Assignment 2 - Sequence-based predictions

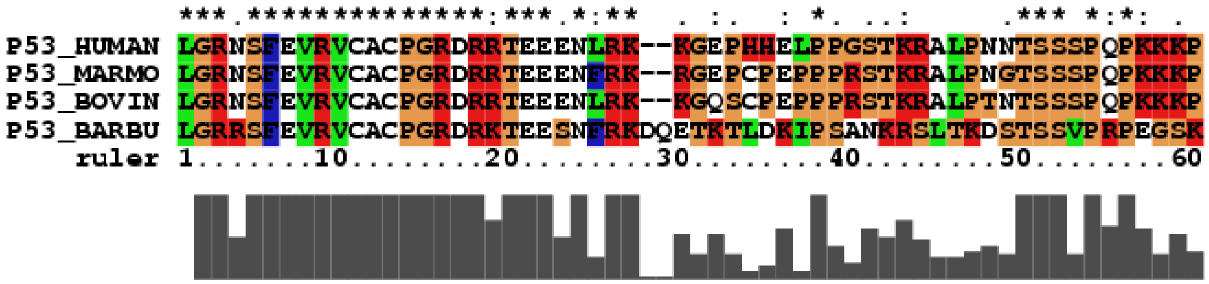
This assignment is heavily based on a previous one made by Bengt Persson, LiU

Introduction

This assignment assumes some familiarity with amino-acids and their one-letter codes.

Domains are the "building blocks" of proteins. They are regions that fold independently and are often interconnected by flexible linker regions. In general, each domain is associated with a distinct function, for example hydrophobic membrane-spanning domains, cofactor-binding domains, and catalytic domains.

To study relationships between proteins, a common first step is to compare the amino acid sequences using a multiple sequence alignment (MSA). A small example is shown here:



```
            ***.**************:***.*:**  . :.  : *. ..:     .*** *:*:  .
 P53_HUMAN  LGRNSFEVRVCACPGRDRRTEEENLRK--KGEPHHELPPGSTKRALPNNTSSSPQPKKKP
 P53_MARMO  LGRNSFEVRVCACPGRDRRTEEENFRK--RGEPCPEPPPRSTKRALPNGTSSSPQPKKKP
 P53_BOVIN  LGRNSFEVRVCACPGRDRRTEEENLRK--KGQSCPEPPPRSTKRALPTNTSSSPQPKKKP
 P53_BARBU  LGRRSFEVRVCACPGRDRKTEESNFRKDQETKTLDKIPSANKRSLTKDSTSSVPRPEGSK
     ruler  1.......10........20........30........40........50........60
```

From the MSA it is possible to determine the conserved regions in the proteins. Proteins often contain clusters of residues that are conserved because of particular requirements on their interactions, either internally in the protein or with the environment. Usually, these residues are of functional importance to the protein, for example for the binding properties or the catalytic activity. The function and structure of proteins can thus be characterized by their conserved sequence motifs, and this can be automated using various computational methods, e.g. machine learning techniques.

In this assignment we will first get acquainted with patterns, which is the most basic method for sequence motif recognition. Then we will expand our view to the more advanced profiles, after which we will move on to the powerful statistical hidden Markov models (HMM), which represent one of the most sensitive classification methods that exist today. Finally, we will look at some tools for predicting other protein features.

Patterns

A pattern is usually a number of consecutive residues important for a specific biological function. These regions include binding sites or enzymatic catalytic sites. Here is one example:

```
[AG]-x-C-x(4)-{DE}
```

This pattern is translated as: Ala or Gly, any, Cys, any, any, any, any, anything but Glu or Asp. Prosite is a well used resource that contains a database of patterns and profiles and also provides web based tools that allows users to analyze proteins online. Use the tools in Prosite to scan the protein RON_HUMAN.

2-1. Does your protein have any known domains? If so, which?

A: Two domains are found, a SEMA domain and a protein kinase domain

2-2. Which pattern matches the ATP binding region? Give both the ID and the actual pattern!
Hint: To see the details for a pattern, click the link next to the pattern ID (on the form PS00000).

A: *PS00107* ***PROTEIN_KINASE_ATP*** Protein kinases ATP-binding region signature

Consensus Pattern: ***[LIV]-G-{P}-G-{P}-[FYWMGSTNH]-[SGA]-{PW}-[LIVCAT]-{PD}-x-[GSTACLIVMFY]-x(5,18)-[LIVMFYWCSTAR]-[AIVP]-[LIVMFAGCKR]-K***
K binds ATP

2-3. Describe how sequences can be matched to this pattern!

Hint: On the documentation page you have the option to "Retrieve an alignment of Swiss-Prot true positive hits", which may help. The Prosite user manual may also help on this and subsequent questions.

A: Using the consensus sequence we can match sequences for this pattern.
[(Leu or Ile or Val)- Gly - anything but Pro – Gly - anything but Pro – (Phe or Tyr or Trp or Met or Gly or Ser or Thr or Asn or His) – (Ser or Gly or Ala) - anything but Pro or Trp – (Leu or Ile or Val or Cys or Ala or Thr) - anything but Pro or Asp - any – (Gly or Ser or Thr or Ala or Cys or Leu or Ile or Val or Met or Phe or Tyr) – any amino acid from five to eighteen positions - (Leu or Ile or Val or Met or Phe or Tyr or Trp or Cys or Ser or Thr or Ala or Arg) – (Ala or Ile or Val or Pro) – (Leu or Ile or Val or Met or Phe or Ala or Gly or Cys or Lys or Arg) – Lys]

2-4. How do you determine whether or not a sequence matches a pattern?

A: Using a database like PROSITE we can get all the possible sequences for a particular pattern. Then we can compare sequences with the pattern for different amino acids in the same location. All amino acids have to meet the restrictions of the pattern.

2-5. Patterns are made from MSAs. Give a plausible explanation as to how this is done!

A: If we have two sequences, one with known function and one unknown, and these show similarities limited to only a few residues, we might need to create a pattern which will give us a "model" for our query sequences. This is done by aligning multiple sequences within a protein family which we know are conserved. Alignment and subsequent matching will create a model that in turn can be used to find more members of this family containing conserved sequences. The pattern is then based on the amino acids within these members. However, this may exclude atypical proteins which might be homologous while containing modifications acquired during evolution which the other protein members didn't acquire.

2-6. Which position in the pattern binds ATP? Answer using a number!

`A:` Position 1114, Lysine (K) in the protein, and the last position in the pattern (K), depending on how many repeats there are at position 13, binds ATP. If we have 5 repeats at position 13, the position (in the pattern) where ATP binds would be at 21.

Consider the following two hypothetical patterns that describe the same conserved region:

`[FW]-C-x (3)-C-[AG]-E-[MLI]-D` and

`[FW]-C-x(3)-C-[ASGPT]-E-[IVLM]-D`

2-7. Which of the patterns is more tolerant?

> *A: The second one, `[FW]-C-x(3)-C-[ASGPT]-E-[IVLM]-D`*

2-8. What are the consequences of using a more tolerant pattern in searches?

> *A: A more tolerant pattern will give more hits, as well as a higher sensitivity but lower selectivity.*

Profiles

Profiles are more sensitive and more robust than patterns and can therefore be used to describe larger sequence features, such as for example domains. Profiles are also called position specific scoring matrices (PSSM), and as the name suggests, they consist of matrices of numbers which are used to score sequences based on which amino acid residue they have where.

Now, getting Prosite to show details on profiles is a bit trickier. Go back to the Prosite scan summary and bring up the documentation page for the first matching domain. On this page, once again click on the link next to the profile ID (should have the exact same text as the one you just clicked, but will not bring you to the same page). If everything went OK, you should now be looking at a page with lots and lots of numbers on it.

2-9. Given that "/M" means match and "/I" means insertion, describe how sequences are matched to this profile!

> *A: The rows correspond to the AA positions of the profile to which a query sequence should be aligned. The columns refer to the AA one letter code (including the ambiguous letters B and Z). Both thus form a matrix showing position specific substititution scores. Furthermore, penalties for matchinsertion transitions (MI) +/- further insertions (I) and match-deletion-transitions (MD) +/- further deletions (D) and internal initiation/termination are defined by default. /M defines matches in the profile sequence, I/ defines insertions in the profile sequence. Alignment of the query sequence with /M position is scored specifically according to the position specific substitution values. /I position oftentimes correspond to gap region between regions of high homology. The best alignments of profile and query sequence are calculated according to an overall score. The corresponding motif is most likely present in the query sequence, when the score exceeds the defined cutoff score.*

2-10. How do you determine whether or not a sequence matches a profile? Are there any differences compared to the pattern case, and if so, how can they be used to improve the reliability of the results?

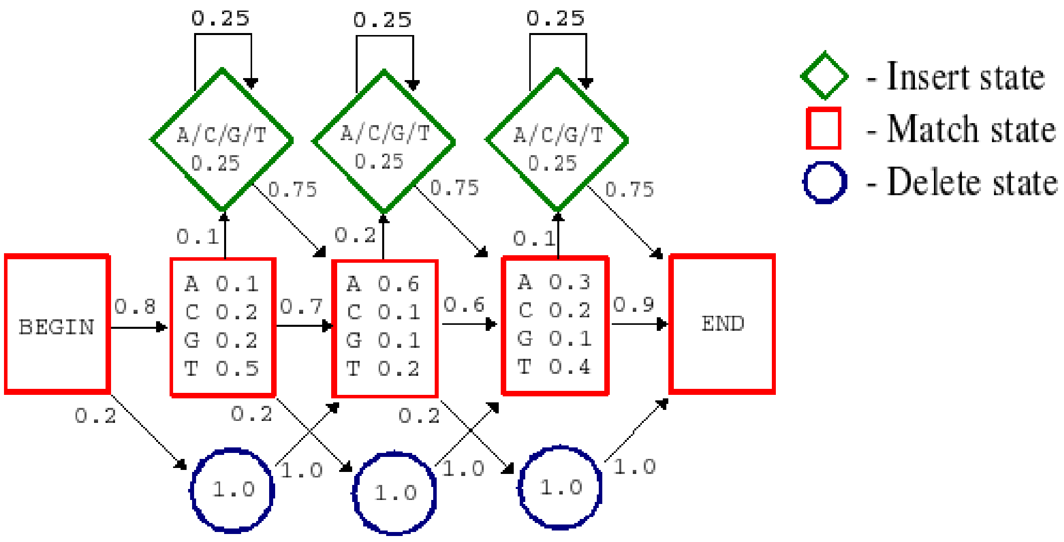
> *A: There are different levels of cut off. At lower cut off gives a score of 215 and the results over 215 can have false negative data. The higher cut off of 323 and the results above this is the false positive data. With this we can compare which sequences are similar. Insertions and deletions in the sequence cannot be recognized by the pattern.*

2-11. Profiles are made from MSAs. Give a plausible explanation to how this is done! (briefly and conceptually, max 100 words)

> *A: Profiles are made from MSA by aligning the amino acid at each position and identifying the position. Then a score is given to each amino acid at the particular position. A higher score indicates a better match. In a deletion state the match state can be omitted, and it receives a position –dependant penalty. Insertion states can also exist and it also gets a postion –dependant penalty.*

Hidden Markov models (HMM)

HMMs are in fact very similar to profiles. Both have match, insert and delete states, and both have specific ranking of amino acid residue types for individual positions. However, while profiles are an empirical attempt to generalize BLAST-like alignment scoring to individual scoring of each alignment position, HMMs have a much more solid base in mathematical statistics and are therefore more reliable. The figure below shows a very simple HMM, which is very short and also uses DNA (simply because DNA has fewer letters than proteins).



In HMMs, everything is a statistically computed probability. The profile scores have been replaced by probabilities; the general insertion and deletion penalties from profiles have been replaced by specific transition probabilities between states; there are probability distributions for the types of amino acid residue present in insertions, and so on. All this makes the HMMs very sensitive and powerful, and they can therefore be used to model very large and widely diverse groups of sequences, such as for example protein superfamilies. A less simplified illustration of an HMM (matching a protein family) is available here:

Less simplified HMM example [SAM drawmodel format, pdf].

Pfam is a database of high quality HMMs designed to reliably identify protein domains. Go to Pfam and look around. Note that Erik Sonnhammer, Stockholm Bioinformatics Center is one of the guys behind this important database. See the HELP pages and use Pfam to scan RON_HUMAN for matching HMMs.

Note: In addition to its own HMMs, Pfam also uses external tools to classify the query protein (e.g. SMART, seg, signalp, etc). For this assignment, however, we will ignore all hits that do not come from Pfam itself.

2-12. Does Pfam find any sequence features that Prosite did not? If so, which?
A: Plexin and IPT/TIG repeats were found.

2-13. Did Prosite find any sequence features that Pfam does not? If so, which?
A: No or Protein Kinase ATP binding domain with a more sensitive setting.

Clicking on a Pfam hit will take you to the corresponding documentation page, which holds both biological and technical information on the hit. For your amusement, you can click on the link "[Download HMM]" far down to the right to see what a HMM really looks like on the inside.

2-14. Using Prosite and Pfam, try to find as much information on RON_HUMAN as possible. What do you think this protein does in the cell? Which parts do what? Please, give your best theories on the protein's function, localisation, interactions, etc...

A:

	Prosite	***Pfam***
RON_HUMAN		
Description	*Macrophage-stimulating protein receptor alpha chain*	*Macrophage-stimulating protein receptor EC=2.7.10.1*
SEMA domain		
Description	*a receptor recognition and binding module found near the N-terminus of the eukaryotic and viral proteins* *Act as repulsive axon guidance cues during development or involved in immune function*	*occurs in a large family of secreted and transmembrane proteins some of which function as repellent signals during axon guidance*
Plexin repeat		
Description	*Plexins, receptors for multiple classes of semaphorins*	*A cysteine rich repeat found in several different extracellular receptors*
IPT/TIG domain		
Description	*NA.*	*cell surface receptors, intracellular transcription factors. Involved in DNA binding*
Protein tyrosine kinase		

Description	*Eukaryotic protein kinases are enzymes belonging to an extensive family of proteins. It share a conserved catalytic core common to both serine/threonine and tyrosine protein kinases. It is involved in ATP binding, catalytic activity of the enzyme*	*Protein kinases are a group of enzymes that possess a catalytic subunit which transfers the gamma phosphate from ATP to one or more amino acid residues in a protein substrate side chain, resulting in a conformational change affecting protein function.*

With this we can say that RON_HUMAN is a macrophage stimulating protein which is localized on the cell membrane. It has ATP binding activity and has SEMA and IPT/TIG domains.

One way of studying protein similarities is to search for homologues in sequence databases, using for example BLAST (compare exercise 1). Another way is to use databases that contain information on sequence patterns and protein family conservation, such as Prosite and Pfam.

2-15. Which are the advantages and disadvantages of using such sequence pattern databases compared to using databases of amino acid sequences and BLAST?

> *A: BLAST is the simpler and more straightforward approach since it searches and identifies only the sequences, such as amino acids or nucleotides in DNA. It is a basic tool. PROSITE and similar (sequence pattern) databases are however more sensitive and generates more data, and we are given more information about the protein in itself. They give us information on domains, families and functional sites, as well as the next level of amino acid sequences; their patterns, signatures and profiles within the protein. However, the usefulness of each database depends on what you want to study; BLAST might be sufficient.*

2-16. If patterns, profiles and HMMs are essentially the same thing (made from the same kind of data, used for the same purpose, etc), how come the older methods are still in use? Or to put it another way, what are the three methods advantages/disadvantages when compared to each other?

A: Patterns and profiles have some intrinsic differences. While patterns describe some short highly conserved subregions of a protein, e.g. the functional/catalytic center of a protein, profiles are used to detect domains which are usually far longer and more variable. These domains could not be detected with a normal pattern engine but only with more sensitive and robust profiles which allow a certain degree of deviation.
Modern profile engines (e.g. PROSITE) increasing begin to employ HMM-like algorithms which are in contrast to classic profiles not only based on experimental results but have a very strong theoretical/mathematical fundament, are more noise-resistant and more reliable as well as flexible than patterns and profiles.

Other tools

There are numerous predictors of protein features available online. We only have time to try a few examples. A good site with multiple machine-learning based predictors is Center for Biological Sequence Analysis in Copenhagen (CBS).

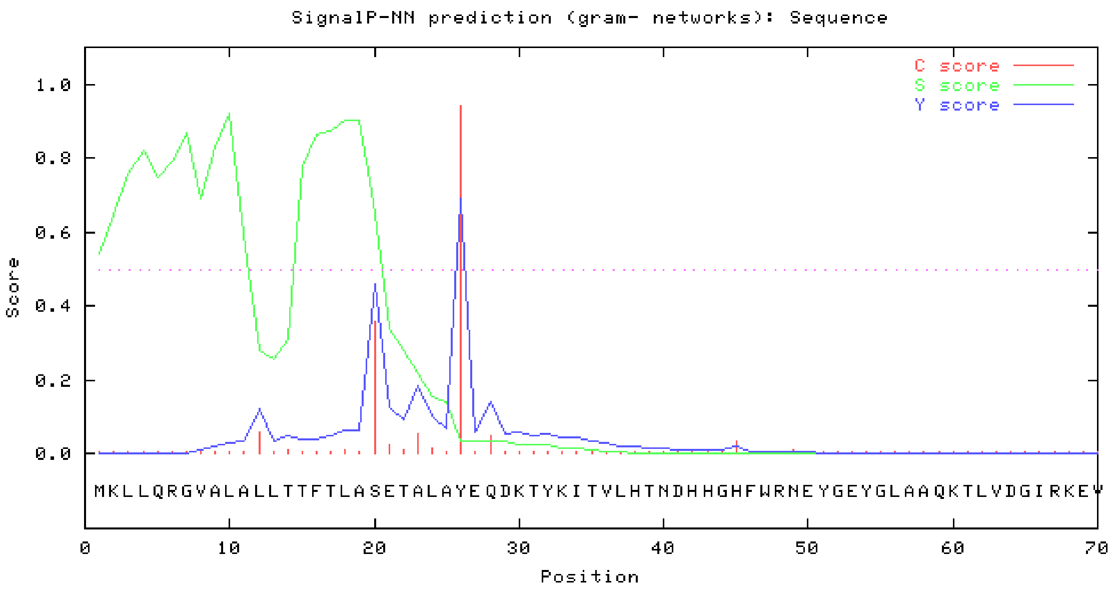
Signal sequences, that direct newly synthesised protein to its final localisation where it should fulfill its function, are often encoded in the N-terminal parts. There are now several predictors available for such predictions and SignalP from CBS is the most well-known. First, you shall use CBS predictors to predict the presence or absence of signal sequences and organelle localisation for three sequences below with given Swissprot IDs.

Hint: You first need to get the amino acid sequence. Go to UniProt. Search for the accession number in question and click FASTA in the upper right corner (orange button).
Use SignalP to predict the signal sequences and TargetP to predict the organelle localisation for the following sequences. Give both the localisation and the actual signal sequence.

2-17. Give both the localisation and the actual signal sequence of USHA_ECOLI ?

- Protein UshA

```
>sp|P07024|USHA_ECOLI Protein UshA OS=Escherichia coli (strain K12) GN=ushA PE=1
SV=2
MKLLQRGVALALLTTFTLASETALAYEQDKTYKITVLHTNDHHGHFWRNEYGEYGLAAQK
TLVDGIRKEVAAEGGSVLLLSGGDINTGVPESDLQDAEPDFRGMNLVGYDAMAIGNHEFD
NPLTVLRQQEKWAKFPLLSANIYQKSTGERLFKPWALFKRQDLKIAVIGLTTDDTAKIGN
PEYFTDIEFRKPADEAKLVIQELQQTEKPDIIIAATHMGHYDNGEHGSNAPGDVEMARAL
PAGSLAMIVGGHSQDPVCMAAENKKQVDYVPGTPCKPDQQNGIWIVQAHEWGKYVGRADF
EFRNGEMKMVNYQLIPVNLKKKVTWEDGKSERVLYTPEIAENQQMISLLSPFQNKGKAQL
EVKIGETNGRLEGDRDKVRFVQTNMGRLILAAQMDRTGADFAVMSGGGIRDSIEAGDISY
KNVLKVQPFGNVVVYADMTGKEVIDYLTAVAQMKPDSGAYPQFANVSFVAKDGKLNDLKI
KGEPVDPAKTYRMATLNFNATGGDGYPRLDNKPGYVNTGFIDAEVLKAYIQKSSPLDVSV
YEPKGEVSWQ
```


SignalP-NN prediction (gram- networks): Sequence

```
>Sequence                 length = 70
# Measure  Position  Value  Cutoff  signal peptide?
  max. C    26       0.940   0.52   YES
  max. Y    26       0.693   0.33   YES
  max. S    10       0.920   0.92   YES
  mean S     1-25    0.607   0.49   YES
```

```
       D     1-25    0.650   0.44   YES
# Most likely cleavage site between pos. 25 and 26: ALA-YE
```

The signal sequence is between positions 1-25, whereas the cutoff (cleavage site) is at position 26, probability 0,940.

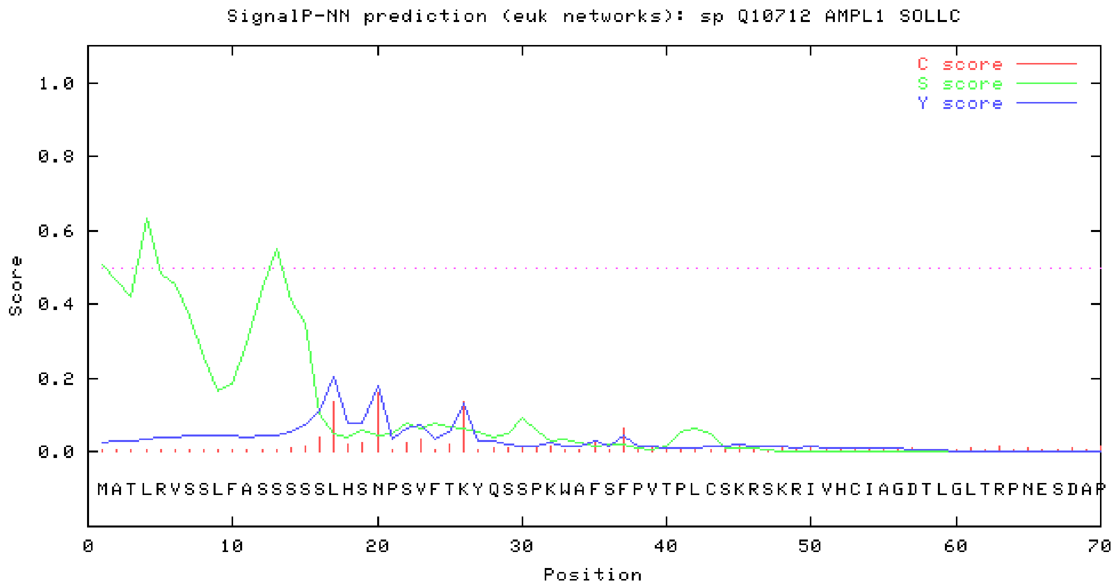
```
Name                      Len            mTP     SP  other  Loc   RC
----------------------------------------------------------------------
Sequence                  550          0.055  0.894  0.044   S     1
----------------------------------------------------------------------
cutoff                                 0.000  0.000  0.000
```

The sequence is in the secretory pathway since the SP-value is the highest, thus indicating that the sequence contains a signal peptide.

2-18. Give both the localisation and the actual signal sequence of AMPL1_SOLLC ?

- Leucine aminopeptidase 1, chloroplastic

```
>sp|Q10712|AMPL1_SOLLC Leucine aminopeptidase 1, chloroplastic OS=Solanum
lycopersicum GN=LAPA1 PE=2 SV=1
MATLRVSSLFASSSSSLHSNPSVFTKYQSSPKWAFSFPVTPLCSKRSKRIVHCIAGDTLG
LTRPNESDAPKISIGAKDTAVVQWQGDLLAIGATENDMARDENSKFKNPLLQQLDSELNG
LLSAASSEEDFSGKSGQSVNLRFPGGRITLVGLGSSASSPTSYHSLGQAAAAAAKSSQAR
NIAVALASTDGLSAESKINSASAIATGVVLGSFEDNRFRSESKKSTLESLDILGLGTGPE
IERKIKYAEHVCAGVILGRELVNAPANIVTPAVLAEEAKKIASTYSDVISVNILDAEQCK
ELKMGAYLAVAAAATENPPYFIHLCFKTPTKERKTKLALVGKGLTFDSGGYNLKVGARSR
IELMKNDMGGAAAVLGAAKALGEIRPSRVEVHFIVAACENMISAEGMRPGDIVTASNGKT
IEVNNTDAEGRLTLADALIYACNQGVEKIIDLATLTGAIMVALGPSVAGAFTPNDDLARE
VVEAAEASGEKLWRMPMEESYWESMKSGVADMINTGPGNGGAITGALFLKQFVDEKVQWL
HLDVAGPVWSDEKKNATGYGVSTLVEWVLRN
```



```
>sp_Q10712_AMPL1_SOLL  length = 70
# Measure  Position  Value   Cutoff   signal peptide?
  max. C     20       0.159   0.32    NO
  max. Y    17        0.205   0.33   NO
  max. S     4        0.632   0.87   NO
  mean S     1-16     0.381   0.48   NO
       D     1-16     0.293   0.43   NO
```

The signal sequence is between positions 1-17. However, there is no cutoff, either at position 17 or 20 since these value don't reach 50%.

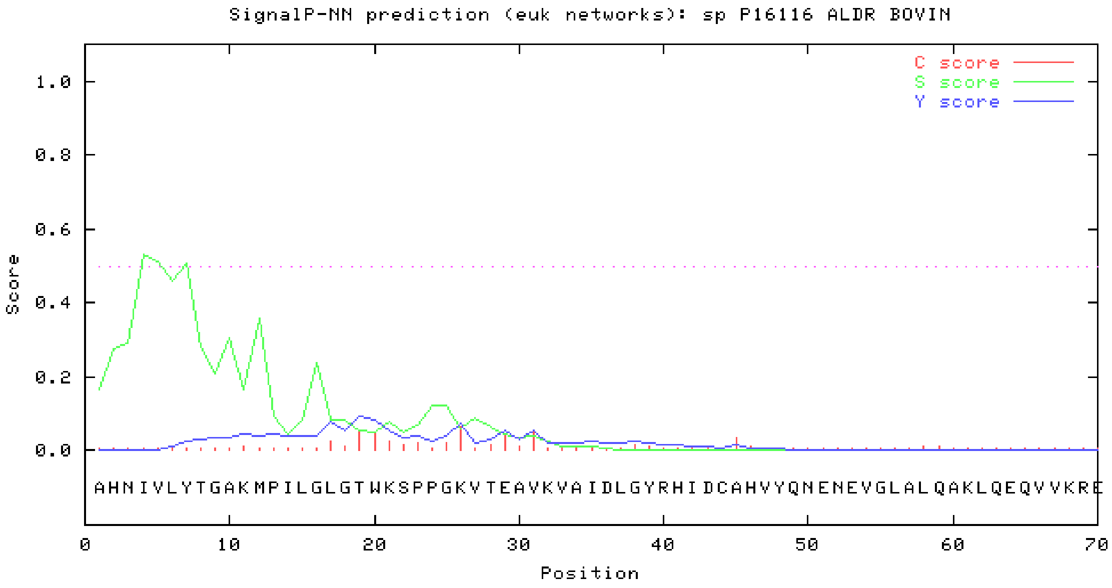
```
Name                      Len      cTP    mTP      SP  other  Loc  RC
----------------------------------------------------------------------
sp_Q10712_AMPL1_SOLL      571    0.831  0.132  0.024  0.032   C    2
----------------------------------------------------------------------
cutoff                           0.000  0.000  0.000  0.000
```

The sequence is localized in the chloroplast and is a chloroplast transit peptide (highest cTP-value).

2-19. Give both the localisation and the actual signal sequence of ALDR_BOVIN ?

- Aldose reductase

```
>sp|P16116|ALDR_BOVIN Aldose reductase OS=Bos taurus GN=AKR1B1 PE=1 SV=1
AHNIVLYTGAKMPILGLGTWKSPPGKVTEAVKVAIDLGYRHIDCAHVYQNENEVGLALQA
KLQEQVVKREDLFIVSKLWCTYHDKDLVKGACQKTLSDLKLDYLDLYLIHWPTGFKPGKD
FFPLDEDGNVIPSEKDFVDTWTAMEELVDEGLVKAIGVSNFNHLQVEKILNKPGLKYKPA
VNQIECHPYLTQEKLIQYCNSKGIVVTAYSPLGSPDRPWAKPEDPSILEDPRIKAIADKY
NKTTAQVLIRFPIQRNLIVIPKSVTPERIAENFQVFDFELDKEDMNTLLSYNRDWRACAL
VSCASHRDYPFHEEF
```



```
>sp_P16116_ALDR_BOVIN  length = 70
# Measure  Position  Value  Cutoff  signal peptide?
  max. C    26        0.065   0.32   NO
  max. Y    19        0.095   0.33   NO
  max. S     4        0.533   0.87   NO
  mean S     1-18     0.261   0.48   NO
       D     1-18     0.178   0.43   NO
```

There is no strong signal sequence from this entry, but whatever signal is present exists at positions 1-19.

```
Name                     Len            mTP     SP  other  Loc  RC
----------------------------------------------------------------------
sp_P16116_ALDR_BOVIN   315            0.085  0.120  0.749   _    2
----------------------------------------------------------------------
cutoff                                0.000  0.000  0.000
```

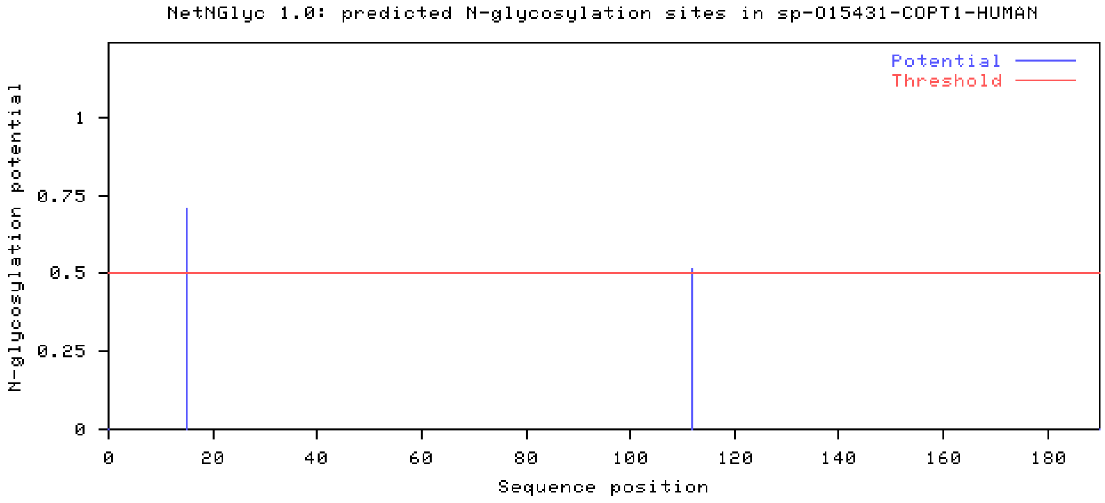
The protein is present in the cytoplasm (any other location than mitochondria, chloroplast and secretory pathway =cytoplasm).

2-22-23. *The major human copper uptake protein, hCTR1 is a transmembrane protein (O15431) which mediates copper uptake through the cell membrane.*

```
>sp|O15431|COPT1_HUMAN High affinity copper uptake protein 1 OS=Homo sapiens
GN=SLC31A1 PE=1 SV=1
MDHSHHMGMSYMDSNSTMQPSHHHPTTSASHSHGGGDSSMMMMPMTFYFGFKNVELLFSG
LVINTAGEMAGAFVAVFLLAMFYEGLKIARESLLRKSQVSIRYNSMPVPGPNGTILMETH
KTVGQQMLSFPHLLQTVLHIIQVVISYFLMLIFMTYNGYLCIAVAAGAGTGYFLFSWKKA
VVVDITEHCH
```

2-22. Are there any more likely N-glycosylation sites except Asn-15?

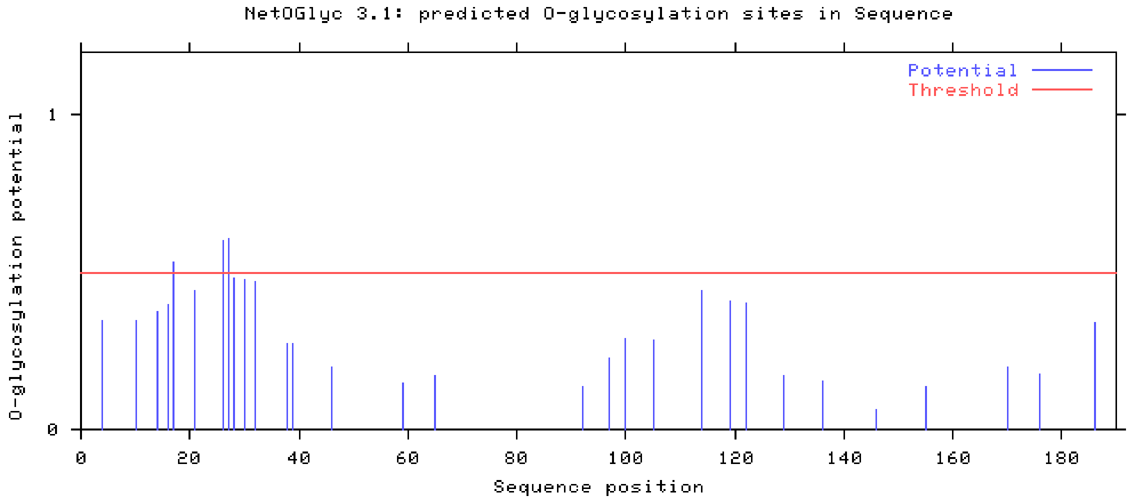
- There is an N-glycolsolated site at residue112. According to pfam, the first transmembrane helix is at residue 60. This means that the first intercellular loop is from res 63-69; the next loop, res 86-130 (including res 112) would be extracellular again. Therefore this site is not to be disregarded.



```
(Threshold=0.5)
----------------------------------------------------------------------
SeqName      Position  Potential   Jury    N-Glyc
                                 agreement  result
----------------------------------------------------------------------
sp_O15431_COPT1_HUMAN   15 NSTM   0.7082     (9/9)    ++
sp_O15431_COPT1_HUMAN  112 NGTI   0.5151     (5/9)    +
----------------------------------------------------------------------
```

2-23. Does mucin-type glycosylation of hCTR1 seem likely?

- No, it doesn't seem so likely since the threshold is very low, but it might occur and if it does it would occur at positions 17, 26 and 27.



```
Name                          S/T   Pos  G-score I-score Y/N  Comment
---------------------------------------------------------------------
Sequence                       T     17   0.532   0.380   T    -
Sequence                       T     26   0.598   0.069   T    -
```

Name	S/T	Pos	G-score	I-score	Y/N
sp_O15431_C	T	17	0.532	0.380	T
sp_O15431_C	T	26	0.598	0.069	T
sp_O15431_C	T	27	0.606	0.034	T

A recent publication showed that Asn-15 is the **only** N-glycosylation site in hCTR1 and that the **key** site of mucin-type O-glycosylation is Thr-27, although additional sites cannot be excluded. Both sites where shown to be important for the copper transport function of the protein.

Links

Pfam
http://pfam.sanger.ac.uk/(search)
http://pfam.sanger.ac.uk/help (help)
Prosite
http://www.expasy.ch/prosite/(search)
http://www.expasy.ch/prosite/prosuser.html (user manual)
SignalP and TargetP
http://www.cbs.dtu.dk/services