1. Chain: a. Synonym: mma methyltransferase 1, mmamt 1. Ec: 2.1.1.-

Source: Methanosarcina barkeri. Organism\_taxid: 2208. Strain: ms

Biol. unit: Hexamer (from PDB file)

Resolution: 1.55Å R-factor: 0.176 R-free: 0.188

Authors: B.Hao,W.Gong,T.K.Ferguson,C.M.James,J.A.Krzycki,M.K.Chan

Key ref: B.Hao et al. (2002). A new UAG-encoded residue in the structure of a methanogen methyltransferase. *Science*, 296, 1462-1466. PubMed id: 12029132

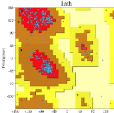
DOI: 10.1126/science.1069556

Date: 30-Jan-03 Release date: 04-Feb-03

Supersedes: 1l2r

Links

PROCHECK



Headers

References

Protein chain

O30642 (MTMB1\_METBA) - Monomethylamine methyltransferase MtmB1

Seq: 458 a.a.

Struc: 457 a.a.

Key: Family PfamA domain Secondary structure CATH domain

\* PDB and UniProt seqs differ at 1 residue position (black cross)

(Enlarge the secondary structure view for the top hits in PDBsum by clicking on the magnifier symbol on the right of Protein Chain view in the red box)

Quality of the blast hit can be explored under the "Click to show/hide details" button in the main page.

If we now hit the "Continue" button on the top of the page, Jpred will predict the secondary structure itself. The Jpred prediction will return a visual representation of the structural elements it has found. A red bar indicates an alpha-helix while a green arrow represents a beta-sheet, the height of the black bars represents the confidence of the prediction. Compare the results of the new Jpred prediction with the secondary structure elements assigned to the top hits and the predictions made with the Chou & Fasman method.

The Jpred tool first tries to search in databases for similar structures. Secondary structure elements from known, similar proteins are often a reliable way of extrapolating secondary structure features. However, in this case the best hit doesn't resemble our protein sequence at all! When we align the sequences of both our query and the hit (click 'show details' under 'Alignment of PDH hits to your sequence'), we see that they only have a few matching amino acids at the beginning of the randomly generated sequence and the end of its top hit (MTMB1\_METBA). Therefore, the "*similar*" hit doesn't make a good template at all, as can be seen in largely differing secondary structure elements.

```
>1nth_A mol:protein length:458 Monomethylamine methyltransferase mtmB1
      Length = 458
```

```
Score = 52.0 bits (123), Expect = 1e-06
Identities = 8/34 (23%), Positives = 13/34 (38%), Gaps = 3/34 (8%)
```

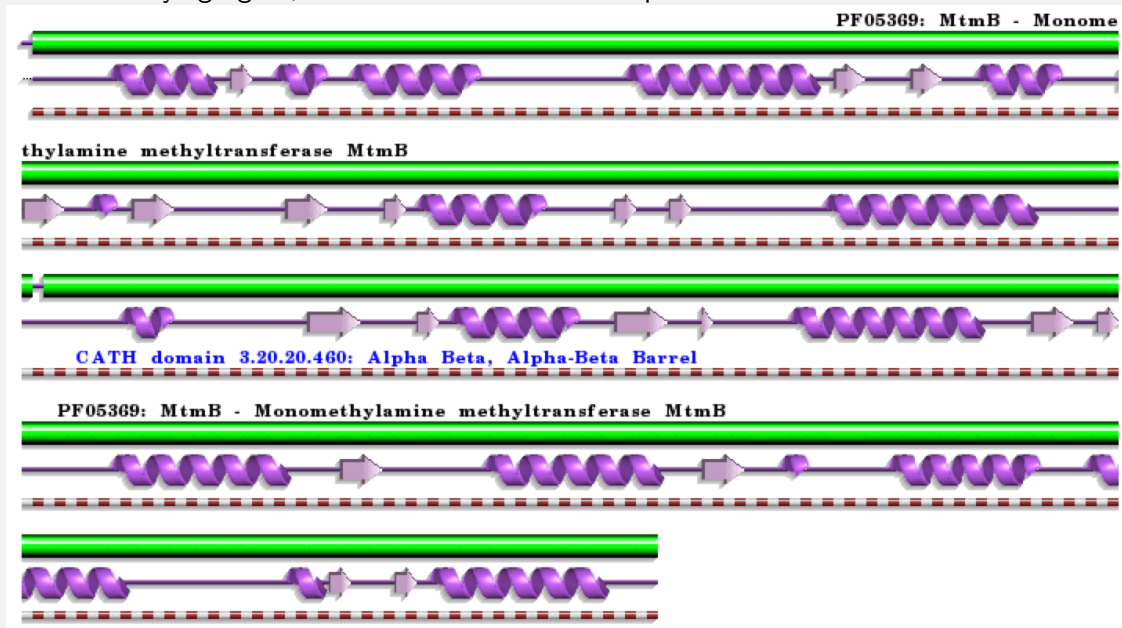
```
Query: 152 PNERFPLEYVYDGFMTLINPALLVAGKG---IP 182
          P + P E +      LI A +AG+      +
Sbjct: 169 PIPKSPYEVLAAKTETRLIKNACAMAGRPGMGV 202
```

```
>112q_A mol:protein length:458 monomethylamine methyltransferase
      Length = 458
```

```
Score = 52.0 bits (123), Expect = 1e-06
Identities = 8/34 (23%), Positives = 13/34 (38%), Gaps = 3/34 (8%)
```

```
Query: 152 PNERFPLEYVYDGFMTLINPALLVAGKG---IP 182
          P + P E +      LI A +AG+      +
Sbjct: 169 PIPKSPYEVLAAKTETRLIKNACAMAGRPGMGV 202
```

In the underlying figure, 100 amino acids are shown per line.



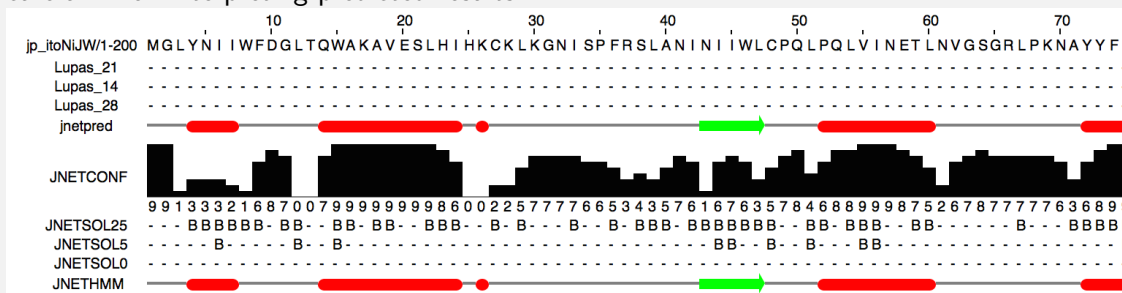
The Jpred prediction for the first 80 amino acids of the *random protein sequence* are:

helix - helix - sheet - helix - helix

while the "*similar*" hit has:

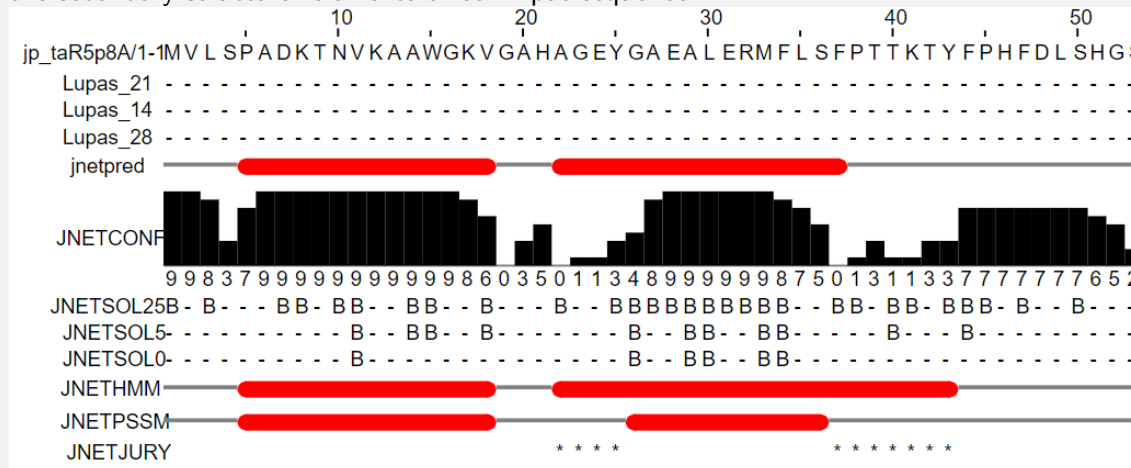
helix - sheet - helix - helix - helix - sheet - sheet

Due to the aforementioned problems with sliding window predictions, it is very hard to compare the predictions of Jpred with those from the Chou & Fasman method. As always, be very careful when interpreting predicted results!



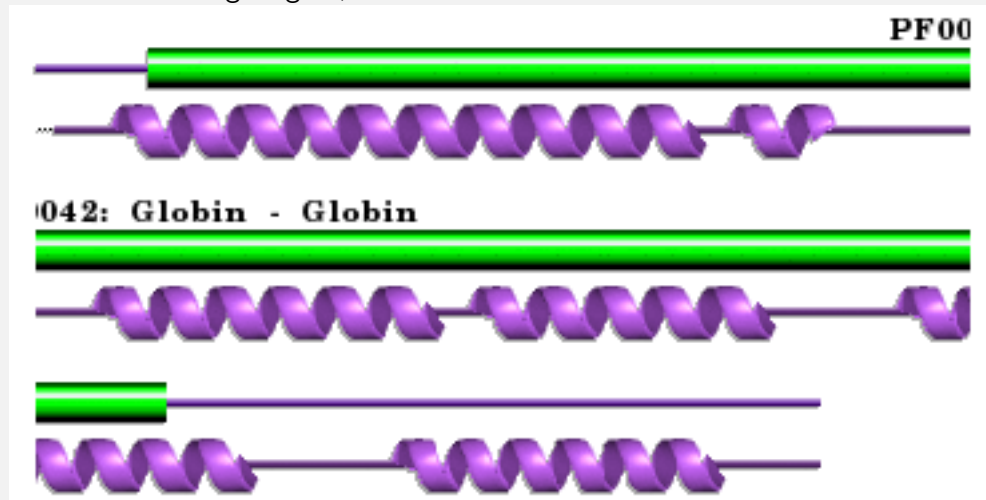
The first 80 amino acids are shown in the above sequence prediction viewer.

For the hemoglobin subunit alpha sequence, the Jpred search for similar hits was a lot more sensible, the top hit is the actual protein. This provides us with a great way of checking how well Jpred's predictions are for this protein. As it turns out, Jpred did a great job at predicting the secondary structural elements of our input sequence.



The first 50 amino acids are shown in the above sequence prediction viewer.

In the following figure, 50 amino acids are shown for the actual protein.



**B)** Very often, the secondary structure of a protein is already known. This information can also be found on UniProt. Go to feature viewer in the UniProt entry to find more information about its secondary structure. Compare the results of the Jpred prediction with the UniProt entry for the hemoglobin protein.

The UniProt entry contains the same secondary structure information as the PDBsum entry, only the beta strands don't match. Therefore, it also compares well with the Jpred predictions.



### Exercise 7.3: Viewing 3D protein structures

**A)** The Protein Data Bank (PDB) can be accessed from <http://www.rcsb.org/pdb/home/home.do>. It contains a large collection of experimentally determined 3D protein structures. You can query the database for protein structures by typing in their name.

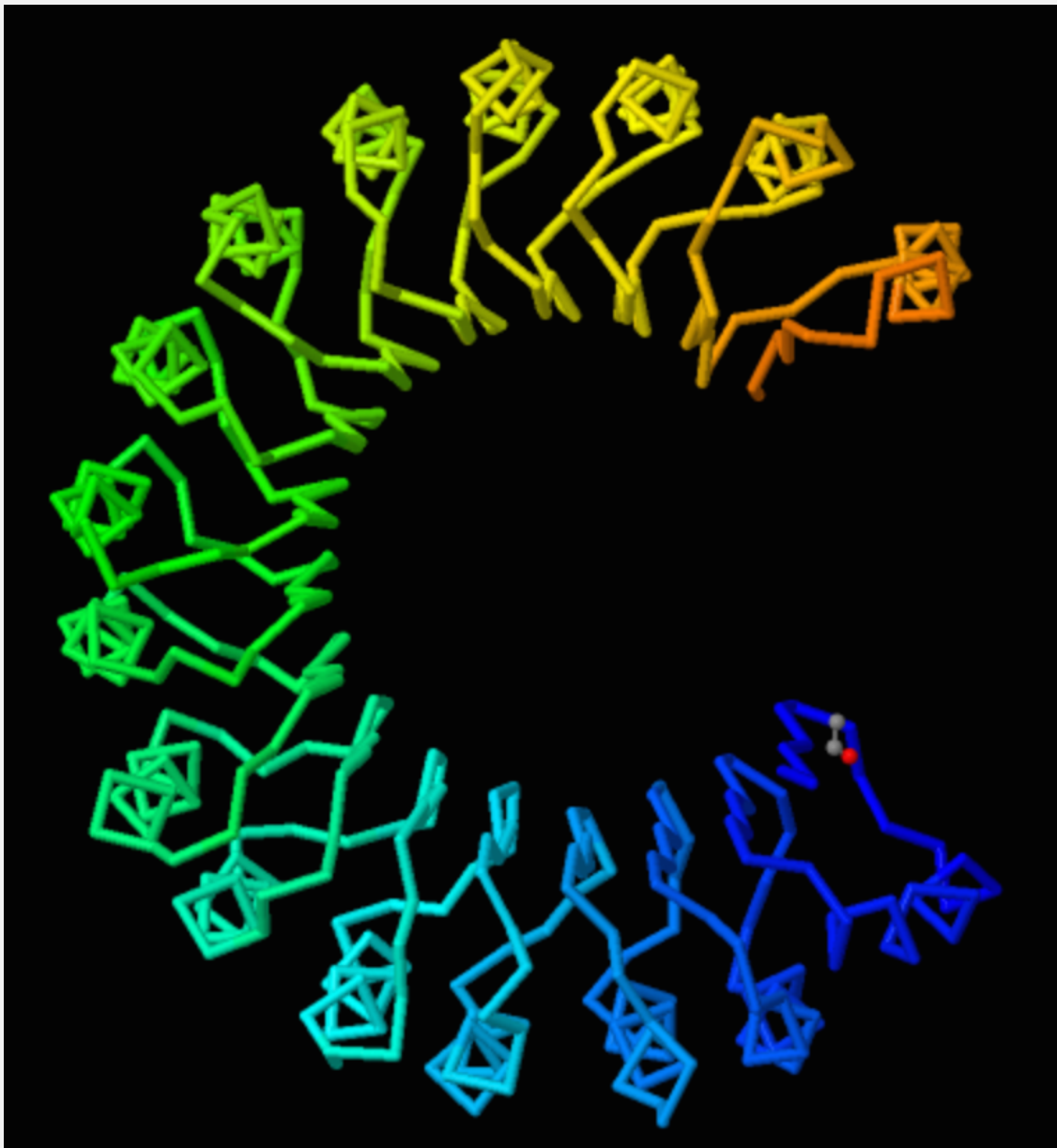
Try looking up the structure of the porcine ribonuclease inhibitor by typing "porcine ribonuclease inhibitor" in the search box and clicking on the only entry that shows up (2BNH). The PDB entry on this protein returns a number of pages containing information on the protein structure, how it was determined, structural annotations, . . . . Click on the 3D view tab and wait for the application to load, then select the **JSmol** viewer on the bottom of the page. Using the display options, make a screenshots from the front and side of the protein coloured by (1) structural elements in cartoon style and (2) rainbow in backbone style. Play around with some of the other options to get a feel of the possibilities available to view the protein structure.



Cartoon, secondary structure view, front side.

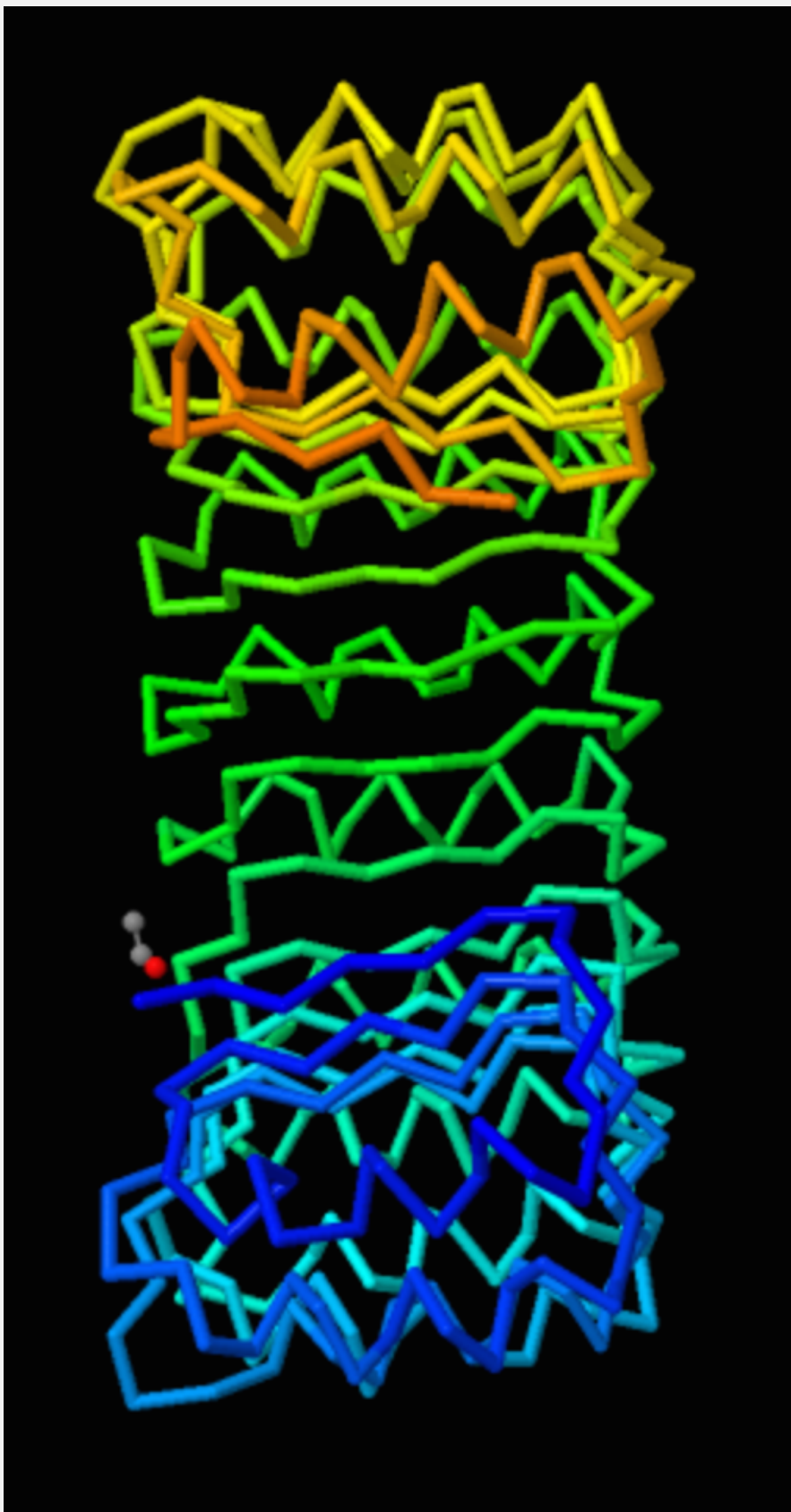


Cartoon, secondary structure view, right side<sub>31</sub>

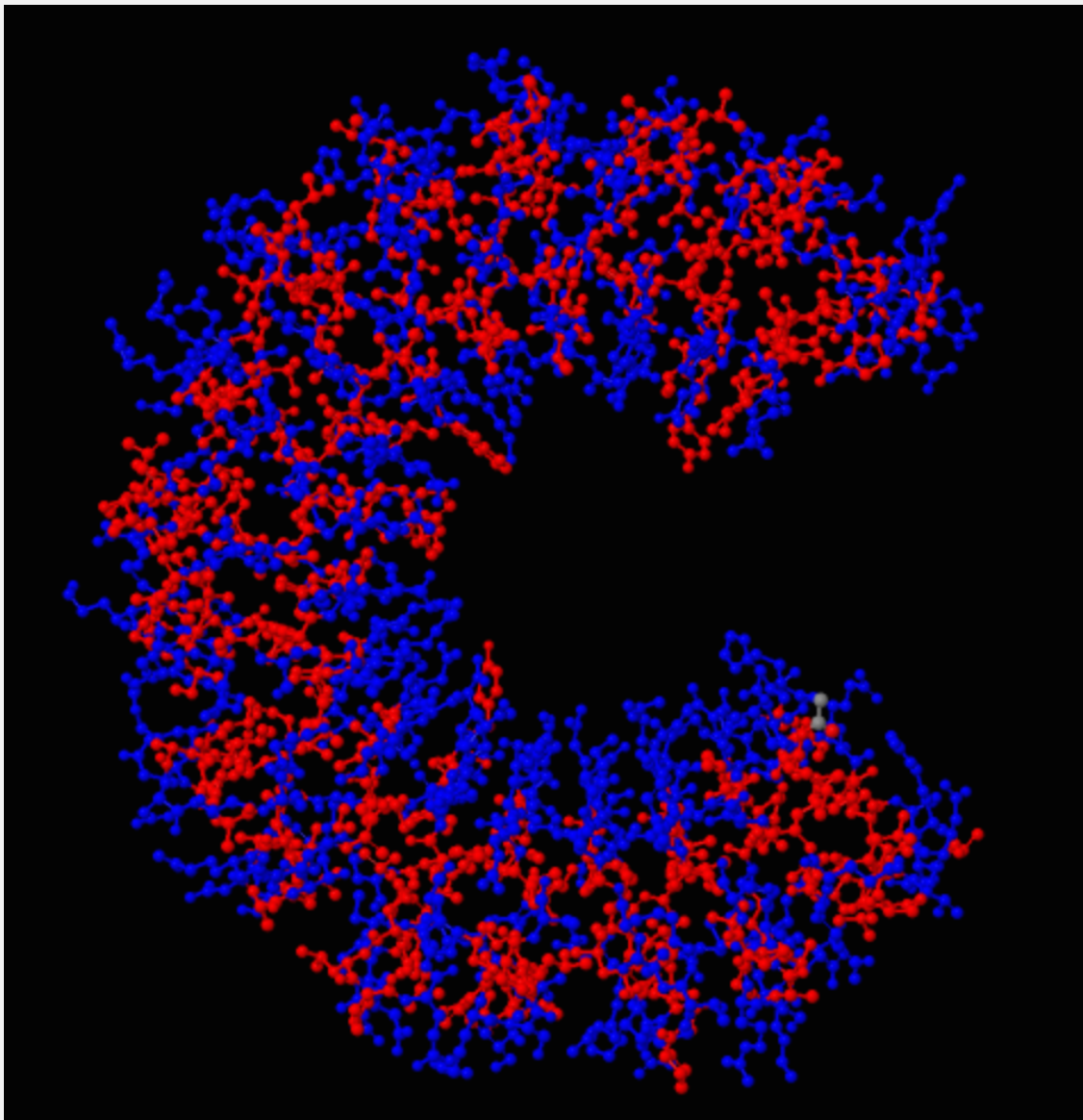


Backbone, rainbow view, front side.





Backbone, rainbow view, right side.



Front view (balls & sticks, hydrophobicity coloured.

## Section 8: Gene ontology

Up to this point, we have learned how to extract information from certain biological sequences of interest. It is important to have a resource that describes and summarises all this information in a structured way. This resource is provided by the gene ontology consortium (more information: <http://geneontology.org/docs/ontology-documentation/>).

### Learning goals

At the end of this section, you should:

- understand why gene ontology is important to understand the function of genes
- have a practical understanding on how gene ontology is used in contemporary research

Gene ontology (GO) annotations are often used to further investigate the outcome of large-scale gene expression experiments comparing different conditions. For example, suppose we have compared the gene expression levels of diseased patients with a number of healthy controls. The output of such an experiment would be a *very* long list of gene names along with an indication of whether they were up- or down-regulated between the groups. But how do we find out what's interesting about these affected genes? Perhaps the majority of them belong to a specific pathway, function, etc. This is termed an *enriched* gene set. In other words, we want to retrieve a functional profile of the differentially expressed genes that helps us understand the underlying biological processes. This is done by checking which GO terms are over- or under represented in the list of differentially expressed genes.

We'll showcase one simple method to analyse whether a specific type of genes (a set) is overrepresented in a larger list of genes.

### Exercise 8.1: GO enrichment

On Blackboard you can find a text file with a list of genes (8-1-GO-enrichment-hgnc-genes.txt) that were up- or down regulated between two groups of people. Use this list as an input to the GO Enrichment Analysis tool on: <http://geneontology.org>. Choose biological process as the type of GO annotation and *Homo sapiens* as the subject organism.

- Look at the different kinds of GO processes that are enriched in the gene set. Does it seem likely that a specific condition or disease was studied in the original experiment? Can you make a more specific guess? You can click on the question mark next to "Results" if you want more information on how to interpret the output table.

After sorting on "False discovery rate (FDR)", you will obtain the following list:

	Homo sapiens (REF)		upload_1 (Hierarchy) <b>NEW!</b> (?)			
GO biological process complete	#	#	expected	Fold Enrichment	+/-	raw P value
<a href="#">response to chemical</a>	<a href="#">3643</a>	<a href="#">79</a>	21.60	3.66	+	1.08E-30
<a href="#">response to xenobiotic stimulus</a>	<a href="#">431</a>	<a href="#">34</a>	2.56	13.31	+	7.54E-29
<a href="#">cellular response to chemical stimulus</a>	<a href="#">1908</a>	<a href="#">59</a>	11.31	5.22	+	5.23E-29
<a href="#">response to oxygen-containing compound</a>	<a href="#">1507</a>	<a href="#">51</a>	8.93	5.71	+	2.53E-26
<a href="#">response to toxic substance</a>	<a href="#">247</a>	<a href="#">25</a>	1.46	17.07	+	6.64E-24
<a href="#">response to stimulus</a>	<a href="#">8080</a>	<a href="#">101</a>	47.90	2.11	+	5.90E-23
<a href="#">response to stress</a>	<a href="#">3411</a>	<a href="#">67</a>	20.22	3.31	+	4.09E-22
<a href="#">dopamine metabolic process</a>	<a href="#">31</a>	<a href="#">13</a>	.18	70.74	+	1.09E-21
<a href="#">response to oxidative stress</a>	<a href="#">362</a>	<a href="#">26</a>	2.15	12.12	+	5.18E-21
<a href="#">behavior</a>	<a href="#">618</a>	<a href="#">31</a>	3.66	8.46	+	2.09E-20
<a href="#">cellular response to xenobiotic stimulus</a>	<a href="#">193</a>	<a href="#">20</a>	1.14	17.48	+	1.87E-19
<a href="#">locomotory behavior</a>	<a href="#">197</a>	<a href="#">20</a>	1.17	17.13	+	2.83E-19
<a href="#">catechol-containing compound metabolic process</a>	<a href="#">47</a>	<a href="#">13</a>	.28	46.66	+	6.87E-19
<a href="#">catecholamine metabolic process</a>	<a href="#">47</a>	<a href="#">13</a>	.28	46.66	+	6.87E-19

The results of the gene ontology analysis show you multiple categories linked to the neural system. A good hypothesis would be that the gene expression data is derived from neuronal cells of a patient with a neurodegenerative disease like Parkinson's disease.

Alternatively, you could have selected "PANTHER Pathways" in stead of "GO biological process complete". The result will show you that around 8 genes are part of the "Parkinson disease" pathway.

- Filter your results on FDR (false discovery rate). Your attention should be drawn by the category dopamine metabolic process; it shows a low false discovery rate, and a very high fold enrichment. You can further inspect the 13 genes of this category that were in your list.

Some of the genes in the list, like Alpha-synuclein, are actively being investigated for their role in neurodegenerative diseases like Parkinson's disease.

## Exercise 8.2: Searching the GO database (GO terms)

In order to know more about the GO category dopamine metabolic process, browse to <http://www.ebi.ac.uk/QuickGO/>, and search for dopamine metabolic process. By clicking on the GO annotation (left) or number of annotated genes (blue oval), you can retrieve some more information about this GO term.

- What is the definition of this GO term? And which GO terms can you find as a child of this term? How many proteins are associated with this GO term?

Searching for "dopamine metabolic process" via QuickGo returns a list of related terms.

## Terms

GO:0042417

P

dopamine metabolic process

13,261 annotations

GO:0042053

P

regulation of dopamine metabolic process

973 annotations

GO:0045963

P

negative regulation of dopamine metabolic process

30 annotations

GO:0045964

P

positive regulation of dopamine metabolic process

313 annotations

GO:0042416

P

dopamine biosynthetic process

4,774 annotations

GO:0042420

P

dopamine catabolic process

5,613 annotations

GO:0015872

P

dopamine transport

7,182 annotations

GO:0090494

P

dopamine uptake

6,005 annotations

GO:0014046

P

dopamine secretion

482 annotations

GO:0035240

F

dopamine binding

1,719 annotations

GO:0008152

P

metabolic process

160,683,423 annotations

GO:0006585

P

dopamine biosynthetic process from tyrosine

3,149 annotations

GO:1903179

P

regulation of dopamine biosynthetic process

221 annotations

GO:0010586

P

miRNA metabolic process

3,886 annotations

GO:0061527

P

dopamine secretion, neurotransmission
















Show all 7,027 results

**Definition:** The chemical reactions and pathways involving dopamine, a catecholamine neurotransmitter and a metabolic precursor of noradrenaline and adrenaline.

The GO-term "dopamine metabolic process" has 5 child terms, as can be seen in the picture below.

### Child Terms

This table lists all terms that are direct descendants (child terms) of GO:0042417

Child Term	Relationship to GO:0042417
GO:0045964    positive regulation of dopamine metabolic process	positively_regulates
GO:0042416    dopamine biosynthetic process	is_a
GO:0042053    regulation of dopamine metabolic process	regulates
GO:0045963    negative regulation of dopamine metabolic process	negatively_regulates
GO:0042420    dopamine catabolic process	is_a

To determine how many proteins are associated with this GO term, click on "13261 annotations" and go to "statistics", the table is automatically filtered on proteins. 2839 distinct proteins are annotated with this GO-term.

- Using the filter options, how many *mouse* proteins can you find with *manual experimental evidence* that are linked to this GO term? How many GO terms are associated with these proteins?

After some additional filtering: number of *mouse* proteins with *experimental evidence used in manual assertion* that are linked to the GO term "dopamine metabolic process": 121 annotations and 101 distinct proteins originating from the mouse have been annotated with "dopamine metabolic process" based on manual, experimental evidence.

### Exercise 8.3: Searching the GO database (protein)

QuickGo also allows to search for a protein. It will then return all the terms annotated to this protein.

- On the homepage of QuickGo, search for one of the proteins that was overrepresented in the first part of this exercise, such as Sodium-dependent dopamine transporter. Scroll to the "Gene products" section of the page, choose an entry and click on 'annotations'. Just by looking at the related GO terms, you will learn more about this protein.

The Gene products section of the search results gives many hits that look very similar at first. Select "Show all results" on the bottom and then select Swiss-Prot. You will see that these are all hits for different species, select the Human variant (by clicking on the '100 annotations') to continue.

Some conclusion you can make: the sodium-dependent dopamine transporter is involved in protein binding (GO molecular function: protein binding), can be found in the plasma membrane (GO cellular component: plasma membrane), and enables the transport of dopamine (GO biological process: dopamine transport).

- How many distinct GO annotations are associated with this protein? By using: Statistics > GO Identifier, you can return a non-redundant list of GO-terms.

Around 58 distinct GO annotations are associated with the sodium-dependent dopamine transporter protein.