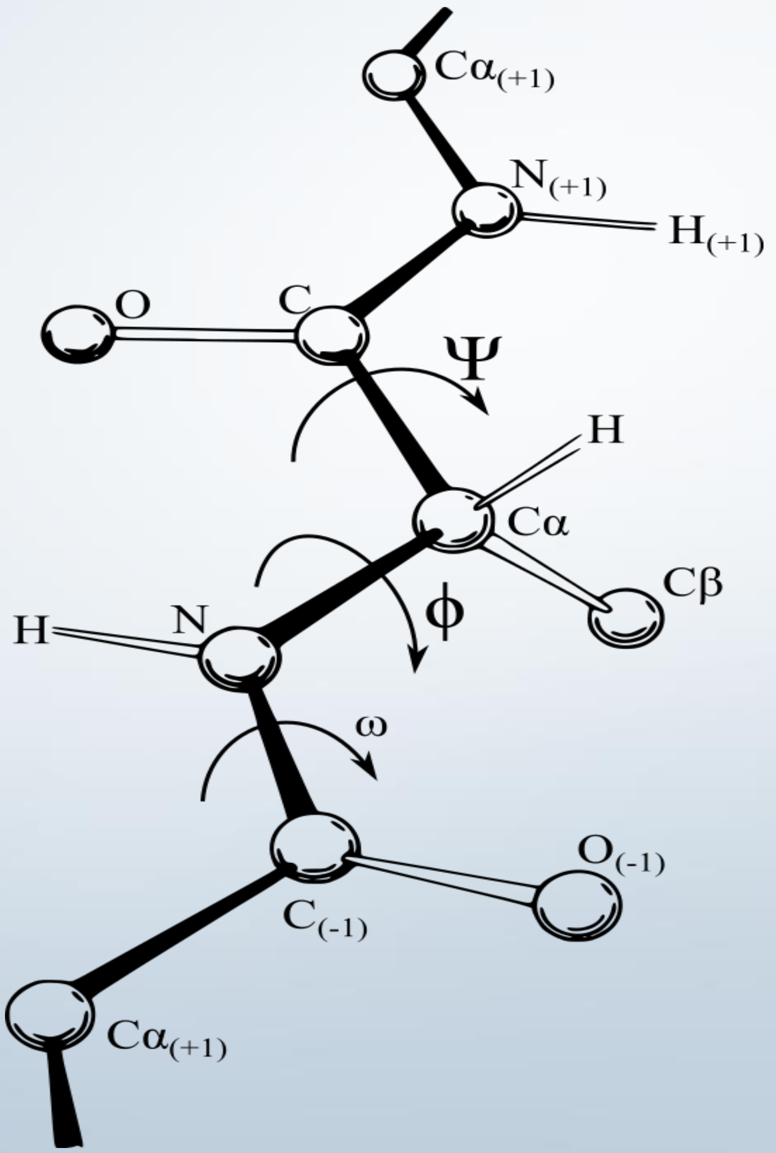
Working with Proteins

Yazdan Asgari

2020

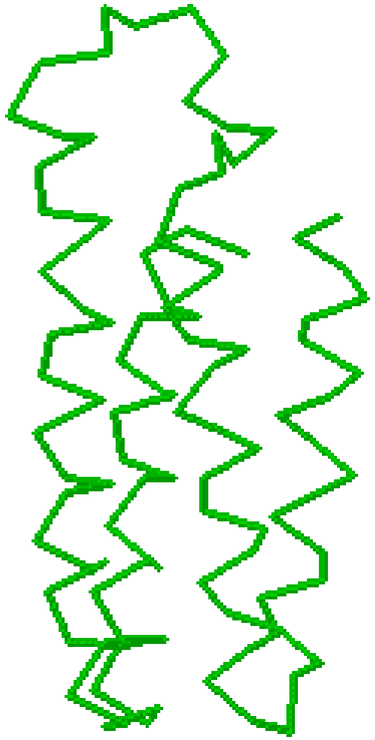
Protein Backbone



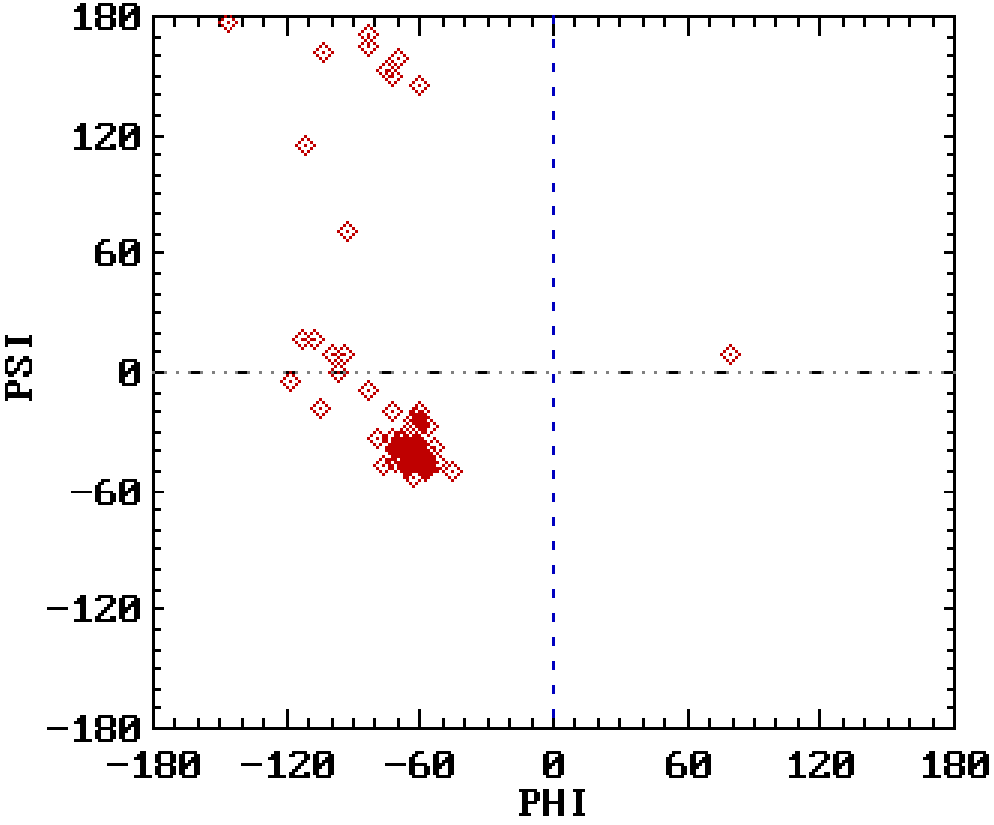
Ramachandran Plot

- Plot of φ vs. ψ
- Repeating values of φ and ψ along the chain result in regular structure
- For example, repeating values of φ ~ -57° and ψ ~ -47° give a right-handed helical fold (the alpha-helix)
- The structure of cytochrome C-256 shows many segments of helix and the Ramachandran plot shows a tight grouping of φ, ψ angles near -50, -50

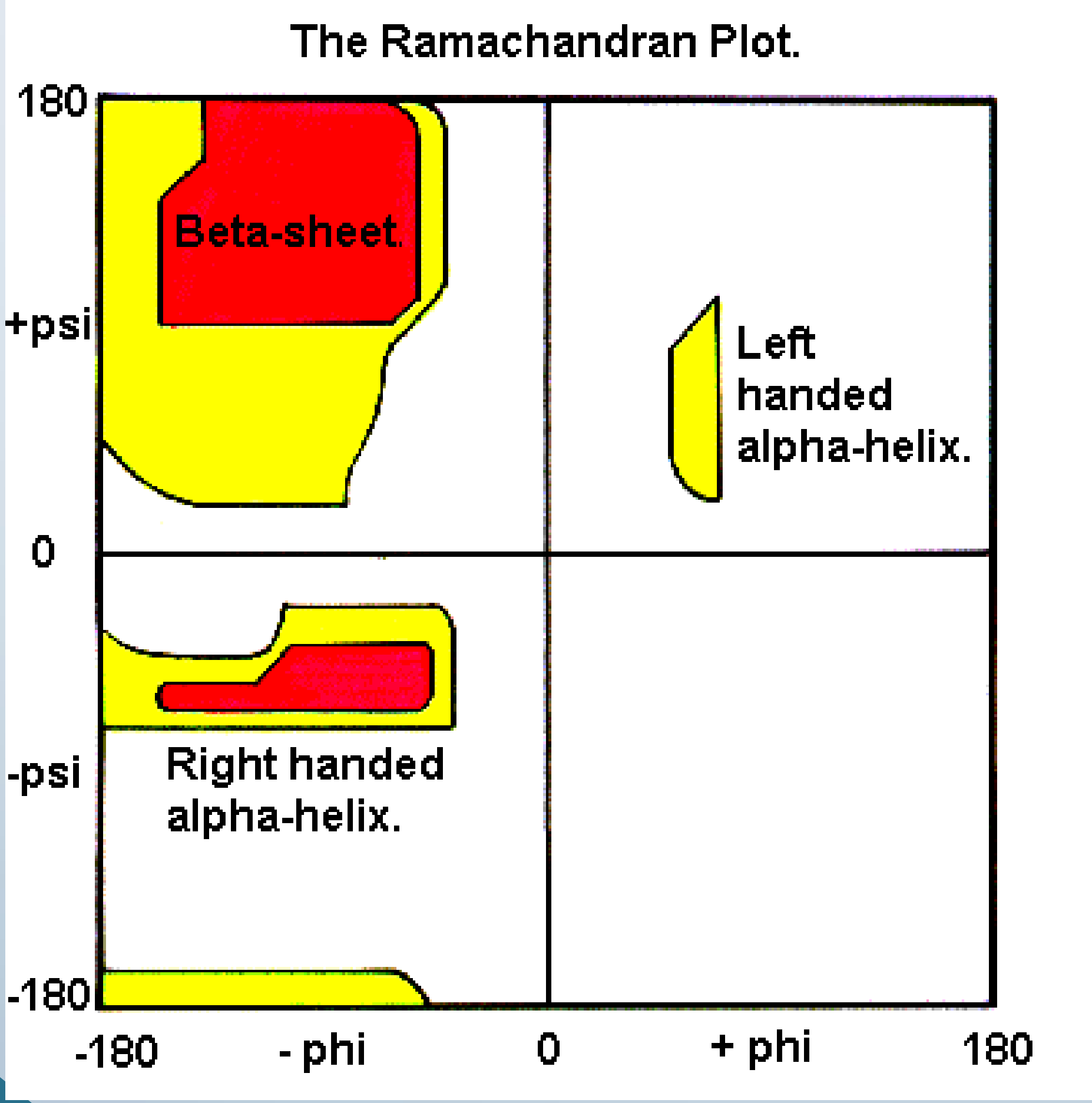
The structure of cytochrome C-256



alpha-helix



cytochrome C-256 Ramachandran plot

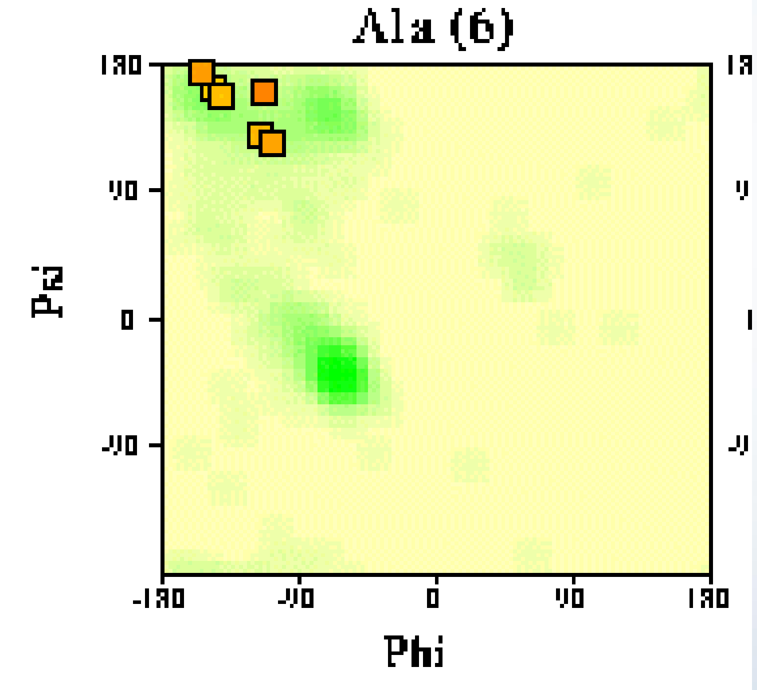


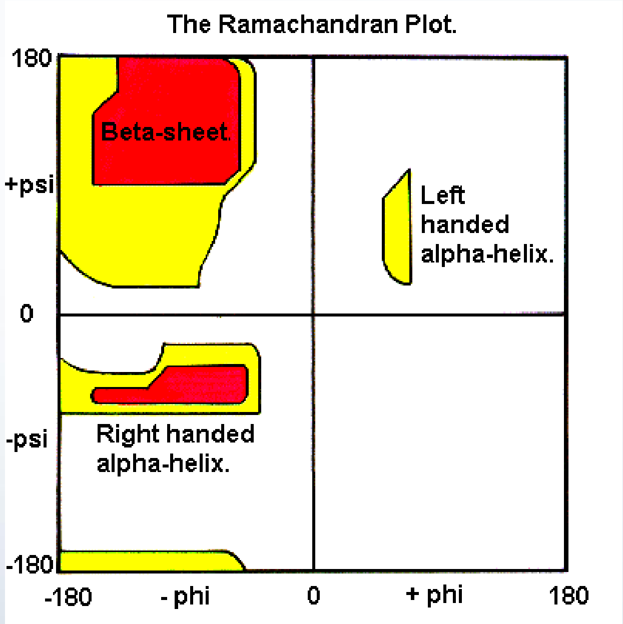
- White = Sterically disallowed conformations (atoms in the polypeptide come closer than the sum of their van der Waals radii)

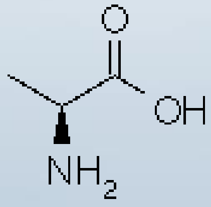
- Red = Sterically allowed regions (namely right-handed alpha helix and beta sheet)

- Yellow = Sterically allowed if shorter radii are used (i.e. atoms allowed closer together; brings out left-handed helix)

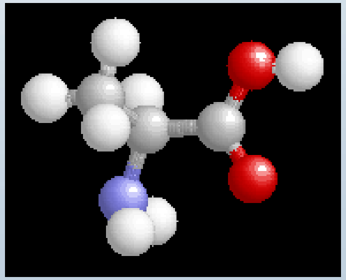
Alanine Ramachandran Plot



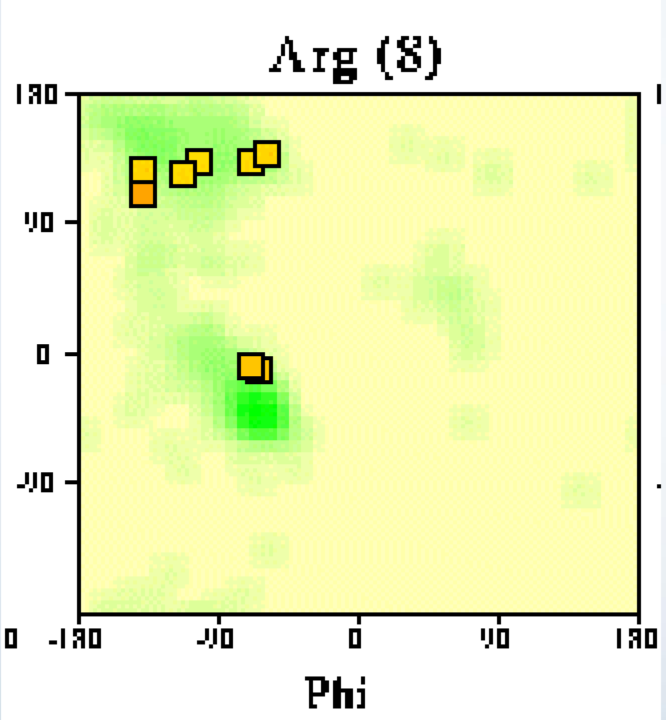


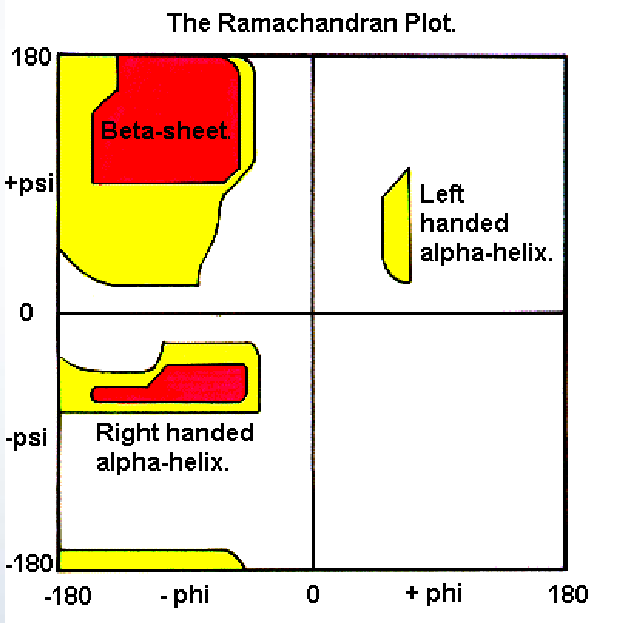


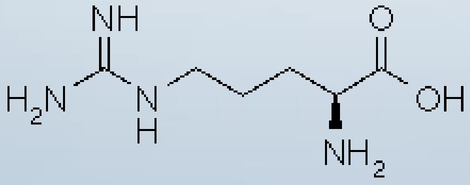
ala a Alanin



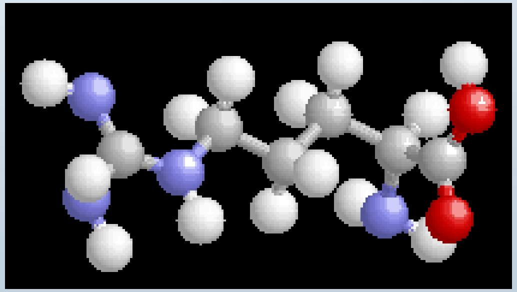
Arginine Ramachandran Plot



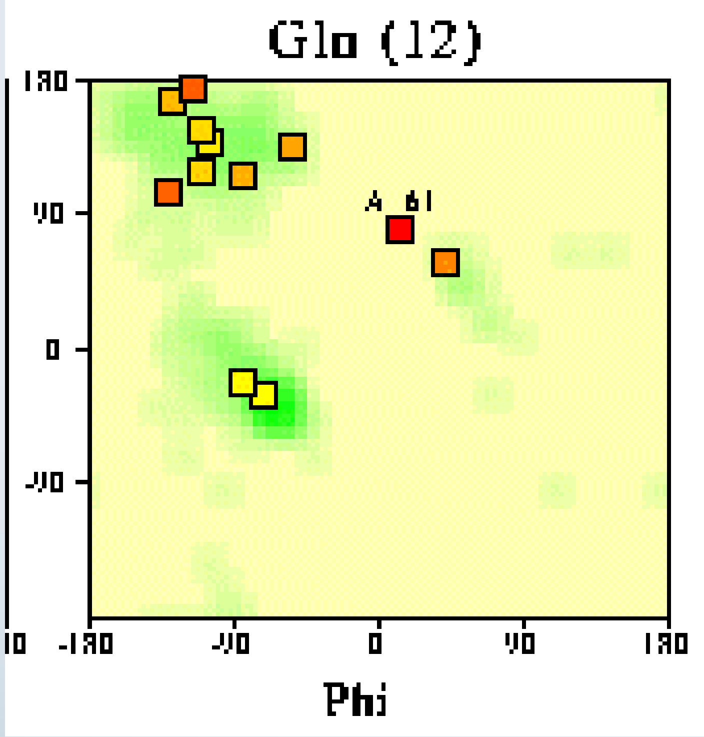


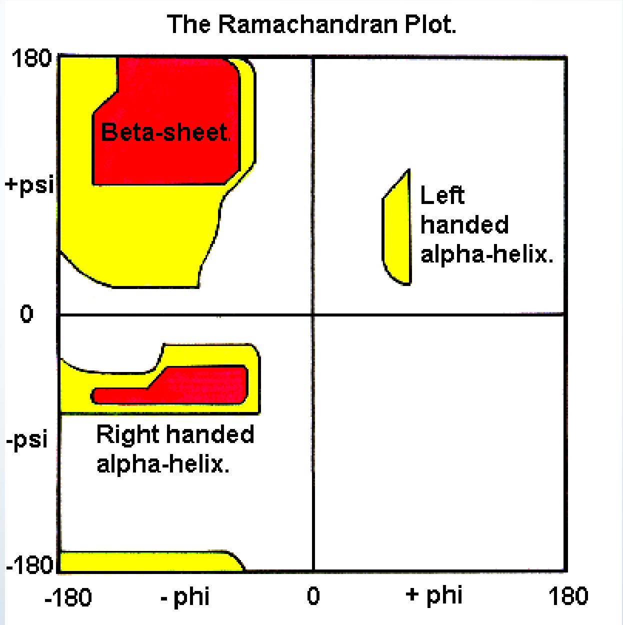


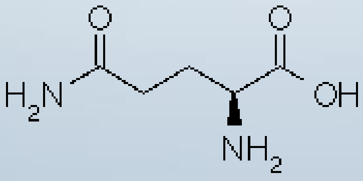
arg r Arginin



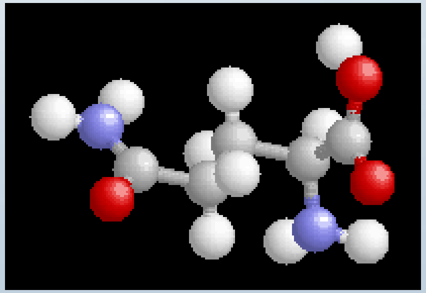
Glutamine Ramachandran Plot



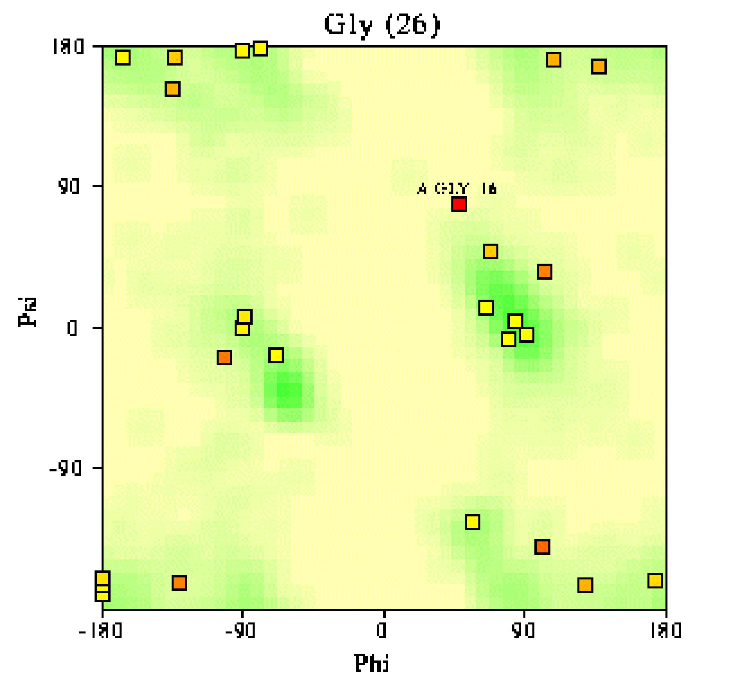


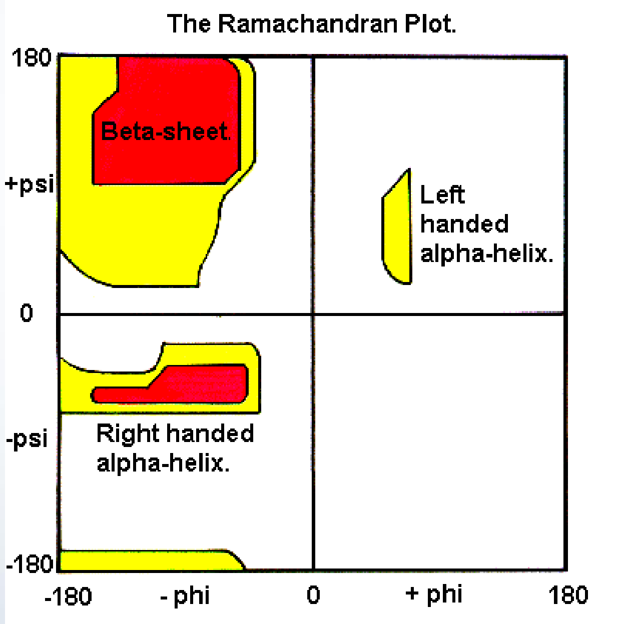


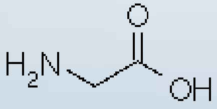
gln q Glutamin



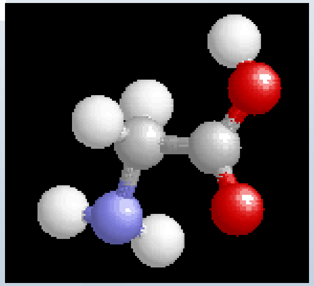
Glycine Ramachandran Plot



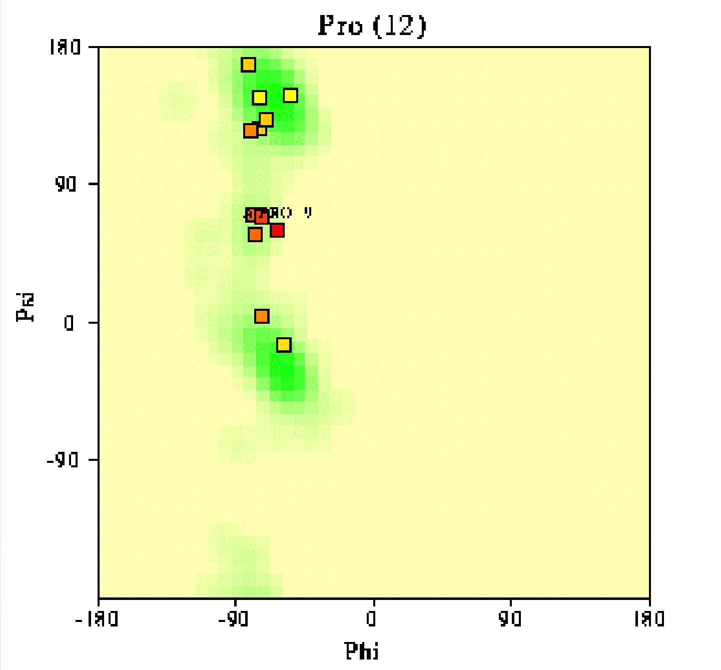


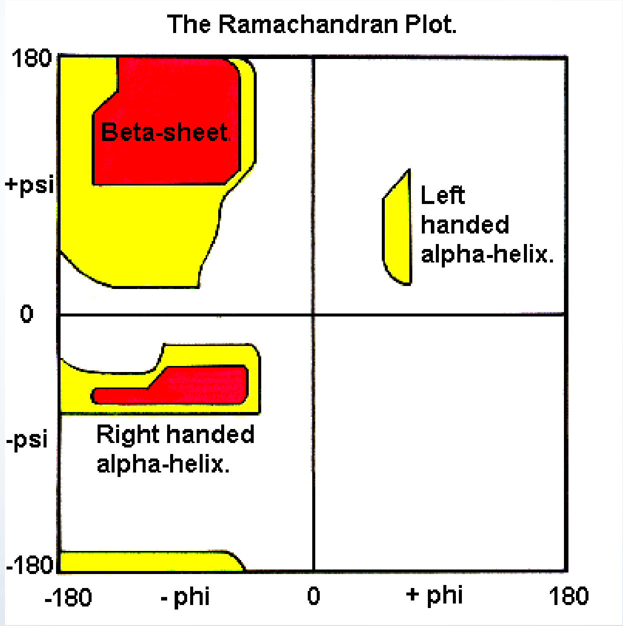


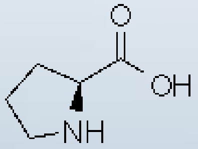
gly g Glycin



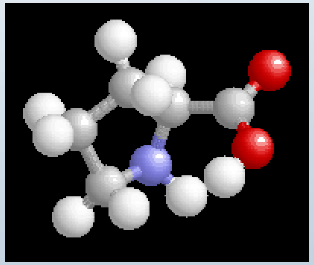
Proline Ramachandran Plot



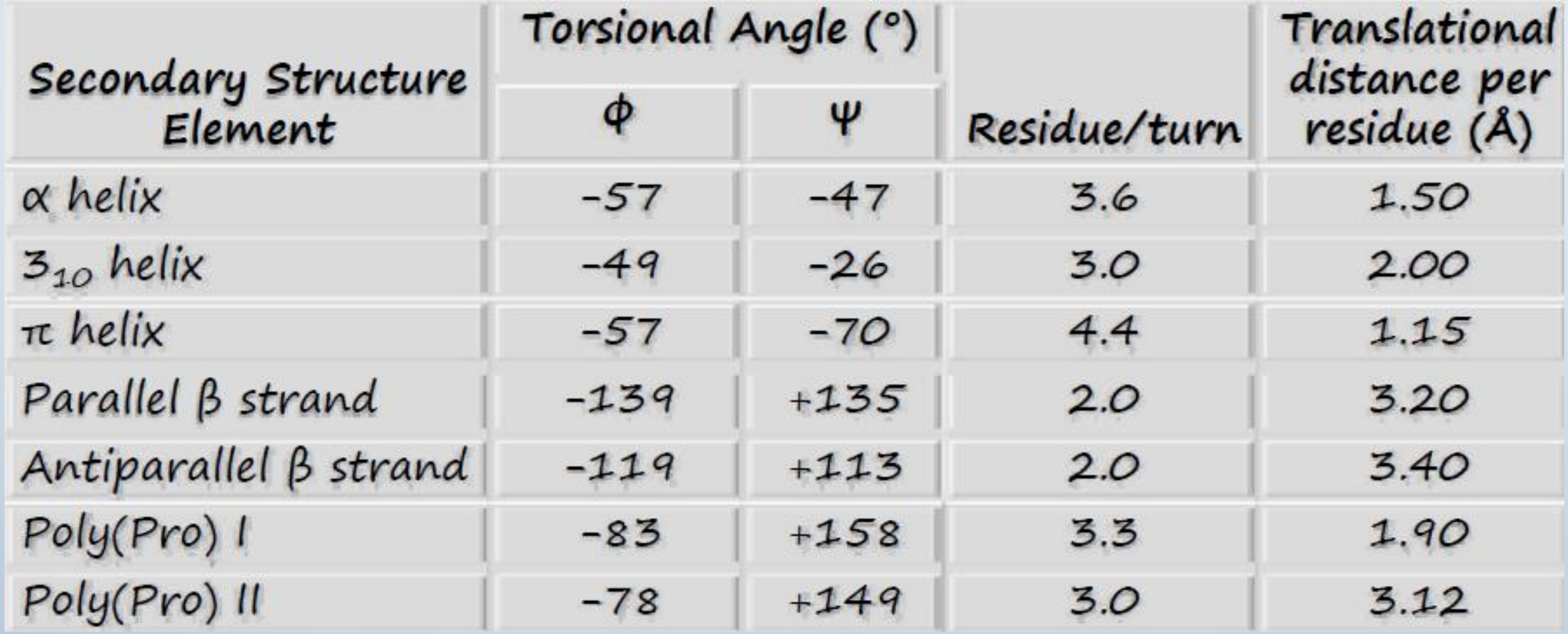




pro p Prolin



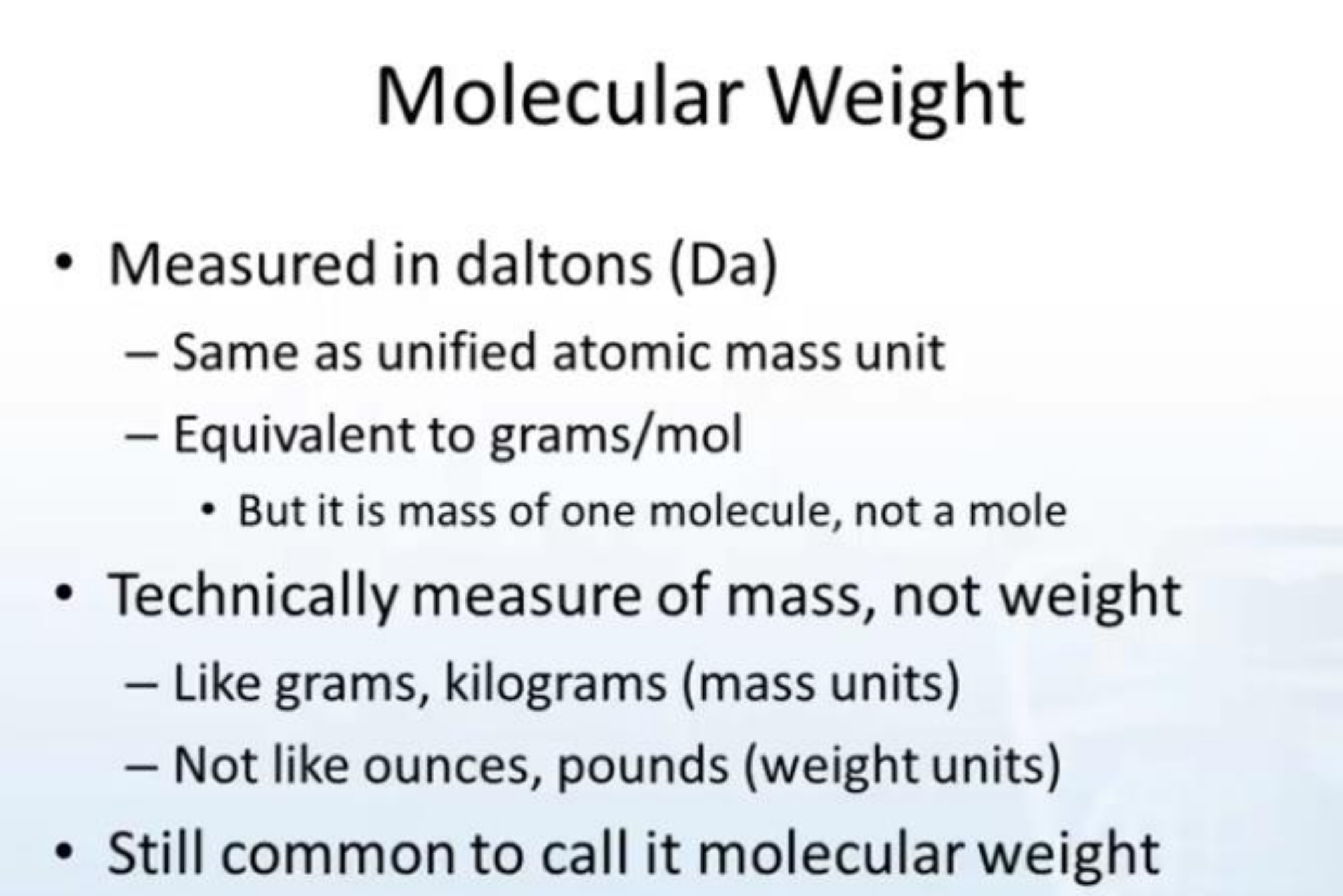
Regular Secondary Structure Conformations



Secondary Structure Element	Torsional Angle (°)		Residue/turn	Translational distance per residue (Å)
	φ	ψ		
α helix	-57	-47	3.6	1.50
3_{10} helix	-49	-26	3.0	2.00
π helix	-57	-70	4.4	1.15
Parallel β strand	-139	+135	2.0	3.20
Antiparallel β strand	-119	+113	2.0	3.40
Poly(Pro) I	-83	+158	3.3	1.90
Poly(Pro) II	-78	+149	3.0	3.12

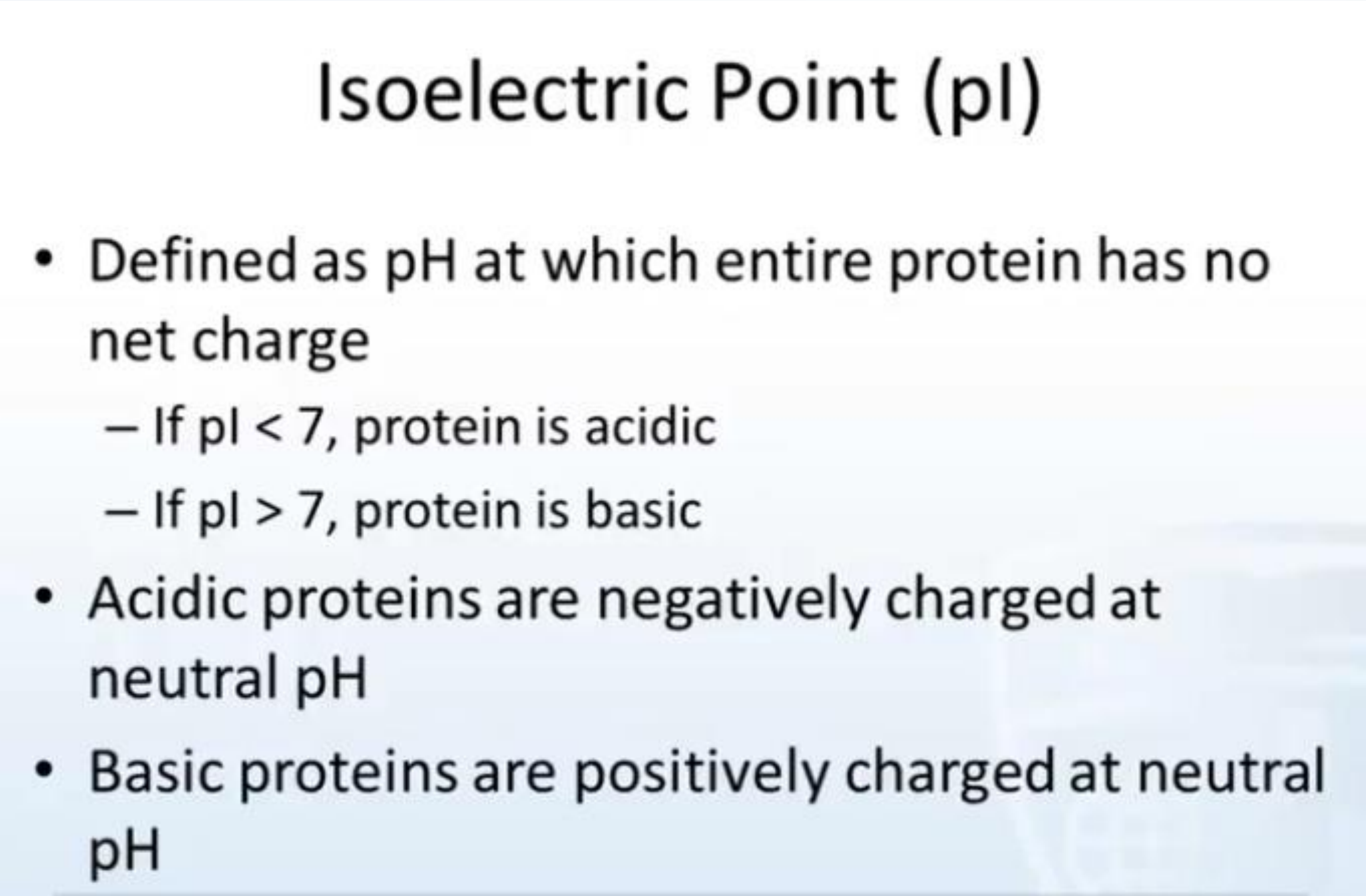
Physicochemical Properties

- Online servers exist to determine many properties of your protein sequences
  - Molecular weight
  - Extinction coefficients
  - Half-life
  - Isoelectric point
- It is also possible to simulate protease digestion



Molecular Weight

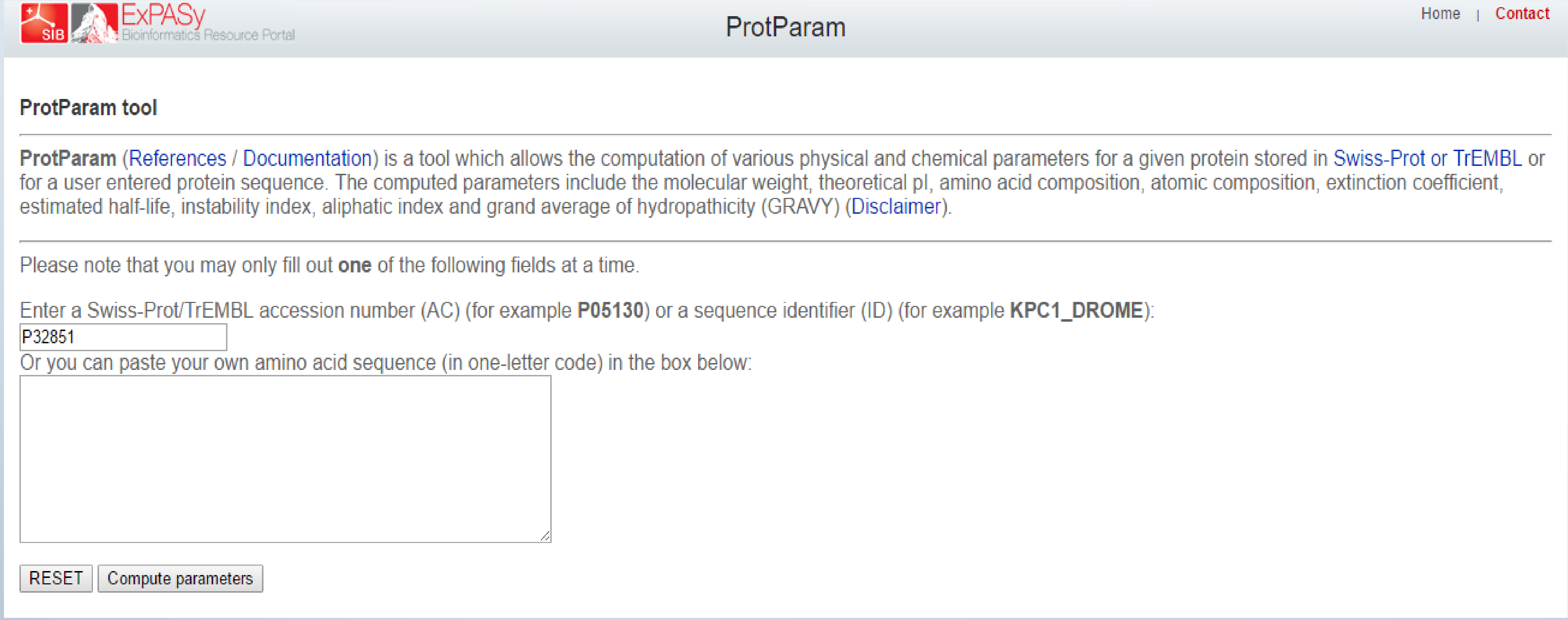
- Measured in daltons (Da)
  - – Same as unified atomic mass unit
  - – Equivalent to grams/mol
    - But it is mass of one molecule, not a mole
- Technically measure of mass, not weight
  - – Like grams, kilograms (mass units)
  - – Not like ounces, pounds (weight units)
- Still common to call it molecular weight



Isoelectric Point (pI)

- Defined as pH at which entire protein has no net charge
  - If pI < 7, protein is acidic
  - If pI > 7, protein is basic
- Acidic proteins are negatively charged at neutral pH
- Basic proteins are positively charged at neutral pH

Protoparam



ProtParam

ProtParam tool

ProtParam (References / Documentation) is a tool which allows the computation of various physical and chemical parameters for a given protein stored in Swiss-Prot or TrEMBL or for a user entered protein sequence. The computed parameters include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY) (Disclaimer).

Please note that you may only fill out **one** of the following fields at a time.

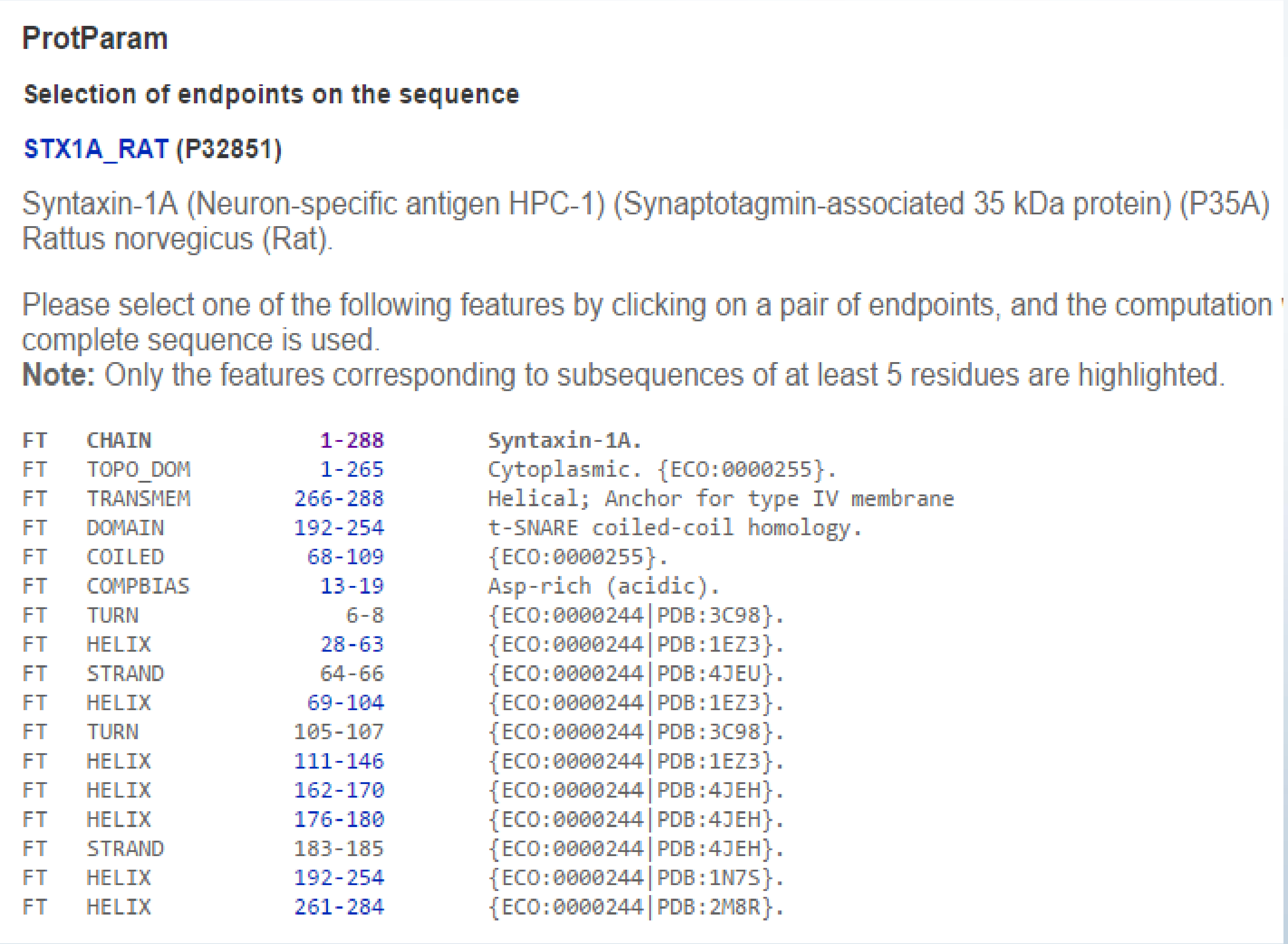
Enter a Swiss-Prot/TrEMBL accession number (AC) (for example **P05130**) or a sequence identifier (ID) (for example **KPC1_DROME**):

P32851

Or you can paste your own amino acid sequence (in one-letter code) in the box below:

http://web.expasy.org/protparam/

Example (P32851 STX1A_RAT)



ProtParam

Selection of endpoints on the sequence

STX1A_RAT (P32851)

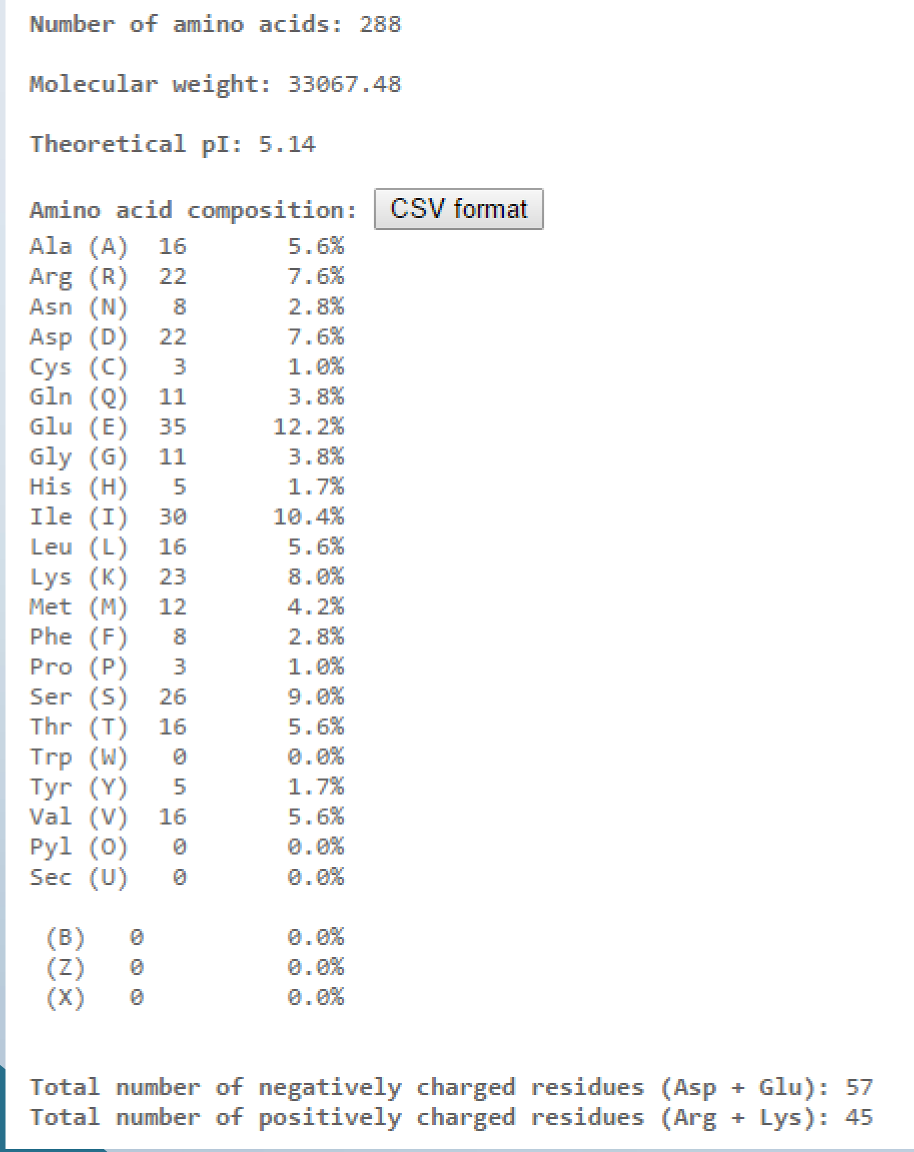
Syntaxin-1A (Neuron-specific antigen HPC-1) (Synaptotagmin-associated 35 kDa protein) (P35A)
Rattus norvegicus (Rat).

Please select one of the following features by clicking on a pair of endpoints, and the computation complete sequence is used.
Note: Only the features corresponding to subsequences of at least 5 residues are highlighted.

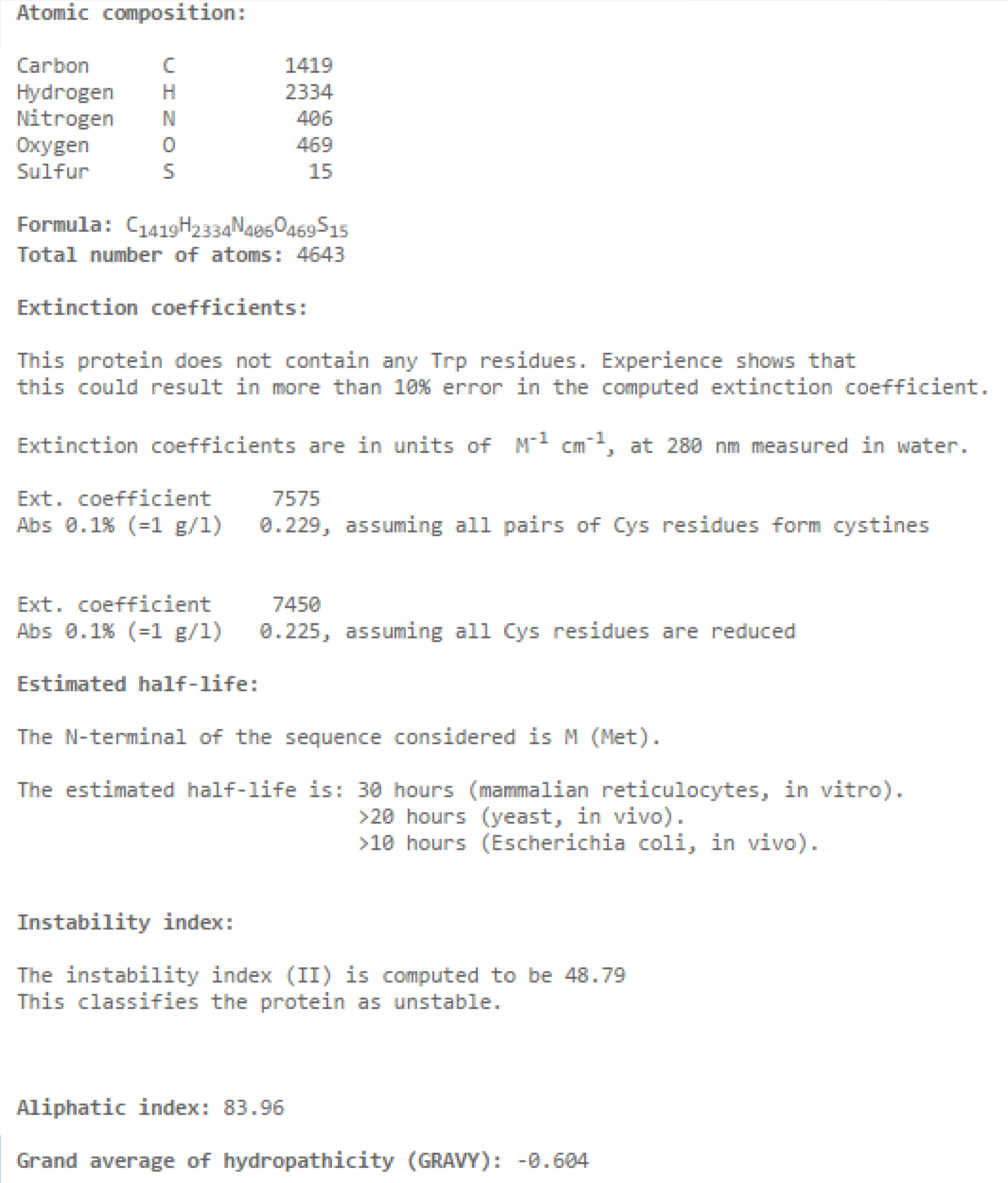
```
FT   CHAIN          1-288    Syntaxin-1A.
FT   TOPO_DOM       1-265    Cytoplasmic. {ECO:0000255}.
FT   TRANSMEM     266-288    Helical; Anchor for type IV membrane
FT   DOMAIN       192-254    t-SNARE coiled-coil homology.
FT   COILED        68-109    {ECO:0000255}.
FT   COMPBIAS       13-19    Asp-rich (acidic).
FT   TURN             6-8    {ECO:0000244|PDB:3C98}.
FT   HELIX          28-63    {ECO:0000244|PDB:1EZ3}.
FT   STRAND         64-66    {ECO:0000244|PDB:4JEU}.
FT   HELIX         69-104    {ECO:0000244|PDB:1EZ3}.
FT   TURN         105-107    {ECO:0000244|PDB:3C98}.
FT   HELIX        111-146    {ECO:0000244|PDB:1EZ3}.
FT   HELIX        162-170    {ECO:0000244|PDB:4JEH}.
FT   HELIX        176-180    {ECO:0000244|PDB:4JEH}.
FT   STRAND       183-185    {ECO:0000244|PDB:4JEH}.
FT   HELIX        192-254    {ECO:0000244|PDB:1N7S}.
FT   HELIX        261-284    {ECO:0000244|PDB:2M8R}.
```

http://web.expasy.org/protparam/

Physicochemical Properties - Example

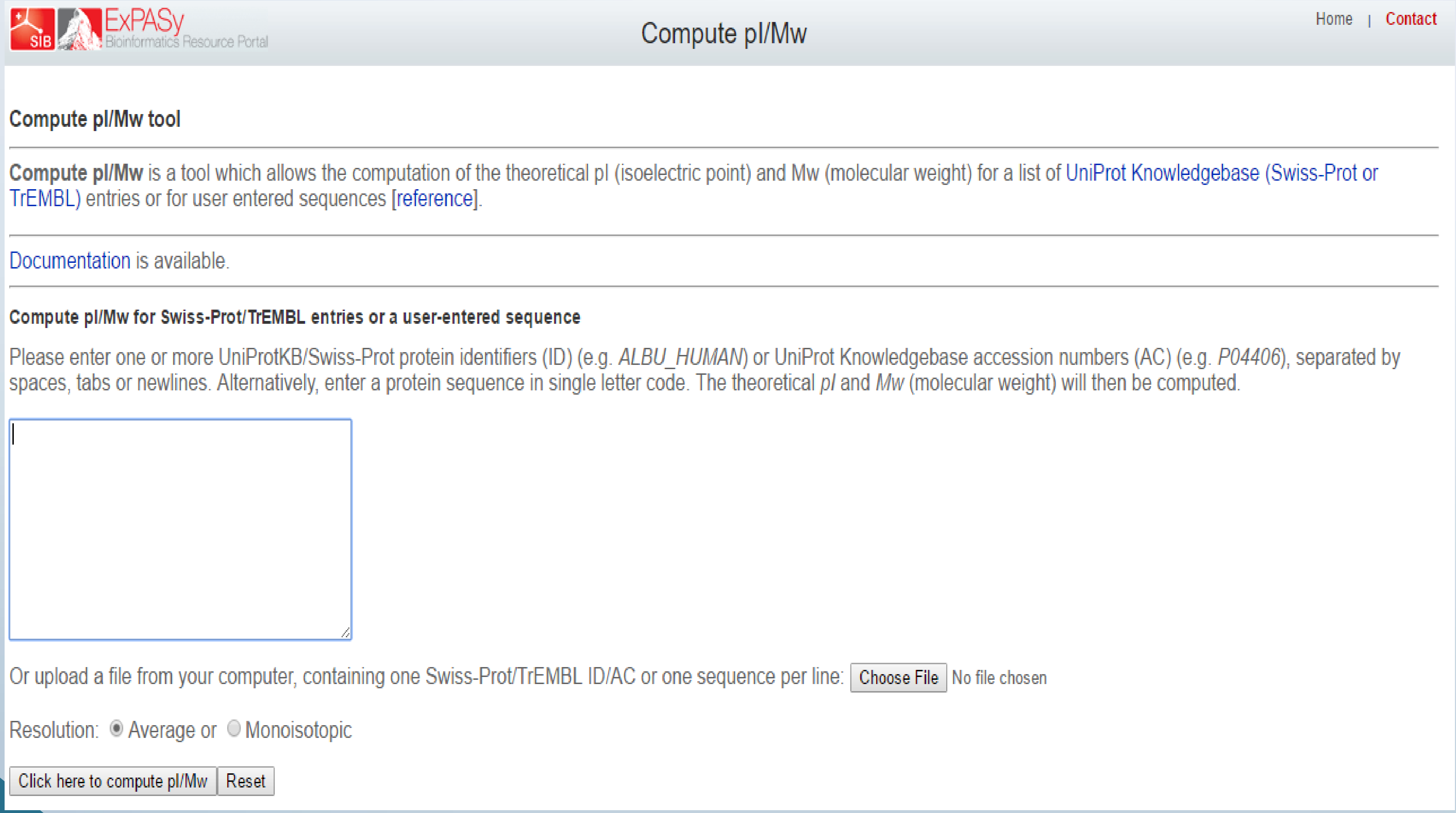


```
Number of amino acids: 288

Molecular weight: 33067.48

Theoretical pI: 5.14

Amino acid composition:  CSV format
Ala (A)  16        5.6%
Arg (R)  22        7.6%
Asn (N)   8        2.8%
Asp (D)  22        7.6%
Cys (C)   3        1.0%
Gln (Q)  11        3.8%
Glu (E)  35       12.2%
Gly (G)  11        3.8%
His (H)   5        1.7%
Ile (I)  30       10.4%
Leu (L)  16        5.6%
Lys (K)  23        8.0%
Met (M)  12        4.2%
Phe (F)   8        2.8%
Pro (P)   3        1.0%
Ser (S)  26        9.0%
Thr (T)  16        5.6%
Trp (W)   0        0.0%
Tyr (Y)   5        1.7%
Val (V)  16        5.6%
Pyl (O)   0        0.0%
Sec (U)   0        0.0%

 (B)   0           0.0%
 (Z)   0           0.0%
 (X)   0           0.0%


Total number of negatively charged residues (Asp + Glu): 57
Total number of positively charged residues (Arg + Lys): 45
```



```
Atomic composition:

Carbon      C        1419
Hydrogen    H        2334
Nitrogen    N         406
Oxygen      O         469
Sulfur      S          15

Formula: $C_{1419}H_{2334}N_{406}O_{469}S_{15}$
Total number of atoms: 4643

Extinction coefficients:

This protein does not contain any Trp residues. Experience shows that
this could result in more than 10% error in the computed extinction coefficient.

Extinction coefficients are in units of  $M^{-1}$ $cm^{-1}$, at 280 nm measured in water.

Ext. coefficient     7575
Abs 0.1% (=1 g/l)   0.229, assuming all pairs of Cys residues form cystines


Ext. coefficient     7450
Abs 0.1% (=1 g/l)   0.225, assuming all Cys residues are reduced

Estimated half-life:

The N-terminal of the sequence considered is M (Met).

The estimated half-life is: 30 hours (mammalian reticulocytes, in vitro).
                            >20 hours (yeast, in vivo).
                            >10 hours (Escherichia coli, in vivo).


Instability index:

The instability index (II) is computed to be 48.79
This classifies the protein as unstable.



Aliphatic index: 83.96

Grand average of hydropathicity (GRAVY): -0.604
```

Compute_pi



Compute pI/Mw

Compute pI/Mw tool

Compute pI/Mw is a tool which allows the computation of the theoretical pI (isoelectric point) and Mw (molecular weight) for a list of UniProt Knowledgebase (Swiss-Prot or TrEMBL) entries or for user entered sequences [reference].

Documentation is available.

Compute pI/Mw for Swiss-Prot/TrEMBL entries or a user-entered sequence

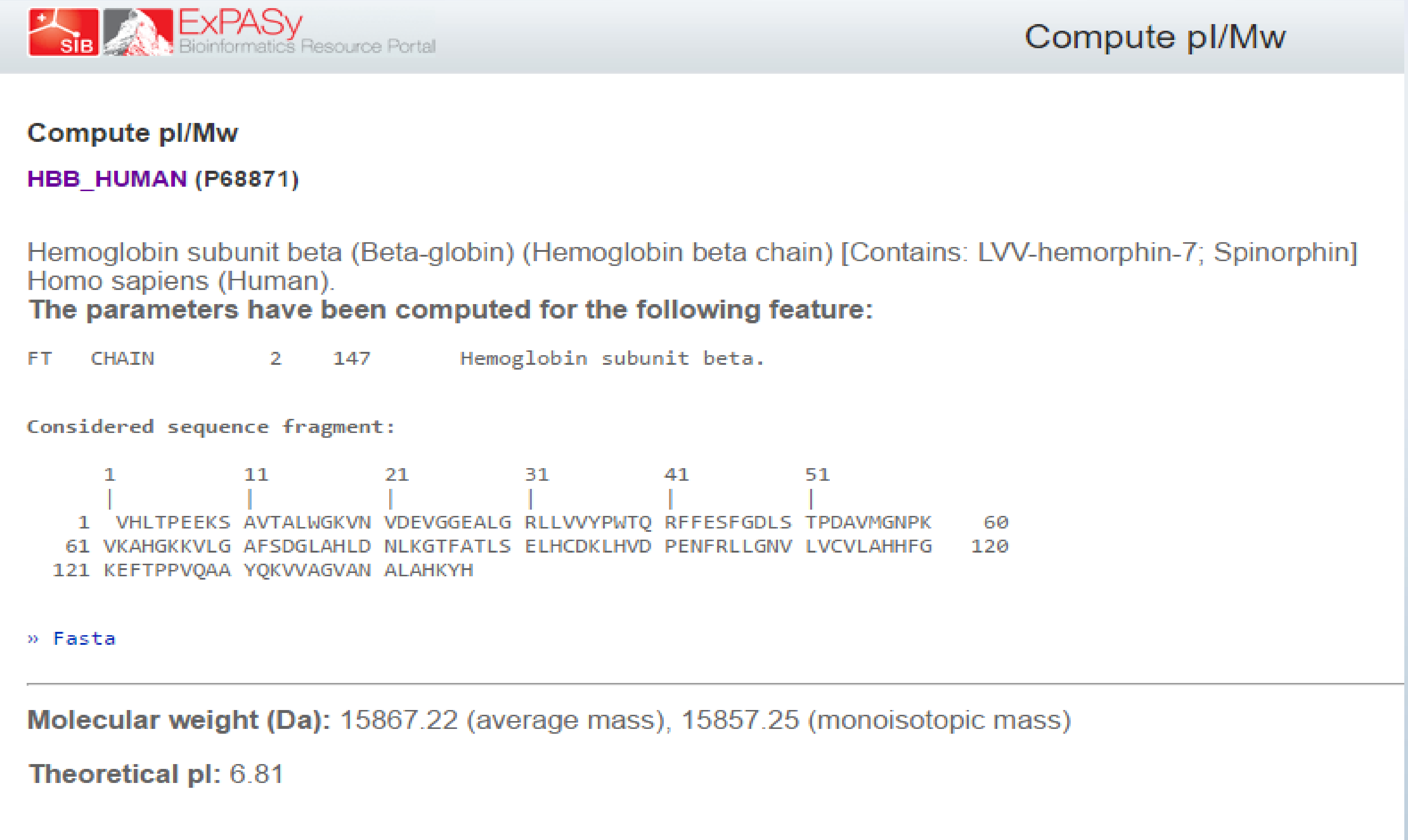
Please enter one or more UniProtKB/Swiss-Prot protein identifiers (ID) (e.g. *ALBU_HUMAN*) or UniProt Knowledgebase accession numbers (AC) (e.g. *P04406*), separated by spaces, tabs or newlines. Alternatively, enter a protein sequence in single letter code. The theoretical *pI* and *Mw* (molecular weight) will then be computed.

Or upload a file from your computer, containing one Swiss-Prot/TrEMBL ID/AC or one sequence per line: Choose File No file chosen

Resolution: ● Average or ○ Monoisotopic

http://web.expasy.org/compute_pi/

Example – (P68871 HBB_HUMAN)



Compute pI/Mw

HBB_HUMAN (P68871)

Hemoglobin subunit beta (Beta-globin) (Hemoglobin beta chain) [Contains: LVV-hemorphin-7; Spinorphin]
Homo sapiens (Human).

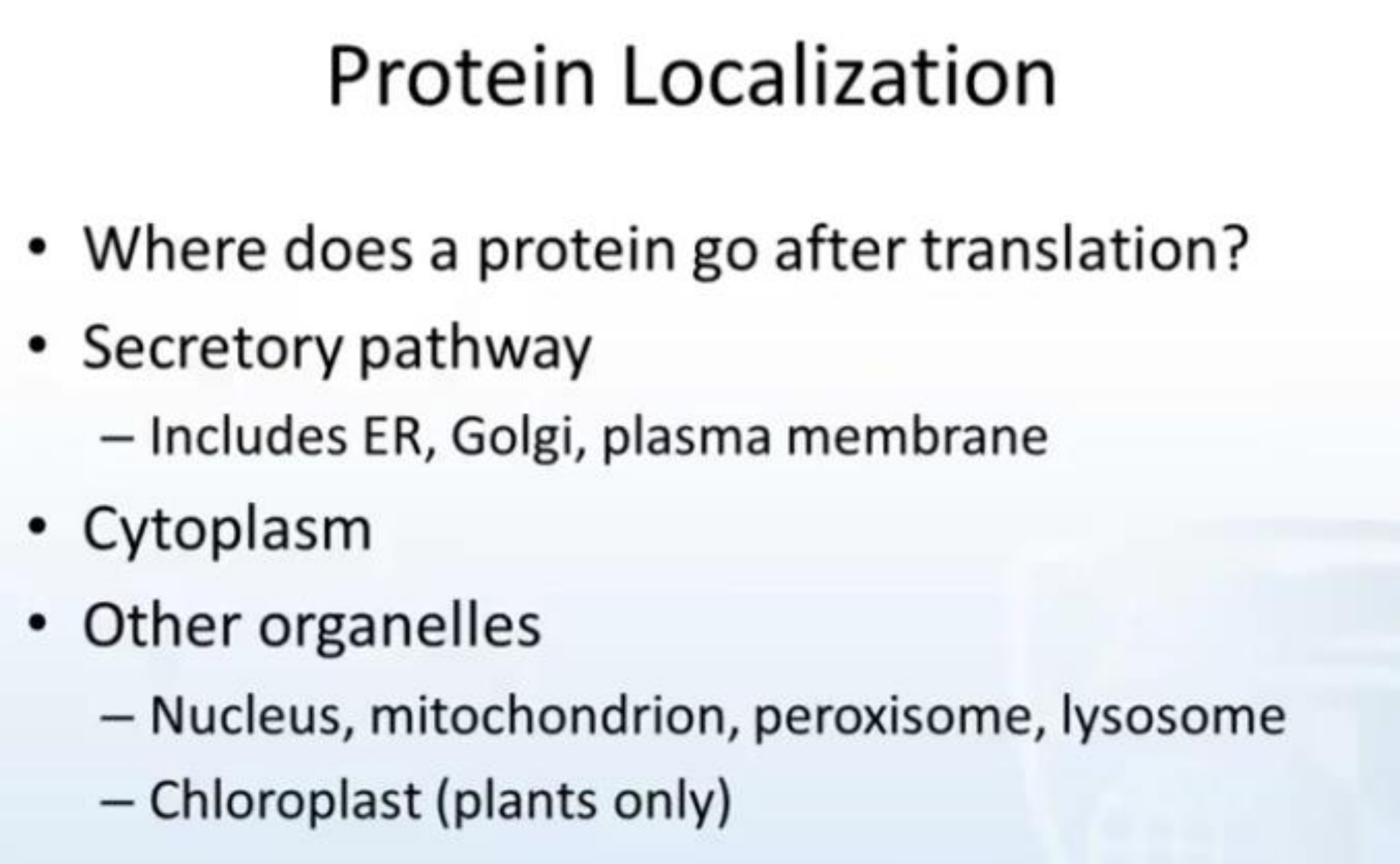
The parameters have been computed for the following feature:

```
FT   CHAIN         2    147       Hemoglobin subunit beta.


Considered sequence fragment:

      1          11         21         31         41         51
      |          |          |          |          |          |
    1  VHLTPEEKS AVTALWGKVN VDEVGGEALG RLLVVYPWTQ RFFESFGDLS TPDAVMGNPK    60
   61 VKAHGKKVLG AFSDGLAHLD NLKGTFATLS ELHCDKLHVD PENFRLLGNV LVCVLAHHFG   120
  121 KEFTPPVQAA YQKVVAGVAN ALAHKYH
```

» Fasta

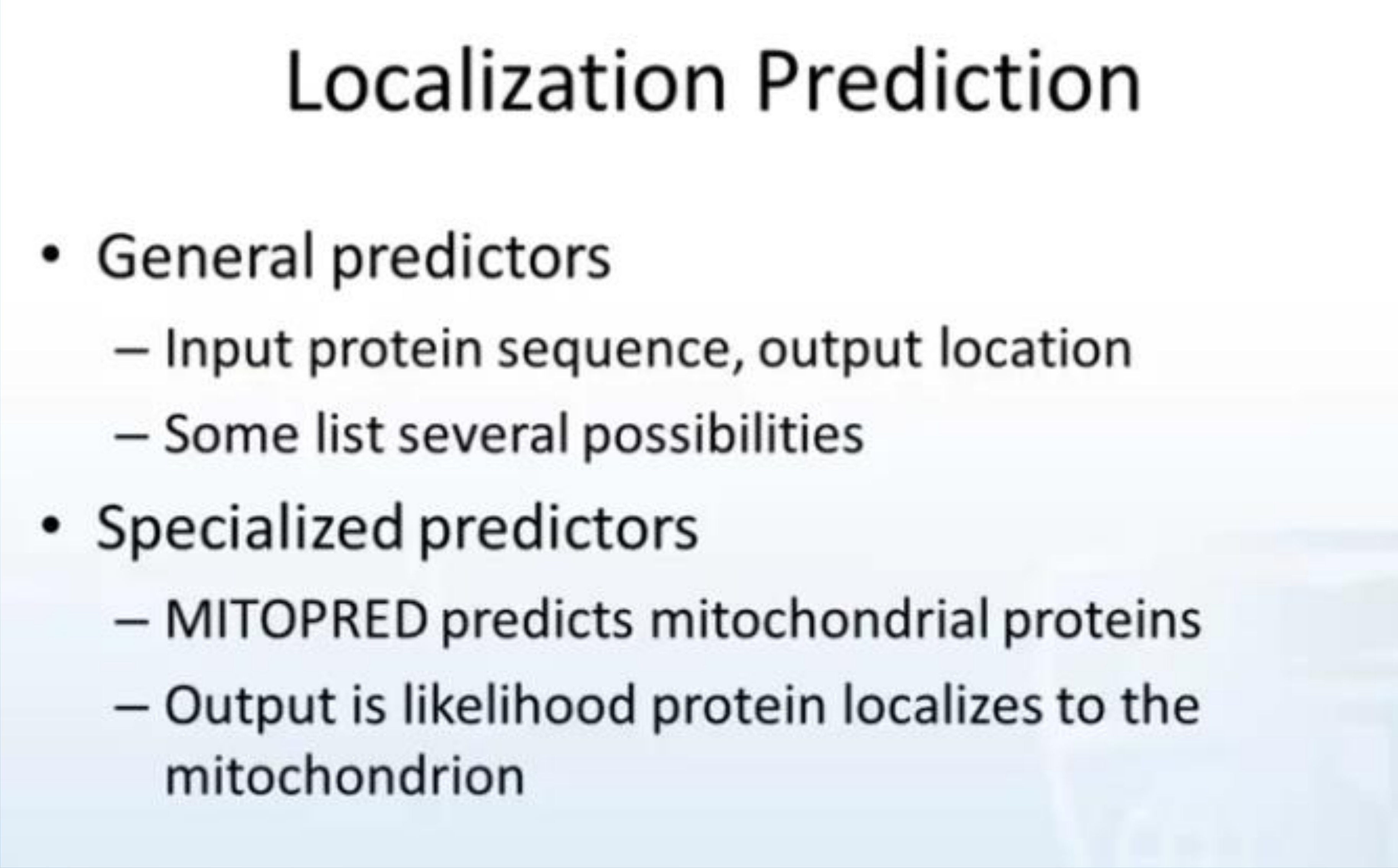
Molecular weight (Da): 15867.22 (average mass), 15857.25 (monoisotopic mass)

Theoretical pI: 6.81



Protein Localization

- Where does a protein go after translation?
- Secretory pathway
  - Includes ER, Golgi, plasma membrane
- Cytoplasm
- Other organelles
  - Nucleus, mitochondrion, peroxisome, lysosome
  - Chloroplast (plants only)



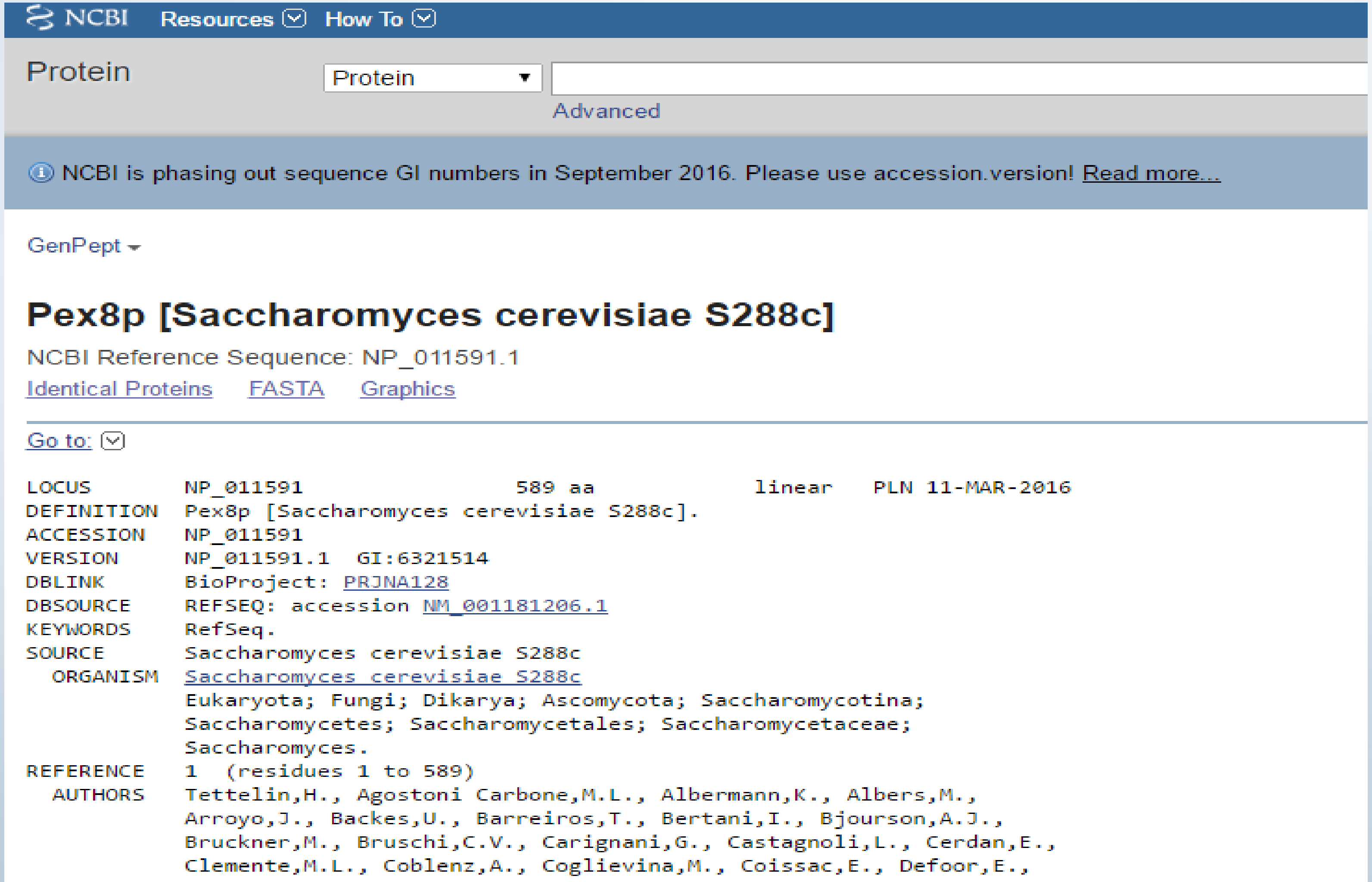
Localization Prediction

- General predictors
  - Input protein sequence, output location
  - Some list several possibilities
- Specialized predictors
  - MITOPRED predicts mitochondrial proteins
  - Output is likelihood protein localizes to the mitochondrion

Some localization servers

- Proloc-GO
- SignalP
- TargetP
- Predotar
- PSORT
- CELLO
- MultiLoc2
- Euk-mPLoc
- LocTree

Example – (TPA: Pex8p [Saccharomyces cerevisiae S288c])



NCBI is phasing out sequence GI numbers in September 2016. Please use accession.version! Read more...

GenPept

Pex8p [Saccharomyces cerevisiae S288c]

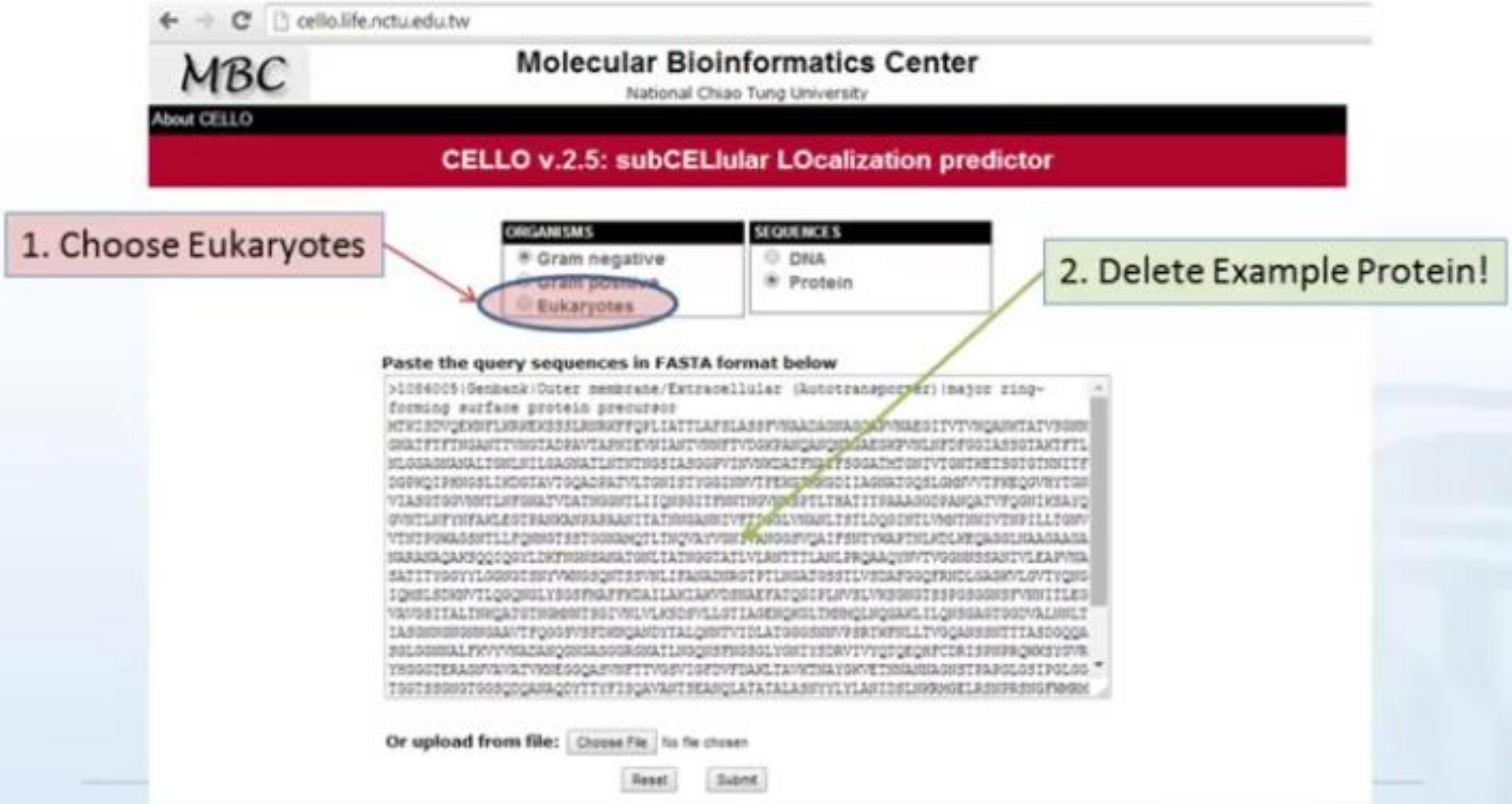
NCBI Reference Sequence: NP_011591.1

Identical Proteins FASTA Graphics

Go to:

```
LOCUS       NP_011591                589 aa            linear   PLN 11-MAR-2016
DEFINITION  Pex8p [Saccharomyces cerevisiae S288c].
ACCESSION   NP_011591
VERSION     NP_011591.1  GI:6321514
DBLINK      BioProject: PRJNA128
DBSOURCE    REFSEQ: accession NM_001181206.1
KEYWORDS    RefSeq.
SOURCE      Saccharomyces cerevisiae S288c
  ORGANISM  Saccharomyces cerevisiae S288c
            Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina;
            Saccharomycetes; Saccharomycetales; Saccharomycetaceae;
            Saccharomyces.
REFERENCE   1  (residues 1 to 589)
  AUTHORS   Tettelin,H., Agostoni Carbone,M.L., Albermann,K., Albers,M.,
            Arroyo,J., Backes,U., Barreiros,T., Bertani,I., Bjourson,A.J.,
            Bruckner,M., Bruschi,C.V., Carignani,G., Castagnoli,L., Cerdan,E.,
            Clemente,M.L., Coblenz,A., Coglievina,M., Coissac,E., Defoor,E.,
```

CELLO



cello.life.nctu.edu.tw

MBC

Molecular Bioinformatics Center

National Chiao Tung University

About CELLO

CELLO v.2.5: subCELlular LOcalization predictor

ORGANISMS

● Gram negative

○ Gram positive

○ Eukaryotes

SEQUENCES

○ DNA

● Protein

Paste the query sequences in FASTA format below

```
>1084005|Genbank|Outer membrane/Extracellular (Autotransporter)|major ring-
forming surface protein precursor
[illegible]
```

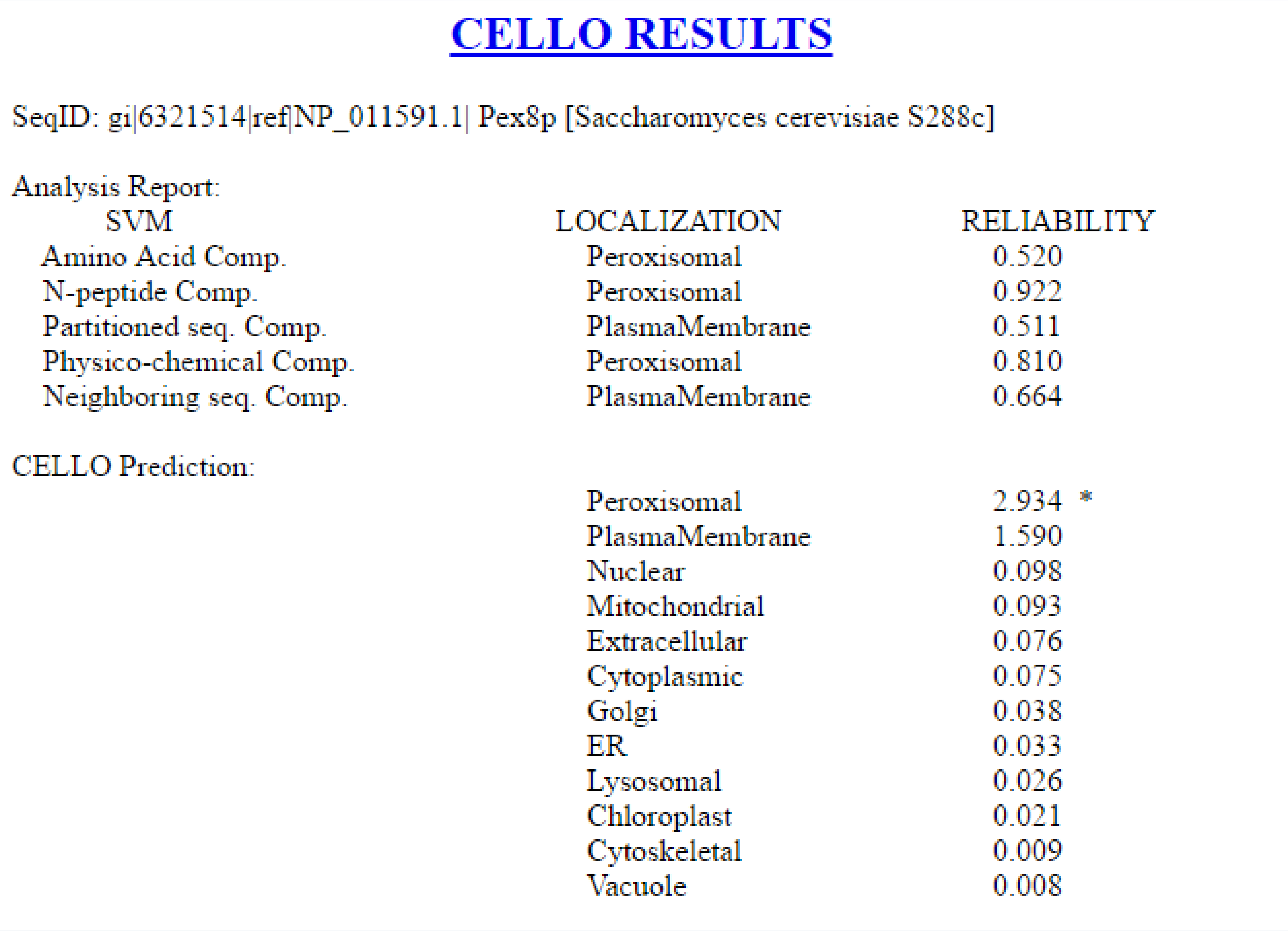
Or upload from file: Choose File No file chosen

Reset Submit

1. Choose Eukaryotes

2. Delete Example Protein!

http://cello.life.nctu.edu.tw/



CELLO RESULTS

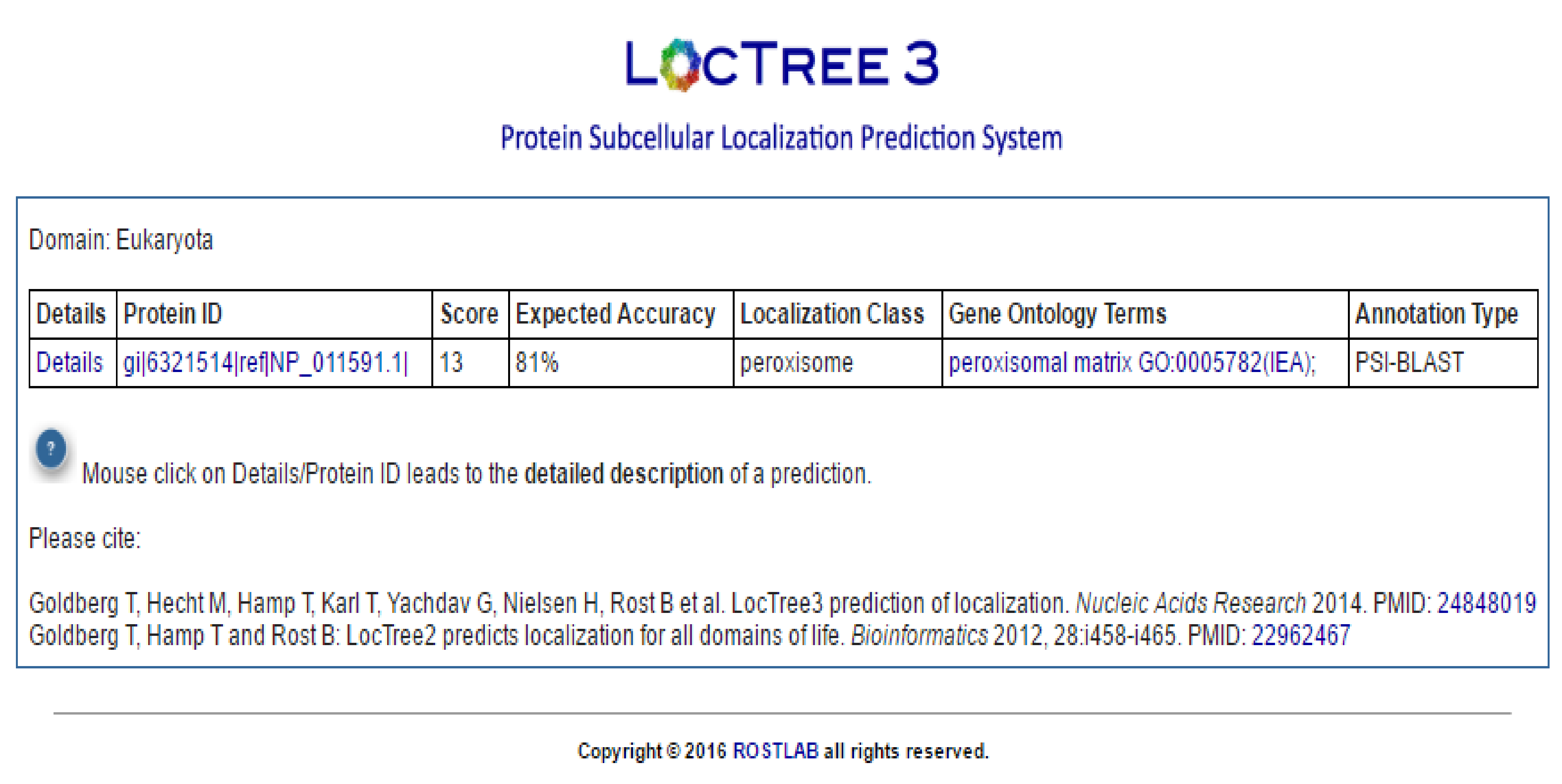
SeqID: gi|6321514|ref|NP_011591.1| Pex8p [Saccharomyces cerevisiae S288c]

Analysis Report:

SVM	LOCALIZATION	RELIABILITY
Amino Acid Comp.	Peroxisomal	0.520
N-peptide Comp.	Peroxisomal	0.922
Partitioned seq. Comp.	PlasmaMembrane	0.511
Physico-chemical Comp.	Peroxisomal	0.810
Neighboring seq. Comp.	PlasmaMembrane	0.664

CELLO Prediction:

	Peroxisomal	2.934	*
	PlasmaMembrane	1.590	
	Nuclear	0.098	
	Mitochondrial	0.093	
	Extracellular	0.076	
	Cytoplasmic	0.075	
	Golgi	0.038	
	ER	0.033	
	Lysosomal	0.026	
	Chloroplast	0.021	
	Cytoskeletal	0.009	
	Vacuole	0.008	



LOCTREE 3

Protein Subcellular Localization Prediction System

Domain: Eukaryota

Details	Protein ID	Score	Expected Accuracy	Localization Class	Gene Ontology Terms	Annotation Type
Details	gi|6321514|ref|NP_011591.1|	13	81%	peroxisome	peroxisomal matrix GO:0005782(IEA);	PSI-BLAST

? Mouse click on Details/Protein ID leads to the **detailed description** of a prediction.

Please cite:

Goldberg T, Hecht M, Hamp T, Karl T, Yachdav G, Nielsen H, Rost B et al. LocTree3 prediction of localization. *Nucleic Acids Research* 2014. PMID: 24848019
Goldberg T, Hamp T and Rost B: LocTree2 predicts localization for all domains of life. *Bioinformatics* 2012, 28:i458-i465. PMID: 22962467

https://rostlab.org/services/loctree2/

Analyzing Local Properties

- Many local properties are important for the function of your protein
  - Hydrophobic regions are potential transmembrane domains
  - Coiled-coiled regions are potential protein-interaction domains
  - Hydrophilic stretches are potential loops

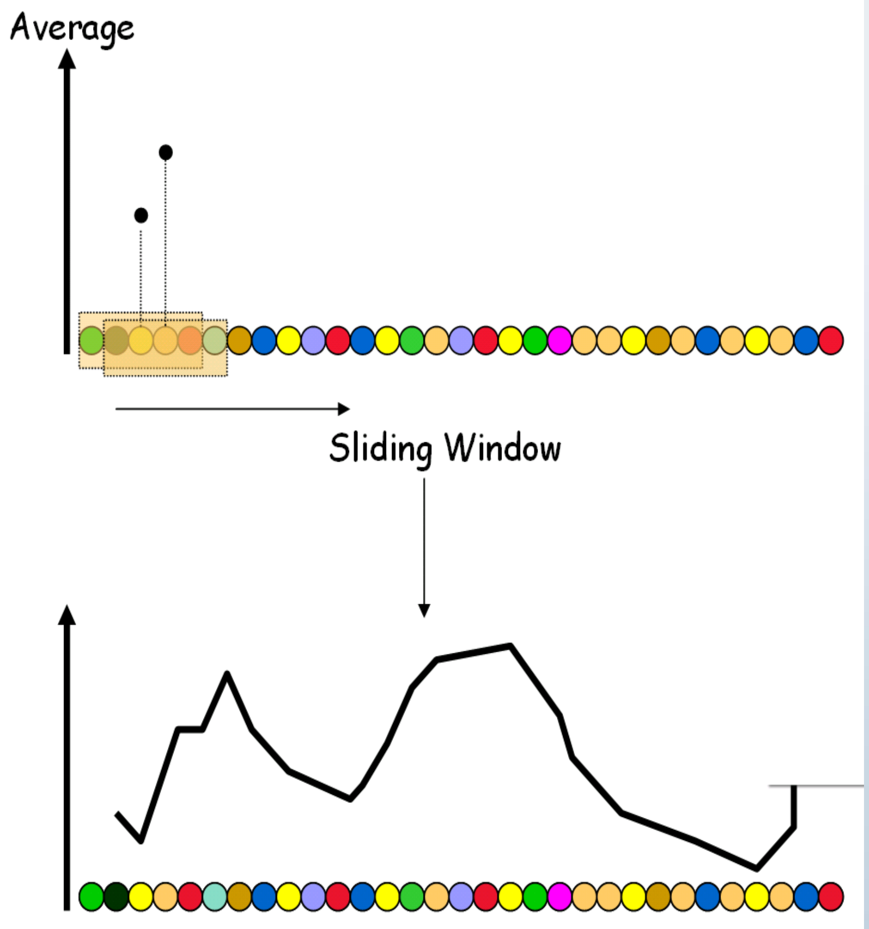
- You can discover these regions
  - Using sliding-widow techniques (easy)
  - Using prediction methods such as hidden Markov Models (more sophisticated)

Sliding-window Techniques

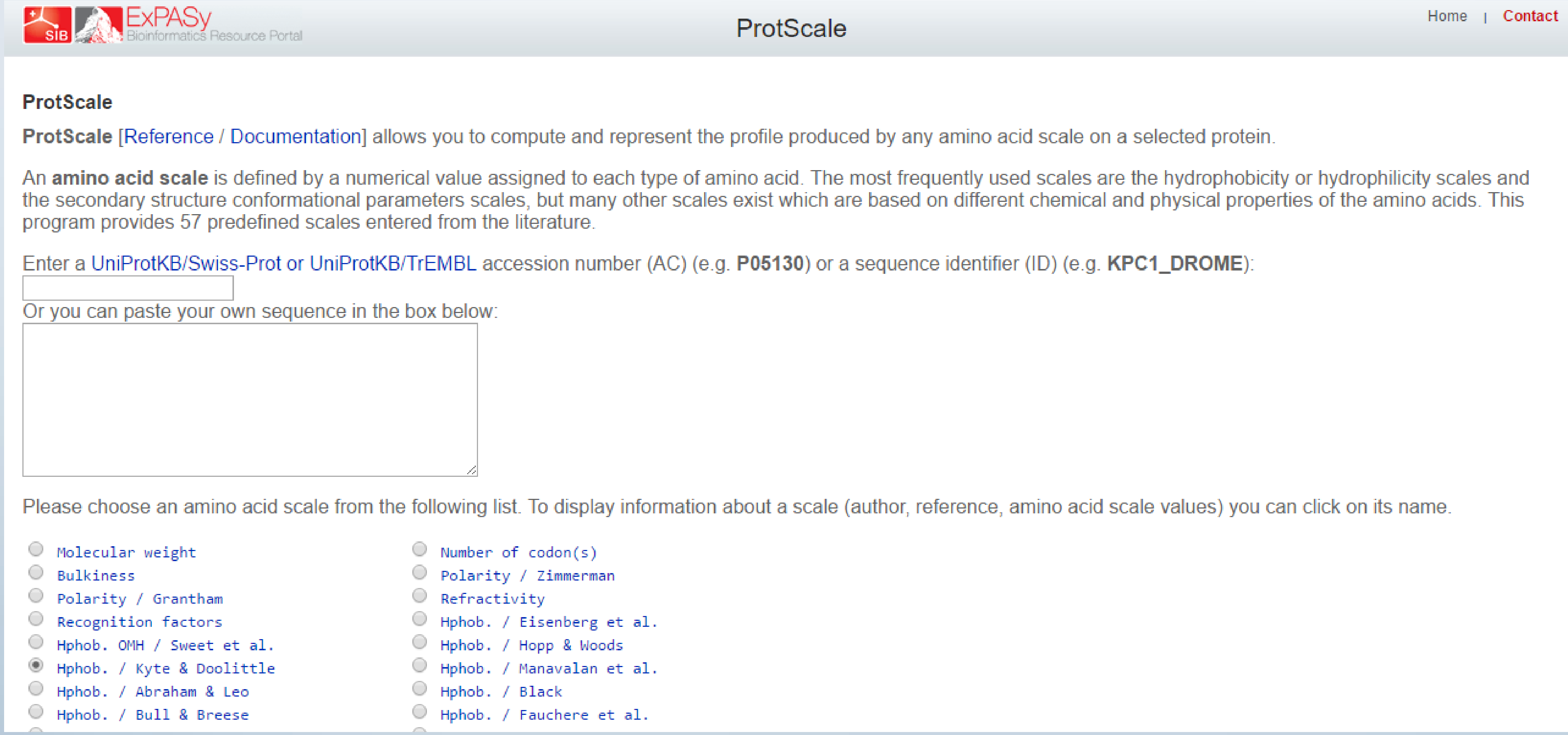
- Ideal for identifying strong signals

- Very simple methods
  - Few artifacts
  - Not very sensitive

- Make the window the same size as the feature you’re looking for



ProtScale



ProtScale

ProtScale [Reference / Documentation] allows you to compute and represent the profile produced by any amino acid scale on a selected protein.

An **amino acid scale** is defined by a numerical value assigned to each type of amino acid. The most frequently used scales are the hydrophobicity or hydrophilicity scales and the secondary structure conformational parameters scales, but many other scales exist which are based on different chemical and physical properties of the amino acids. This program provides 57 predefined scales entered from the literature.

Enter a UniProtKB/Swiss-Prot or UniProtKB/TrEMBL accession number (AC) (e.g. **P05130**) or a sequence identifier (ID) (e.g. **KPC1_DROME**):

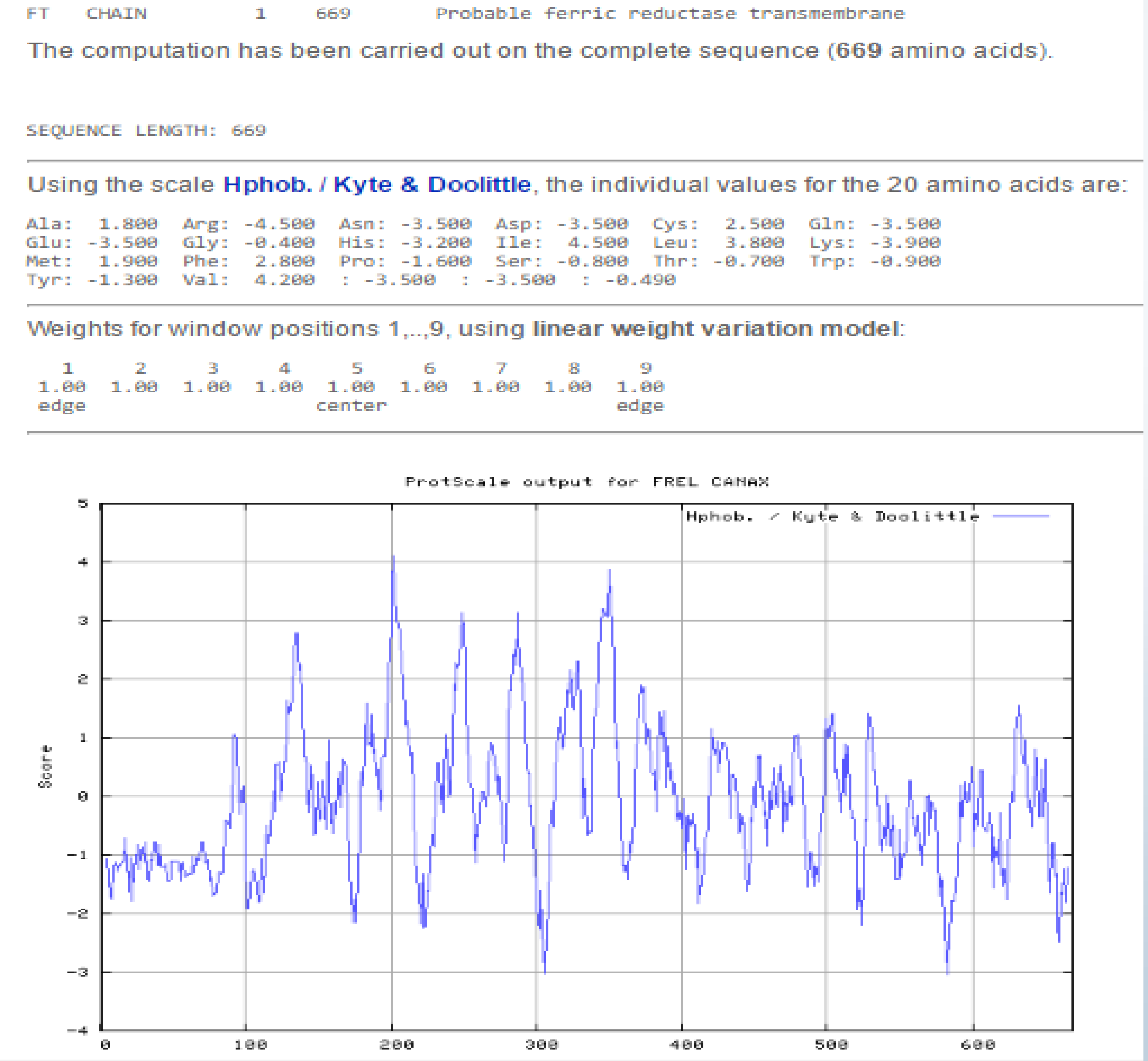
Or you can paste your own sequence in the box below:

Please choose an amino acid scale from the following list. To display information about a scale (author, reference, amino acid scale values) you can click on its name.

- ○ Molecular weight
- ○ Bulkiness
- ○ Polarity / Grantham
- ○ Recognition factors
- ○ Hphob. OMH / Sweet et al.
- ● Hphob. / Kyte & Doolittle
- ○ Hphob. / Abraham & Leo
- ○ Hphob. / Bull & Breese
- ○ Number of codon(s)
- ○ Polarity / Zimmerman
- ○ Refractivity
- ○ Hphob. / Eisenberg et al.
- ○ Hphob. / Hopp & Woods
- ○ Hphob. / Manavalan et al.
- ○ Hphob. / Black
- ○ Hphob. / Fauchere et al.

http://web.expasy.org/protscale/

ProtScale – Example (P78588)



```
FT   CHAIN         1     669       Probable ferric reductase transmembrane
```

The computation has been carried out on the complete sequence (**669** amino acids).

```
SEQUENCE LENGTH: 669
```

Using the scale **Hphob. / Kyte & Doolittle**, the individual values for the 20 amino acids are:

```
Ala:  1.800  Arg: -4.500  Asn: -3.500  Asp: -3.500  Cys:  2.500  Gln: -3.500
Glu: -3.500  Gly: -0.400  His: -3.200  Ile:  4.500  Leu:  3.800  Lys: -3.900
Met:  1.900  Phe:  2.800  Pro: -1.600  Ser: -0.800  Thr: -0.700  Trp: -0.900
Tyr: -1.300  Val:  4.200   : -3.500   : -3.500   : -0.490
```

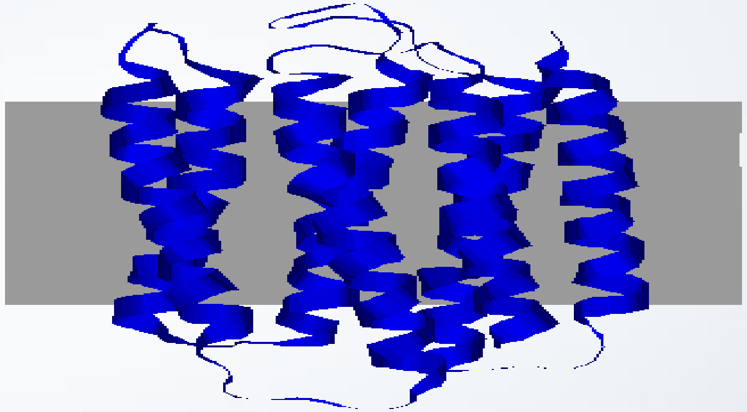
Weights for window positions 1,..,9, using **linear weight variation model**:

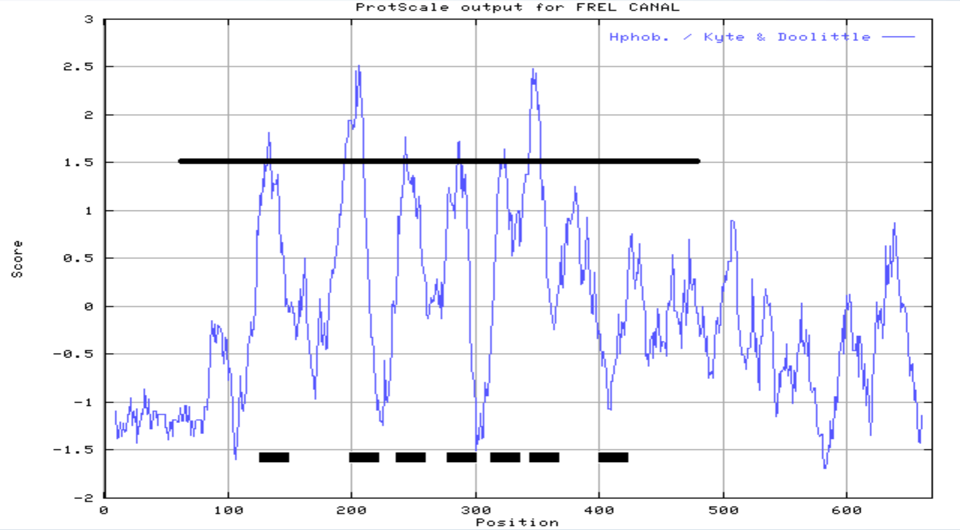
```
  1     2     3     4     5     6     7     8     9
 1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00
 edge                   center                   edge
```

http://web.expasy.org/protscale/

Transmembrane Domains

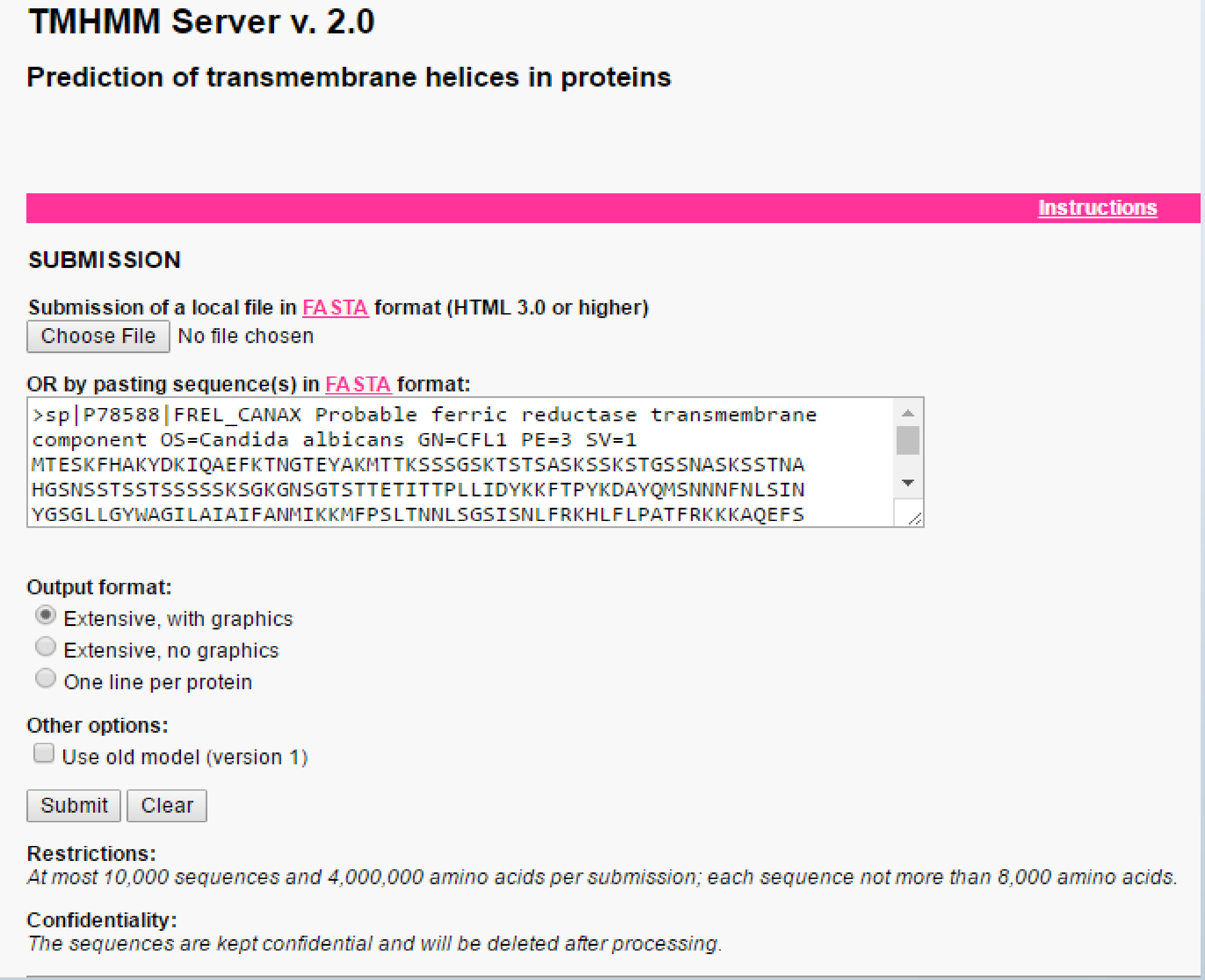
- Discovering a transmembrane domain tells you a lot about your protein
- Many important receptors have 7 transmembrane domains
- Transmembrane segments can be found using ProtScale
- The most accurate predictions come from using TMHMM





Transmembrane Domains – TMHMM Method

- TMHMM is the best method for predicting transmembrane domains
- TMHMM uses an HMM
- Its principle is very different from that of ProtScale
- TMHMM output is a prediction



TMHMM Server v. 2.0

Prediction of transmembrane helices in proteins

SUBMISSION

Submission of a local file in FASTA format (HTML 3.0 or higher)

Choose File No file chosen

OR by pasting sequence(s) in FASTA format:

```
>sp|P78588|FREL_CANAX Probable ferric reductase transmembrane
component OS=Candida albicans GN=CFL1 PE=3 SV=1
MTESKFHAKYDKIQAEFKTNGTEYAKMTTKSSSGSKTSTSASKSSKSTGSSNASKSSTNA
HGSNSSTSSTSSSSSKSGKGNSGTSTTETITTPLLIDYKKFTPYKDAYQMSNNNFNLSIN
YGSGLLGYWAGILAIAIFANMIKKMFPSLTNNLSGSISNLFRKHLFLPATFRKKKAQEFS
```

Output format:

● Extensive, with graphics

○ Extensive, no graphics

○ One line per protein

Other options:

□ Use old model (version 1)

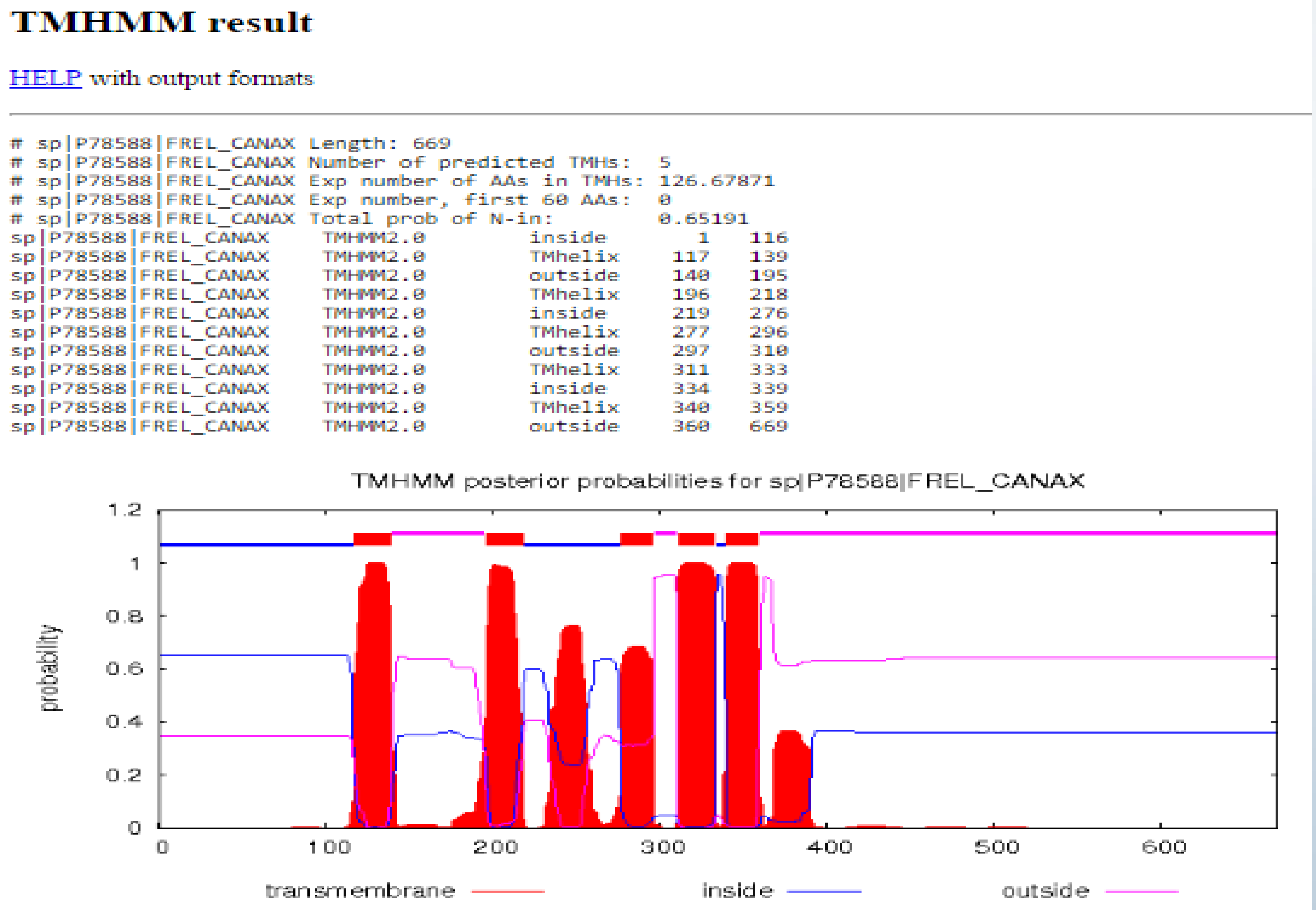
Submit Clear

Restrictions:
At most 10,000 sequences and 4,000,000 amino acids per submission; each sequence not more than 8,000 amino acids.

Confidentiality:
The sequences are kept confidential and will be deleted after processing.

http://www.cbs.dtu.dk/services/TMHMM/

Predicting Transmembrane using TMHMM Example (P78588)

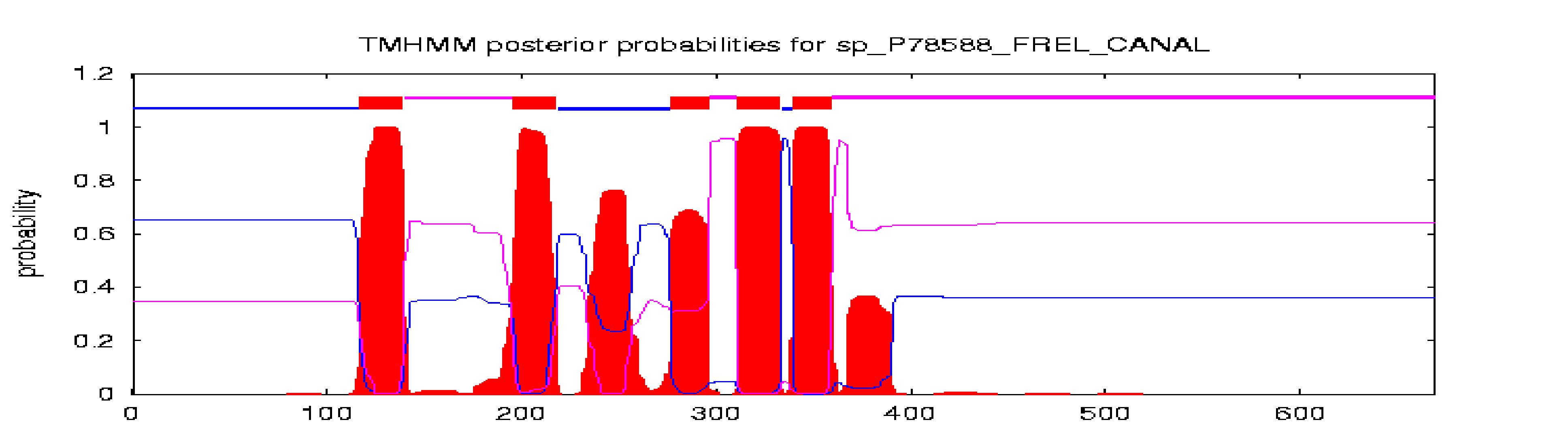


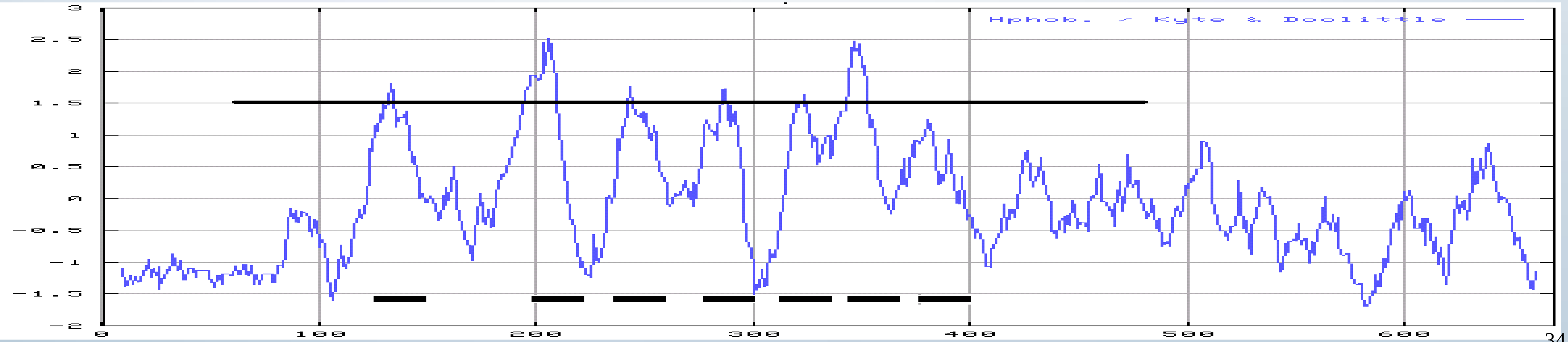
TMHMM result

HELP with output formats

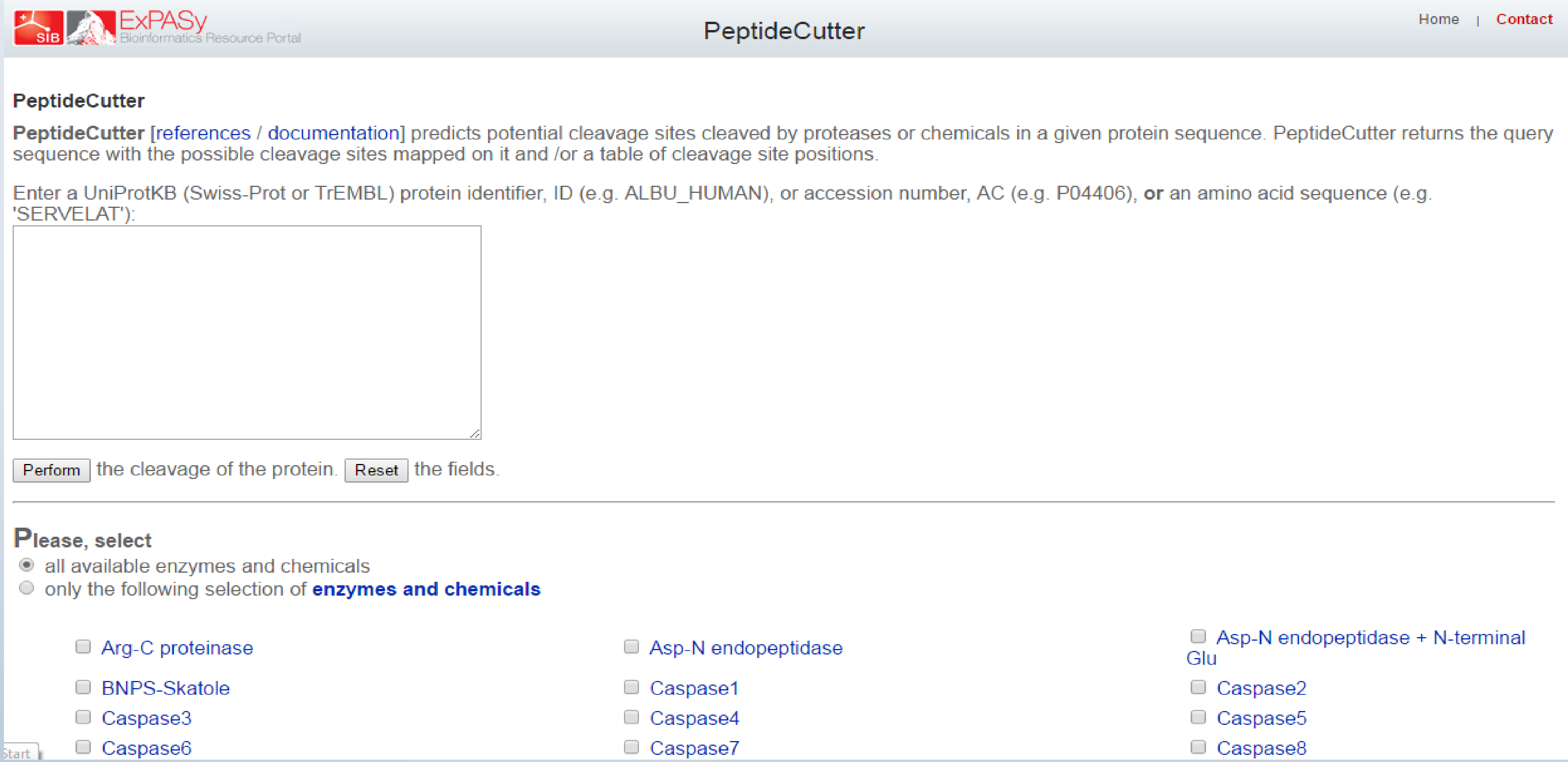
```
# sp|P78588|FREL_CANAX Length: 669
# sp|P78588|FREL_CANAX Number of predicted TMHs:  5
# sp|P78588|FREL_CANAX Exp number of AAs in TMHs: 126.67871
# sp|P78588|FREL_CANAX Exp number, first 60 AAs:  0
# sp|P78588|FREL_CANAX Total prob of N-in:        0.65191
sp|P78588|FREL_CANAX    TMHMM2.0        inside       1    116
sp|P78588|FREL_CANAX    TMHMM2.0        TMhelix    117    139
sp|P78588|FREL_CANAX    TMHMM2.0        outside    140    195
sp|P78588|FREL_CANAX    TMHMM2.0        TMhelix    196    218
sp|P78588|FREL_CANAX    TMHMM2.0        inside     219    276
sp|P78588|FREL_CANAX    TMHMM2.0        TMhelix    277    296
sp|P78588|FREL_CANAX    TMHMM2.0        outside    297    310
sp|P78588|FREL_CANAX    TMHMM2.0        TMhelix    311    333
sp|P78588|FREL_CANAX    TMHMM2.0        inside     334    339
sp|P78588|FREL_CANAX    TMHMM2.0        TMhelix    340    359
sp|P78588|FREL_CANAX    TMHMM2.0        outside    360    669
```

TMHMM vs. ProtScale





Potential Cleavage Sites



PeptideCutter

PeptideCutter [references / documentation] predicts potential cleavage sites cleaved by proteases or chemicals in a given protein sequence. PeptideCutter returns the query sequence with the possible cleavage sites mapped on it and /or a table of cleavage site positions.

Enter a UniProtKB (Swiss-Prot or TrEMBL) protein identifier, ID (e.g. ALBU_HUMAN), or accession number, AC (e.g. P04406), **or** an amino acid sequence (e.g. 'SERVELAT'):

Perform the cleavage of the protein. Reset the fields.

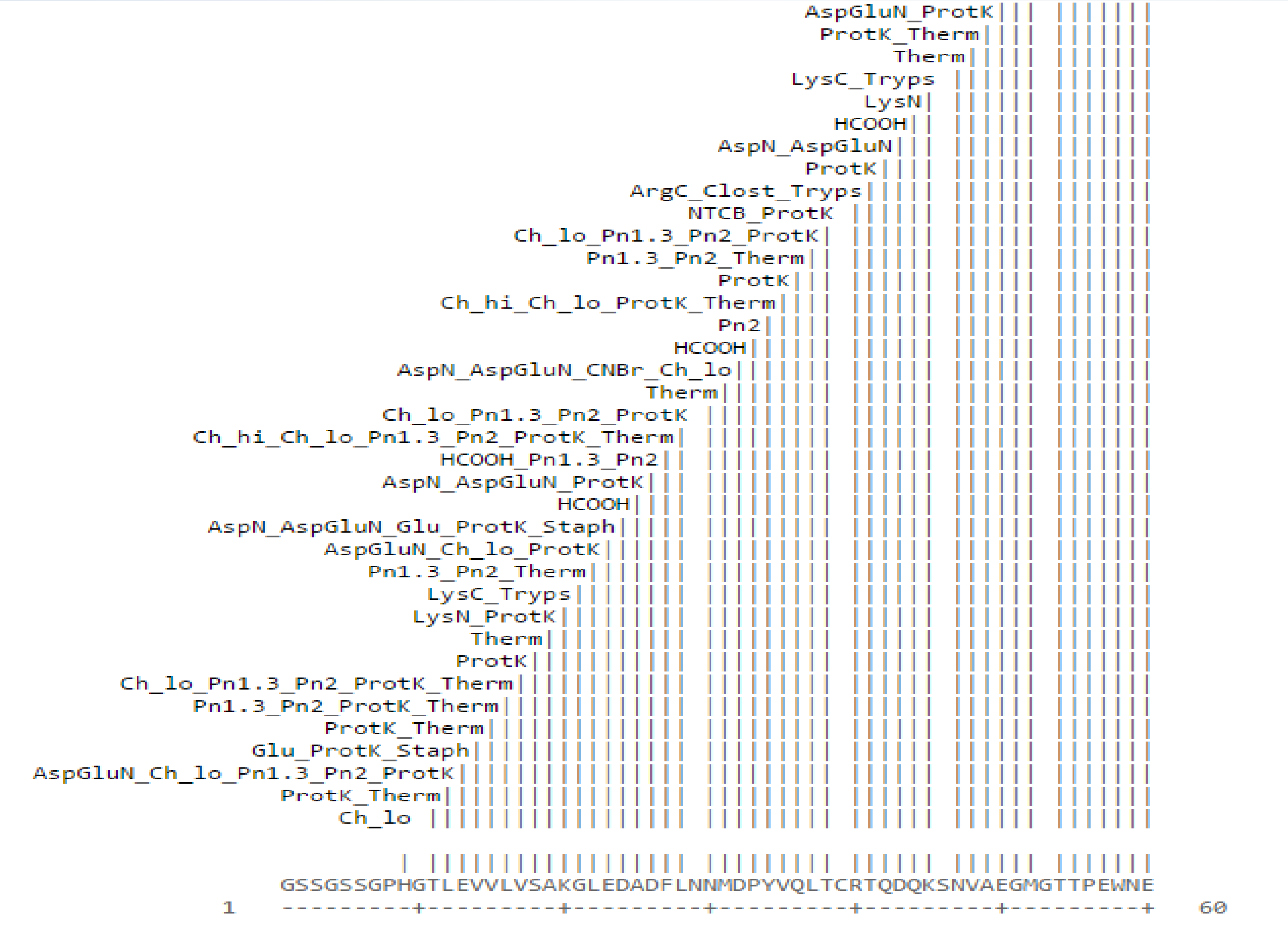
Please, select

● all available enzymes and chemicals
○ only the following selection of **enzymes and chemicals**

□ Arg-C proteinase
□ Asp-N endopeptidase
□ Asp-N endopeptidase + N-terminal Glu
□ BNPS-Skatole
□ Caspase1
□ Caspase2
□ Caspase3
□ Caspase4
□ Caspase5
□ Caspase6
□ Caspase7
□ Caspase8

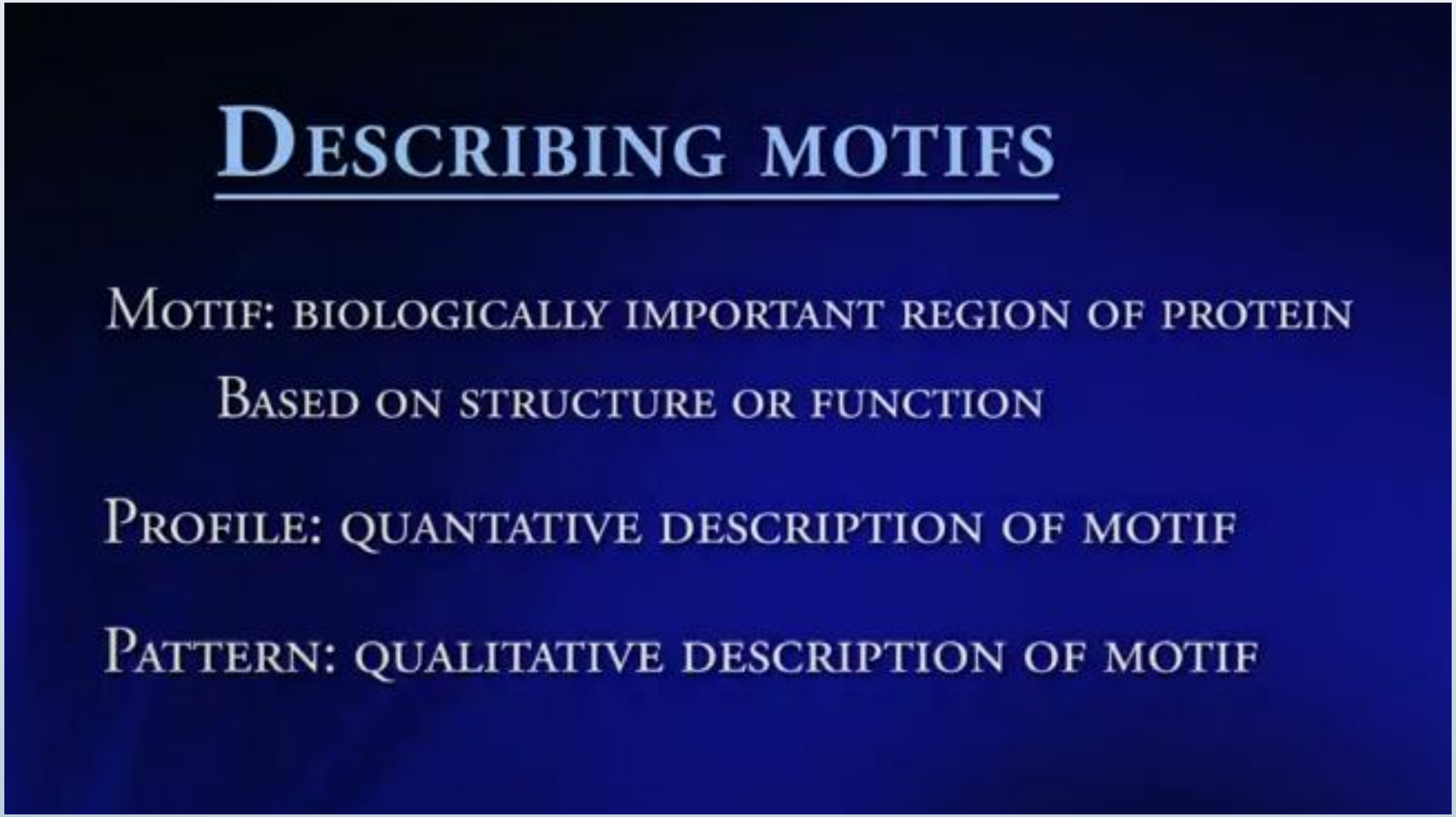
http://web.expasy.org/peptide_cutter/

Example (plant protein) - Results



```
                                                        AspGluN_ProtK||| |||||||
                                                         ProtK_Therm|||| |||||||
                                                              Therm||||| |||||||
                                                       LysC_Tryps |||||| |||||||
                                                            LysN| |||||| |||||||
                                                          HCOOH|| |||||| |||||||
                                                  AspN_AspGluN||| |||||| |||||||
                                                        ProtK|||| |||||| |||||||
                                            ArgC_Clost_Tryps||||| |||||| |||||||
                                                NTCB_ProtK |||||| |||||| |||||||
                                    Ch_lo_Pn1.3_Pn2_ProtK| |||||| |||||| |||||||
                                         Pn1.3_Pn2_Therm|| |||||| |||||| |||||||
                                                  ProtK||| |||||| |||||| |||||||
                               Ch_hi_Ch_lo_ProtK_Therm|||| |||||| |||||| |||||||
                                                  Pn2||||| |||||| |||||| |||||||
                                               HCOOH|||||| |||||| |||||| |||||||
                            AspN_AspGluN_CNBr_Ch_lo||||||| |||||| |||||| |||||||
                                             Therm|||||||| |||||| |||||| |||||||
                           Ch_lo_Pn1.3_Pn2_ProtK ||||||||| |||||| |||||| |||||||
              Ch_hi_Ch_lo_Pn1.3_Pn2_ProtK_Therm| ||||||||| |||||| |||||| |||||||
                               HCOOH_Pn1.3_Pn2|| ||||||||| |||||| |||||| |||||||
                           AspN_AspGluN_ProtK||| ||||||||| |||||| |||||| |||||||
                                       HCOOH|||| ||||||||| |||||| |||||| |||||||
               AspN_AspGluN_Glu_ProtK_Staph||||| ||||||||| |||||| |||||| |||||||
                       AspGluN_Ch_lo_ProtK|||||| ||||||||| |||||| |||||| |||||||
                          Pn1.3_Pn2_Therm||||||| ||||||||| |||||| |||||| |||||||
                              LysC_Tryps|||||||| ||||||||| |||||| |||||| |||||||
                             LysN_ProtK||||||||| ||||||||| |||||| |||||| |||||||
                                 Therm|||||||||| ||||||||| |||||| |||||| |||||||
                                ProtK||||||||||| ||||||||| |||||| |||||| |||||||
         Ch_lo_Pn1.3_Pn2_ProtK_Therm|||||||||||| ||||||||| |||||| |||||| |||||||
              Pn1.3_Pn2_ProtK_Therm||||||||||||| ||||||||| |||||| |||||| |||||||
                       ProtK_Therm|||||||||||||| ||||||||| |||||| |||||| |||||||
                  Glu_ProtK_Staph||||||||||||||| ||||||||| |||||| |||||| |||||||
   AspGluN_Ch_lo_Pn1.3_Pn2_ProtK|||||||||||||||| ||||||||| |||||| |||||| |||||||
                    ProtK_Therm||||||||||||||||| ||||||||| |||||| |||||| |||||||
                        Ch_lo |||||||||||||||||| ||||||||| |||||| |||||| |||||||

                            | |||||||||||||||||| ||||||||| |||||| |||||| |||||||
                    GSSGSSGPHGTLEVVLVSAKGLEDADFLNNMDPYVQLTCRTQDQKSNVAEGMGTTPEWNE
                1   ---------+---------+---------+---------+---------+---------+   60
```

Motifs – Patterns



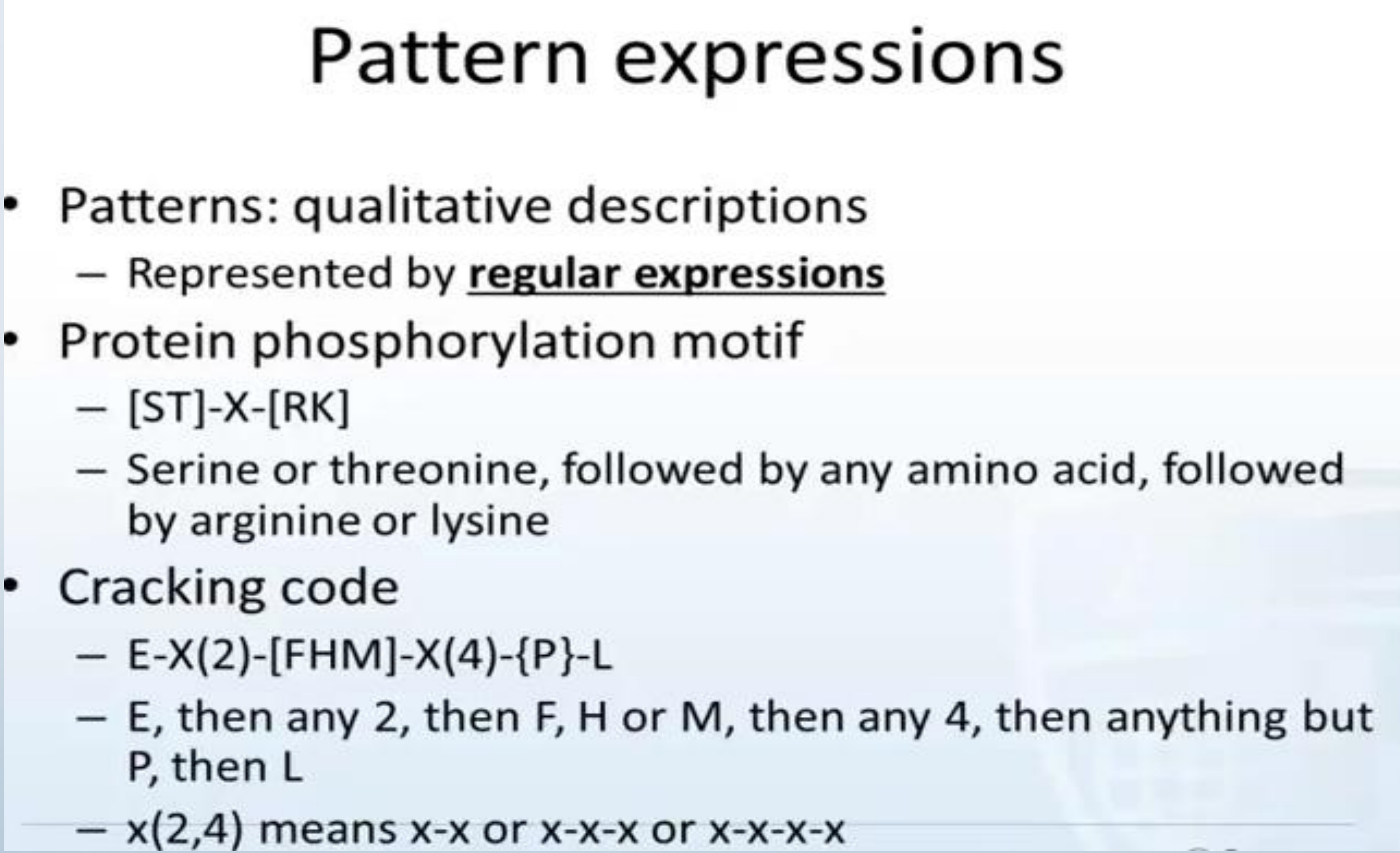
Describing motifs

Motif: biologically important region of protein

Based on structure or function

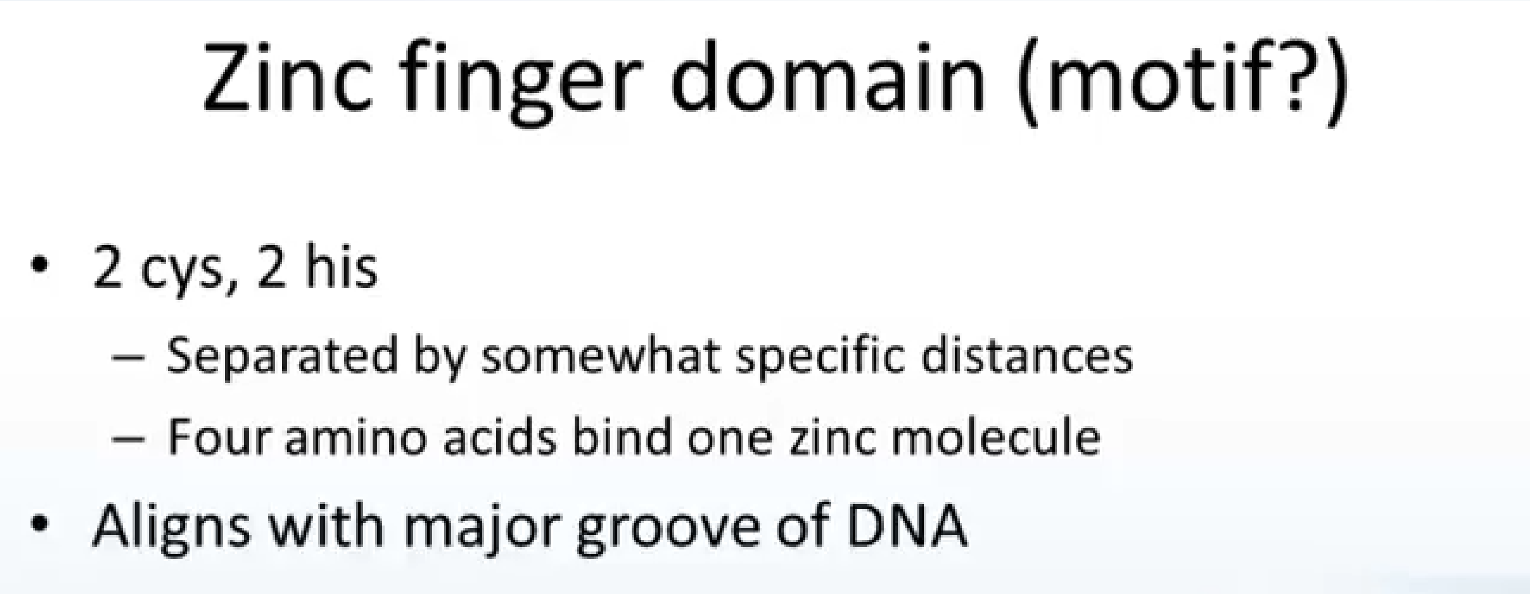
Profile: quantative description of motif

Pattern: qualitative description of motif



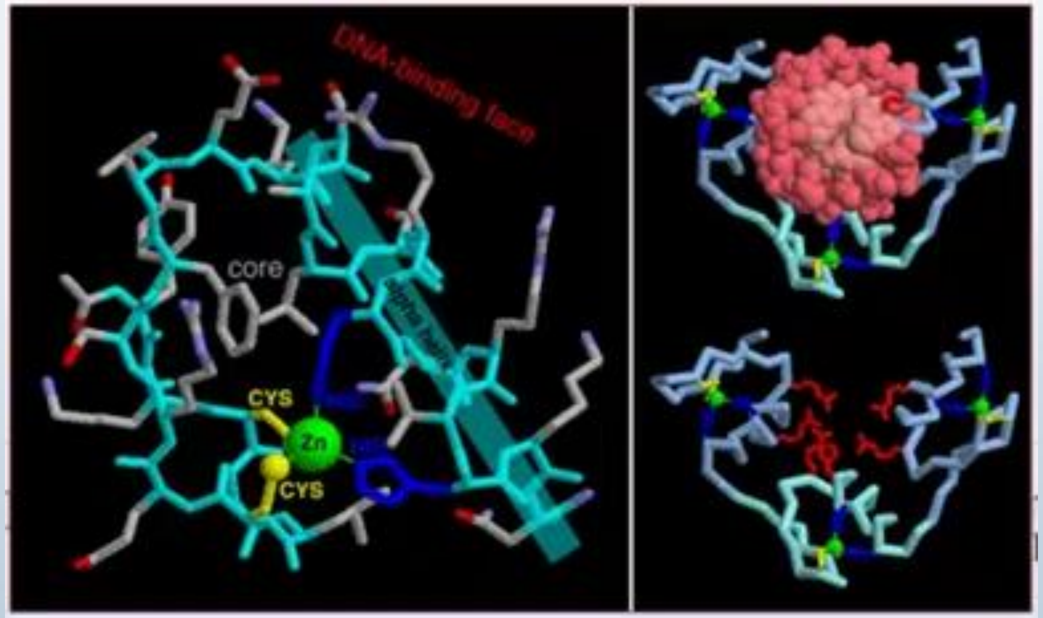
Pattern expressions

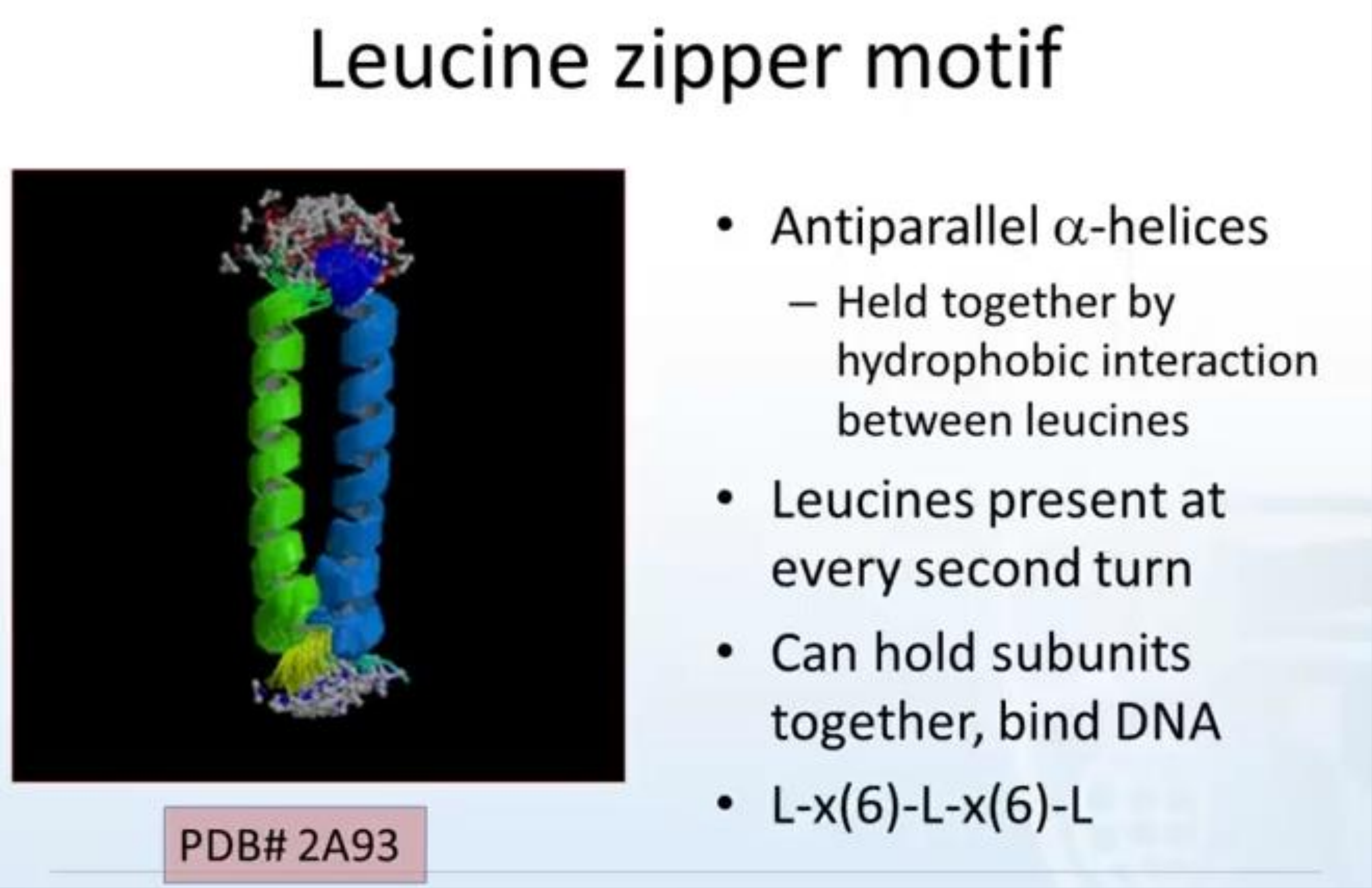
- Patterns: qualitative descriptions
  - Represented by ***regular expressions***
- Protein phosphorylation motif
  - [ST]-X-[RK]
  - Serine or threonine, followed by any amino acid, followed by arginine or lysine
- Cracking code
  - E-X(2)-[FHM]-X(4)-{P}-L
  - E, then any 2, then F, H or M, then any 4, then anything but P, then L
  - x(2,4) means x-x or x-x-x or x-x-x-x



Zinc finger domain (motif?)

- 2 cys, 2 his
  - Separated by somewhat specific distances
  - Four amino acids bind one zinc molecule
- Aligns with major groove of DNA



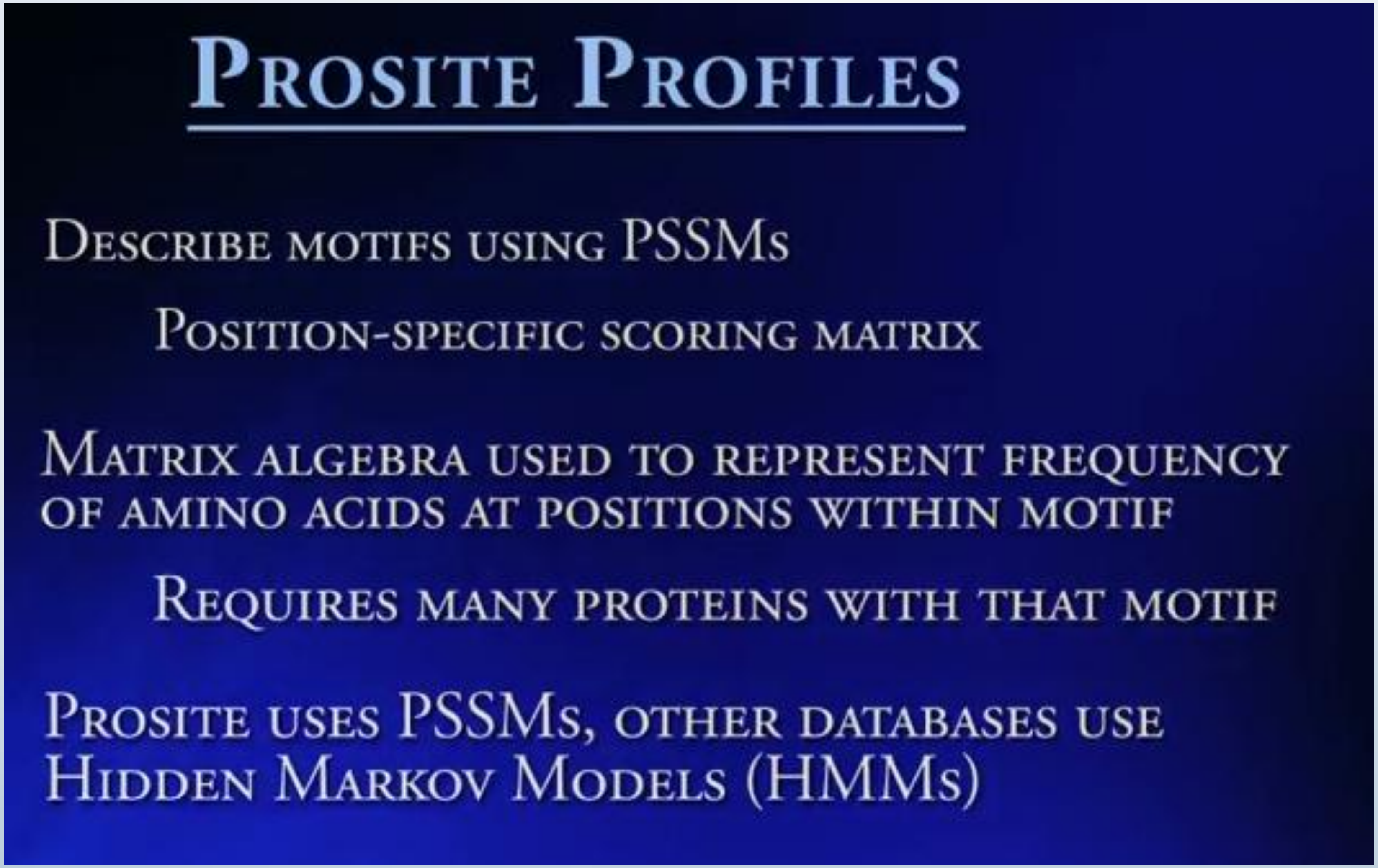


Leucine zipper motif

- Antiparallel α-helices
  - Held together by hydrophobic interaction between leucines
- Leucines present at every second turn
- Can hold subunits together, bind DNA
- L-x(6)-L-x(6)-L

PDB# 2A93

Motifs - PROSITE



Prosite Profiles

Describe motifs using PSSMs

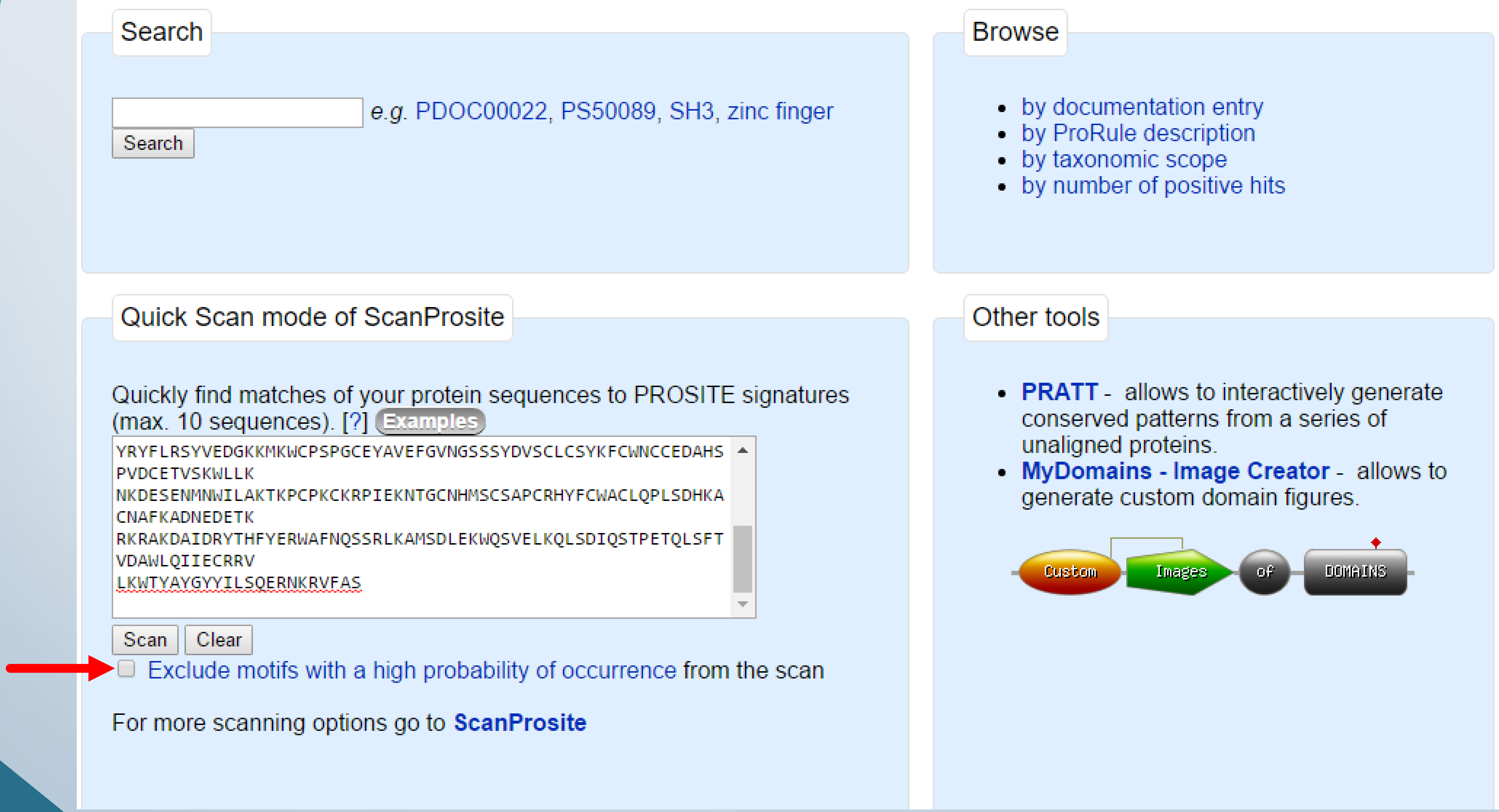
- Position-specific scoring matrix

Matrix algebra used to represent frequency of amino acids at positions within motif

- Requires many proteins with that motif

Prosite uses PSSMs, other databases use Hidden Markov Models (HMMs)

Example (NP_180737.3)



Search

e.g. PDOC00022, PS50089, SH3, zinc finger

Browse

- by documentation entry
- by ProRule description
- by taxonomic scope
- by number of positive hits

Quick Scan mode of ScanProsite

Quickly find matches of your protein sequences to PROSITE signatures (max. 10 sequences). [?] Examples

```
YRYFLRSYVEDGKKMKWCPSPGCEYAVEFGVNGSSSYDVSCLCSYKFCWNCCEDAHS
PVDCETVSKWLLK
NKDESENMNWILAKTKPCPKCKRPIEKNTGCNHMSCSAPCRHYFCWACLQPLSDHKA
CNAFKADNEDETK
RKRAKDAIDRYTHFYERWAFNQSSRLKAMSDLEKWQSVELKQLSDIQSTPETQLSFT
VDAWLQIIECRRV
LKWTYAYGYYILSQERNKRVFAS
```

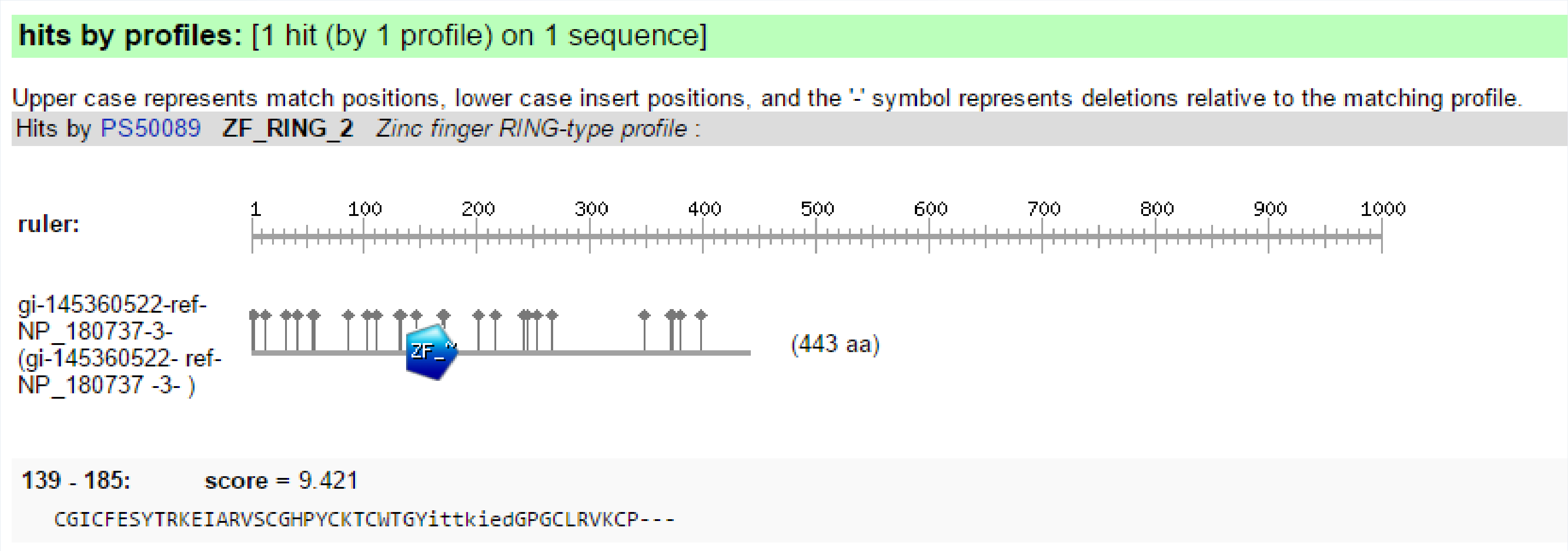
□ Exclude motifs with a high probability of occurrence from the scan

For more scanning options go to **ScanProsite**

Other tools

- **PRATT** - allows to interactively generate conserved patterns from a series of unaligned proteins.
- **MyDomains - Image Creator** - allows to generate custom domain figures.

http://prosite.expasy.org/



hits by profiles: [1 hit (by 1 profile) on 1 sequence]

Upper case represents match positions, lower case insert positions, and the '-' symbol represents deletions relative to the matching profile.

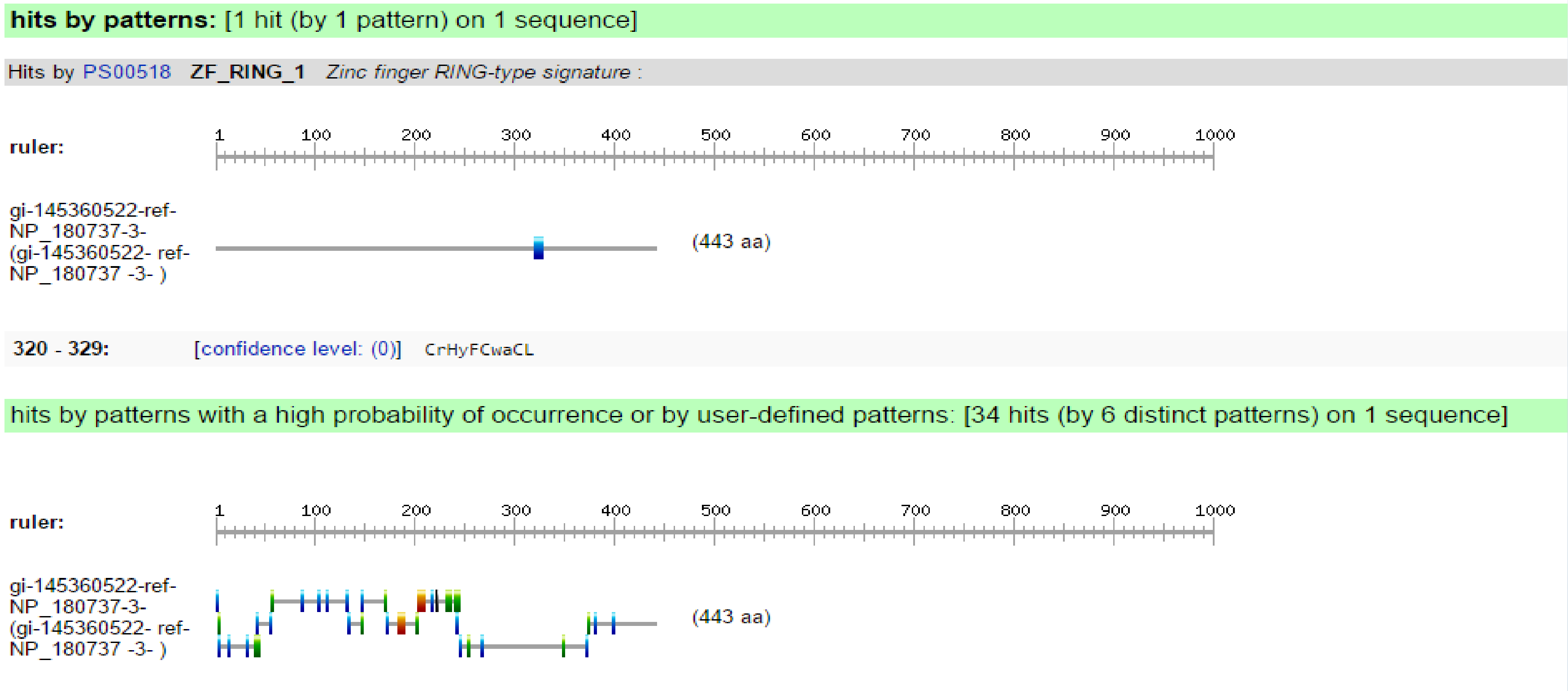
Hits by PS50089 **ZF_RING_2** *Zinc finger RING-type profile* :

ruler:

gi-145360522-ref-NP_180737-3- (gi-145360522- ref-NP_180737 -3-) (443 aa)

139 - 185: **score** = 9.421

```
CGICFESYTRKEIARVSCGHPYCKTCWTGYittkiedGPGCLRVKCP---
```



hits by patterns: [1 hit (by 1 pattern) on 1 sequence]

Hits by PS00518 **ZF_RING_1** *Zinc finger RING-type signature* :

ruler:

gi-145360522-ref-NP_180737-3- (gi-145360522- ref-NP_180737 -3-) (443 aa)

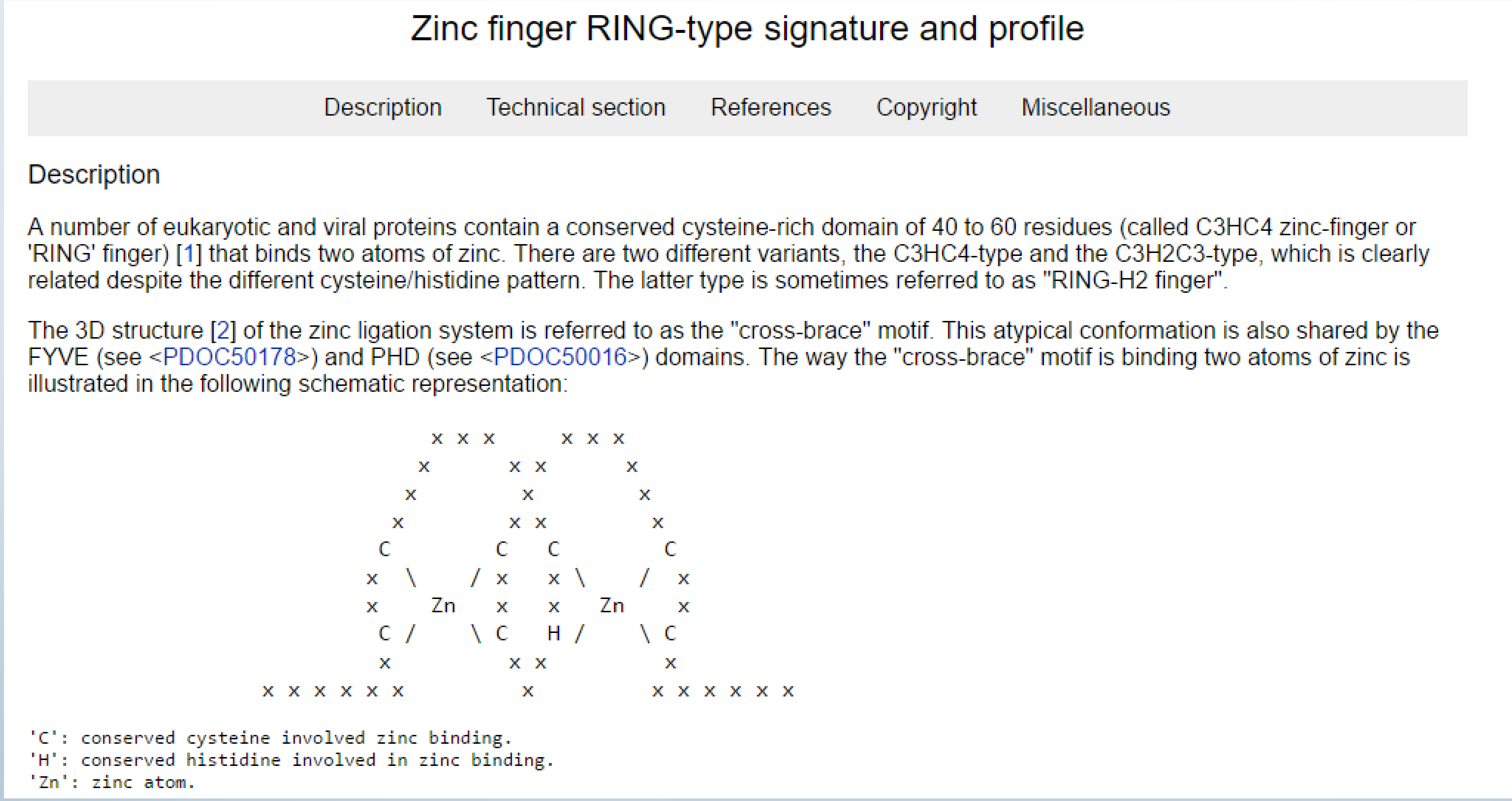
320 - 329: [confidence level: (0)] `CrHyFCwaCL`

hits by patterns with a high probability of occurrence or by user-defined patterns: [34 hits (by 6 distinct patterns) on 1 sequence]

ruler:

gi-145360522-ref-NP_180737-3- (gi-145360522- ref-NP_180737 -3-) (443 aa)

Result



Zinc finger RING-type signature and profile

Description Technical section References Copyright Miscellaneous

Description

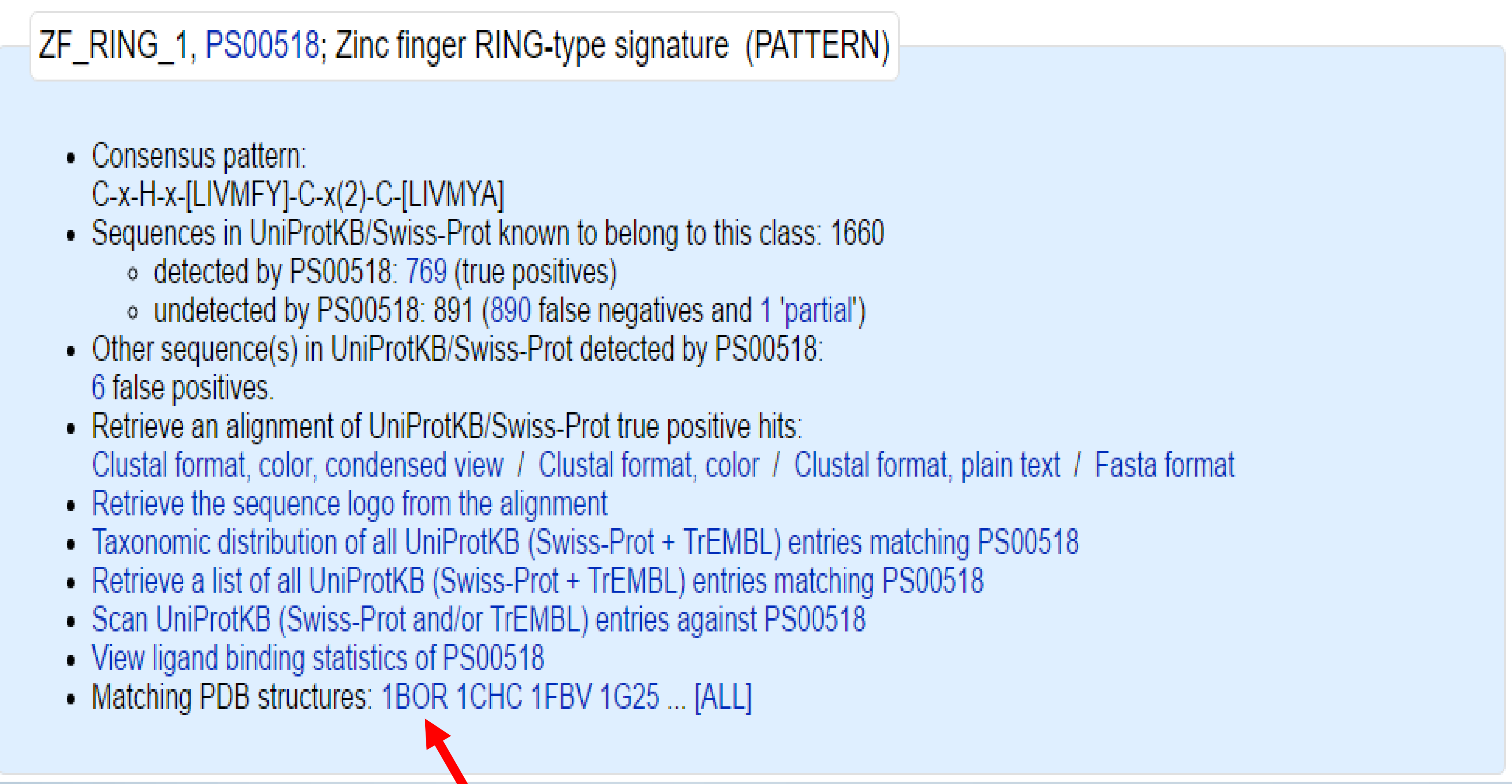
A number of eukaryotic and viral proteins contain a conserved cysteine-rich domain of 40 to 60 residues (called C3HC4 zinc-finger or 'RING' finger) [1] that binds two atoms of zinc. There are two different variants, the C3HC4-type and the C3H2C3-type, which is clearly related despite the different cysteine/histidine pattern. The latter type is sometimes referred to as "RING-H2 finger".

The 3D structure [2] of the zinc ligation system is referred to as the "cross-brace" motif. This atypical conformation is also shared by the FYVE (see <PDOC50178>) and PHD (see <PDOC50016>) domains. The way the "cross-brace" motif is binding two atoms of zinc is illustrated in the following schematic representation:

```
             x x x     x x x
            x      x x      x
           x        x        x
          x        x x        x
         C        C   C        C
        x  \    / x   x \    /  x
        x    Zn   x   x   Zn    x
         C /    \ C   H /    \ C
         x         x x         x
x x x x x x         x         x x x x x x
```

```
'C': conserved cysteine involved zinc binding.
'H': conserved histidine involved in zinc binding.
'Zn': zinc atom.
```

Result

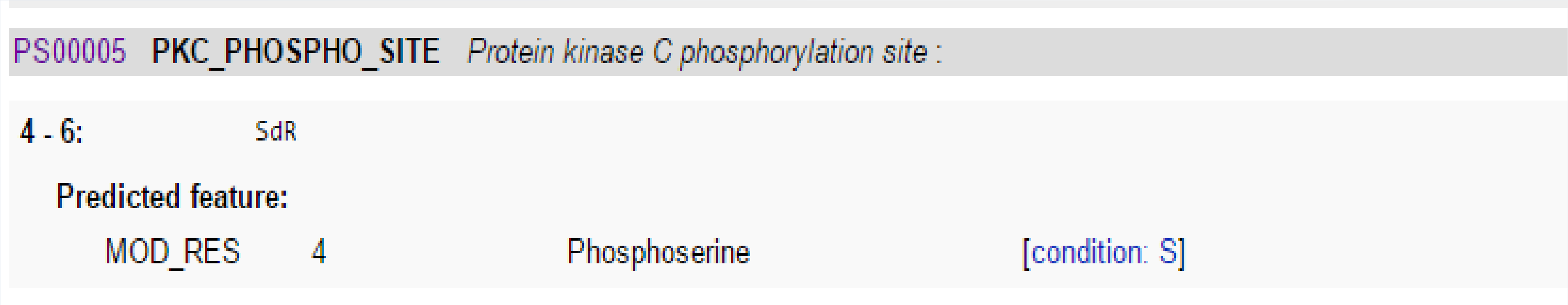


ZF_RING_1, PS00518; Zinc finger RING-type signature (PATTERN)

- Consensus pattern:
  C-x-H-x-[LIVMFY]-C-x(2)-C-[LIVMYA]
- Sequences in UniProtKB/Swiss-Prot known to belong to this class: 1660
  - detected by PS00518: 769 (true positives)
  - undetected by PS00518: 891 (890 false negatives and 1 'partial')
- Other sequence(s) in UniProtKB/Swiss-Prot detected by PS00518:
  6 false positives.
- Retrieve an alignment of UniProtKB/Swiss-Prot true positive hits:
  Clustal format, color, condensed view / Clustal format, color / Clustal format, plain text / Fasta format
- Retrieve the sequence logo from the alignment
- Taxonomic distribution of all UniProtKB (Swiss-Prot + TrEMBL) entries matching PS00518
- Retrieve a list of all UniProtKB (Swiss-Prot + TrEMBL) entries matching PS00518
- Scan UniProtKB (Swiss-Prot and/or TrEMBL) entries against PS00518
- View ligand binding statistics of PS00518
- Matching PDB structures: 1BOR 1CHC 1FBV 1G25 ... [ALL]

Link to view the Structure in JSmol

Result

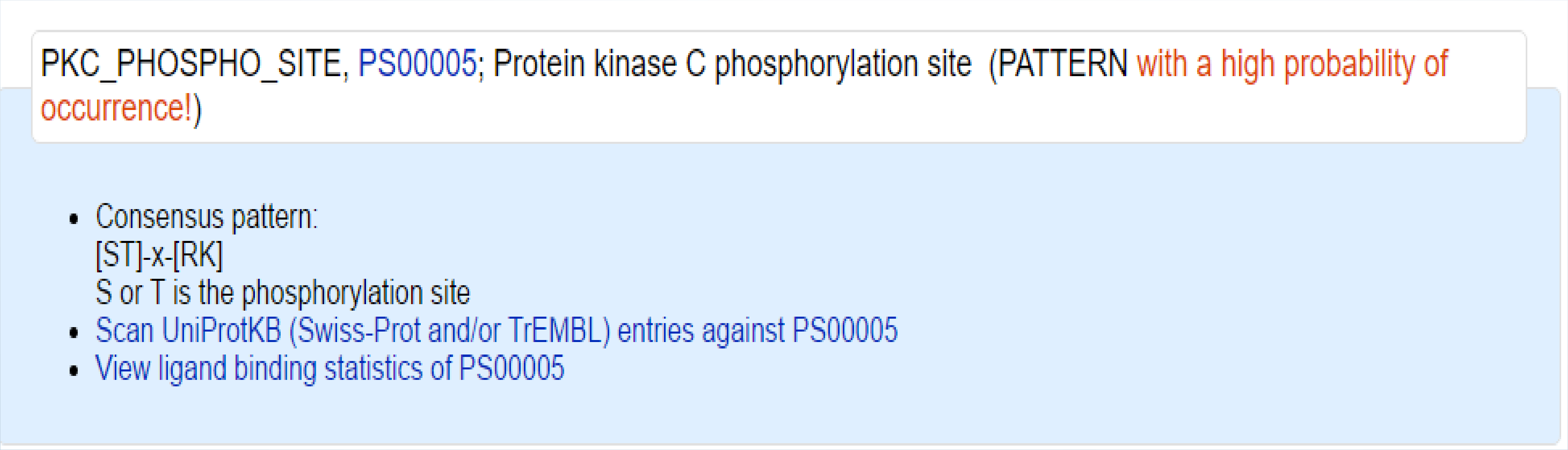


PS00005 **PKC_PHOSPHO_SITE** *Protein kinase C phosphorylation site :*

4 - 6: SdR

Predicted feature:

MOD_RES 4 Phosphoserine [condition: S]



PKC_PHOSPHO_SITE, PS00005; Protein kinase C phosphorylation site (PATTERN with a high probability of occurrence!)

- Consensus pattern:
  [ST]-x-[RK]
  S or T is the phosphorylation site
- Scan UniProtKB (Swiss-Prot and/or TrEMBL) entries against PS00005
- View ligand binding statistics of PS00005

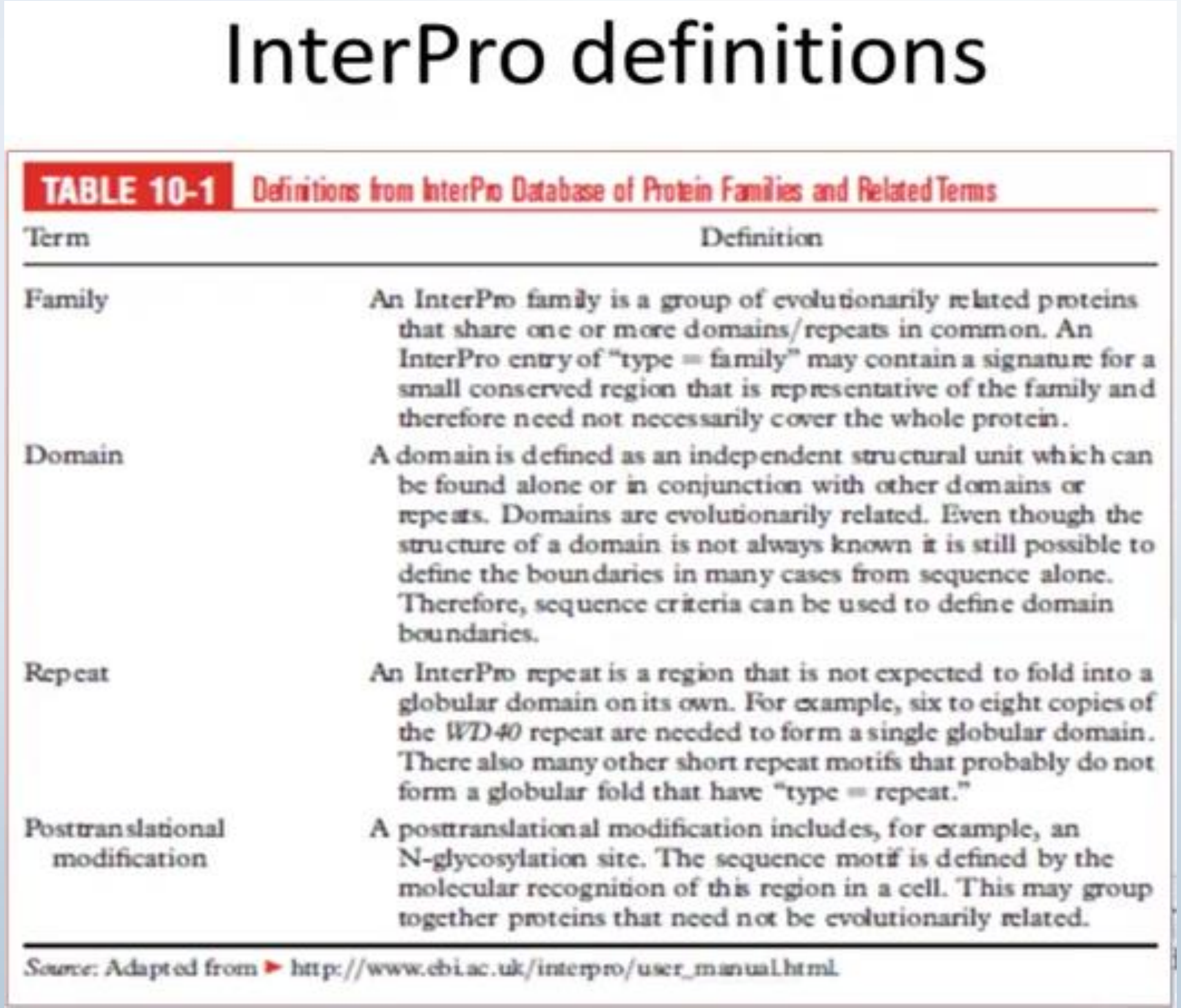
Patterns and Domains

- Patterns are usually the most striking feature of the more general motifs (called domains)

- Domains are less conserved than patterns but usually longer

- In proteins, domain analysis is gradually replacing pattern analysis

Different Definitions



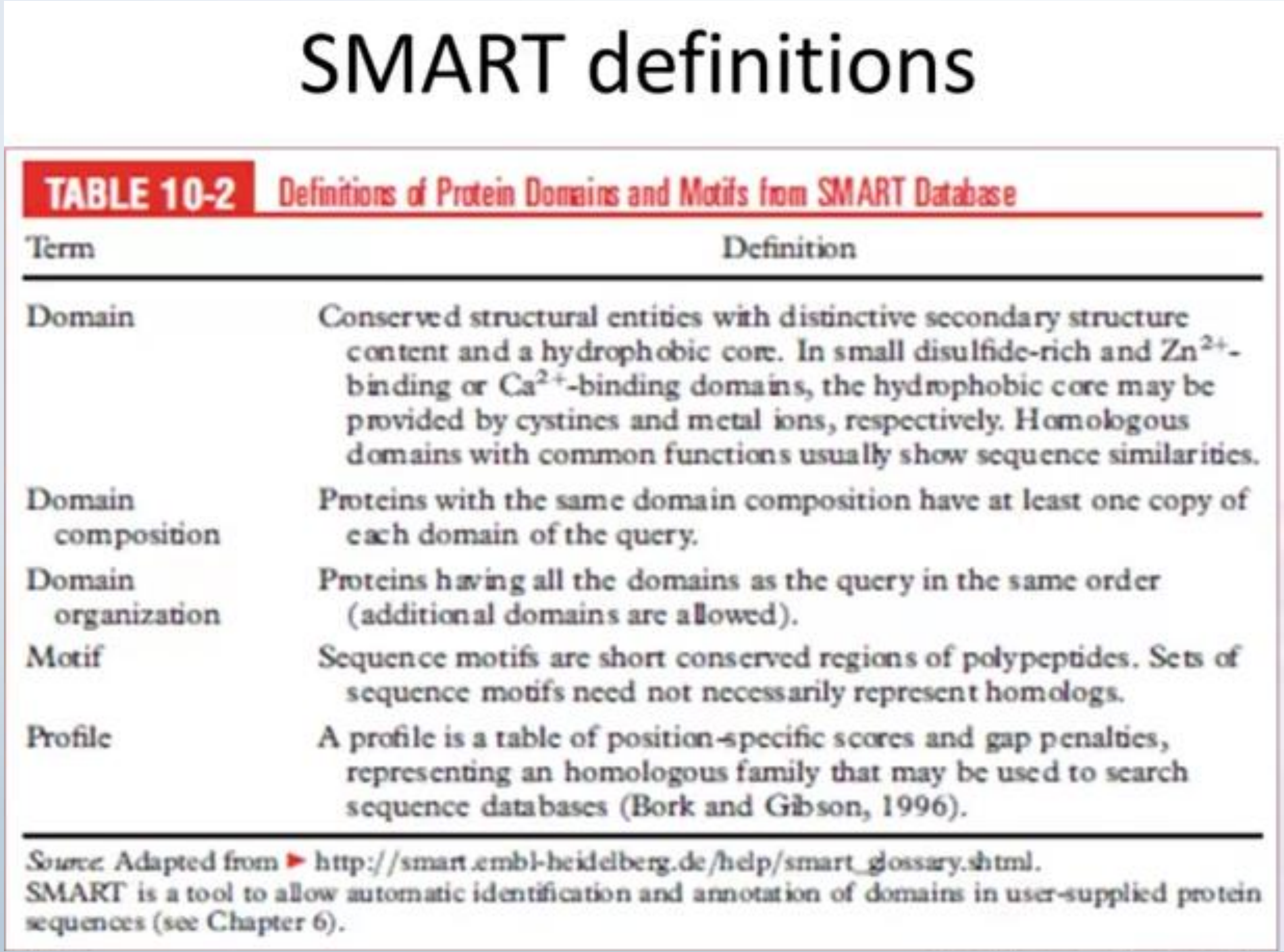
InterPro definitions

TABLE 10-1 Definitions from InterPro Database of Protein Families and Related Terms

Term	Definition
Family	An InterPro family is a group of evolutionarily related proteins that share one or more domains/repeats in common. An InterPro entry of "type = family" may contain a signature for a small conserved region that is representative of the family and therefore need not necessarily cover the whole protein.
Domain	A domain is defined as an independent structural unit which can be found alone or in conjunction with other domains or repeats. Domains are evolutionarily related. Even though the structure of a domain is not always known it is still possible to define the boundaries in many cases from sequence alone. Therefore, sequence criteria can be used to define domain boundaries.
Repeat	An InterPro repeat is a region that is not expected to fold into a globular domain on its own. For example, six to eight copies of the *WD40* repeat are needed to form a single globular domain. There also many other short repeat motifs that probably do not form a globular fold that have "type = repeat."
Posttranslational modification	A posttranslational modification includes, for example, an N-glycosylation site. The sequence motif is defined by the molecular recognition of this region in a cell. This may group together proteins that need not be evolutionarily related.

Source: Adapted from ► http://www.ebi.ac.uk/interpro/user_manual.html.

Different Definitions



SMART definitions

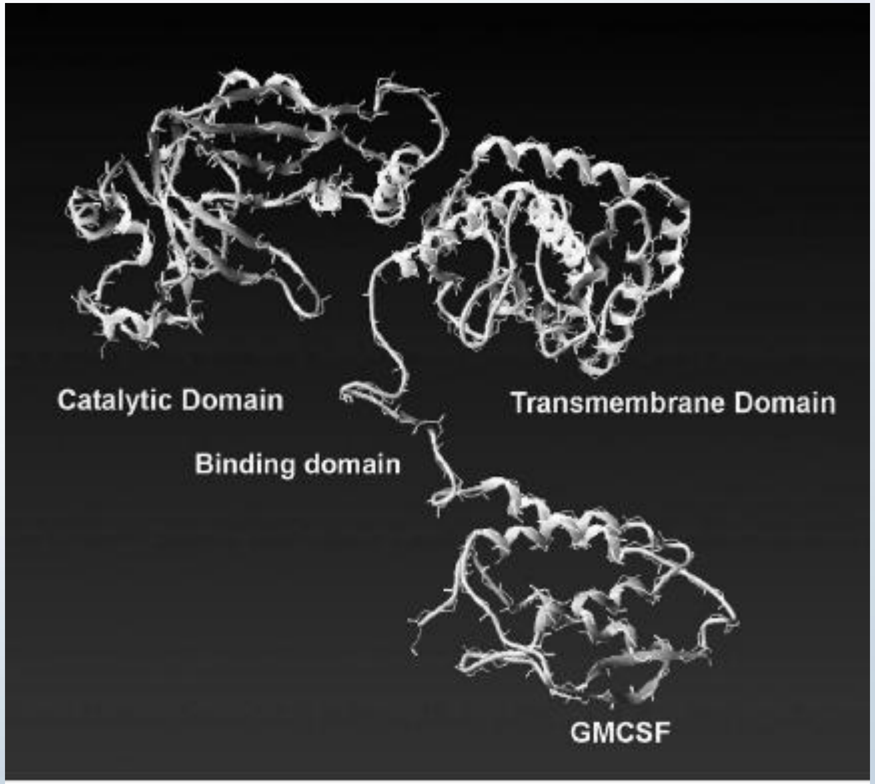
TABLE 10-2 Definitions of Protein Domains and Motifs from SMART Database

Term	Definition
Domain	Conserved structural entities with distinctive secondary structure content and a hydrophobic core. In small disulfide-rich and Zn^{2+}-binding or Ca^{2+}-binding domains, the hydrophobic core may be provided by cystines and metal ions, respectively. Homologous domains with common functions usually show sequence similarities.
Domain composition	Proteins with the same domain composition have at least one copy of each domain of the query.
Domain organization	Proteins having all the domains as the query in the same order (additional domains are allowed).
Motif	Sequence motifs are short conserved regions of polypeptides. Sets of sequence motifs need not necessarily represent homologs.
Profile	A profile is a table of position-specific scores and gap penalties, representing an homologous family that may be used to search sequence databases (Bork and Gibson, 1996).

Source: Adapted from ► http://smart.embl-heidelberg.de/help/smart_glossary.shtml.
SMART is a tool to allow automatic identification and annotation of domains in user-supplied protein sequences (see Chapter 6).

Protein Domains

- Proteins are usually made of domains
- A domain is an autonomous folding unit
- Domains are more than 50 amino acids long
- It’s common to find these together:
  - A regulatory domain
  - A binding domain
  - A catalytic domain



Crystal structure of DTGM.

Discovering Domains

- Researchers discover domains by
  - Comparing proteins that have similar functions
  - Aligning those proteins
  - Identifying conserved segments

- A domain is a multiple-sequence alignment formulated as a profile

- For each column, a domain indicates which amino acid is more likely to occur

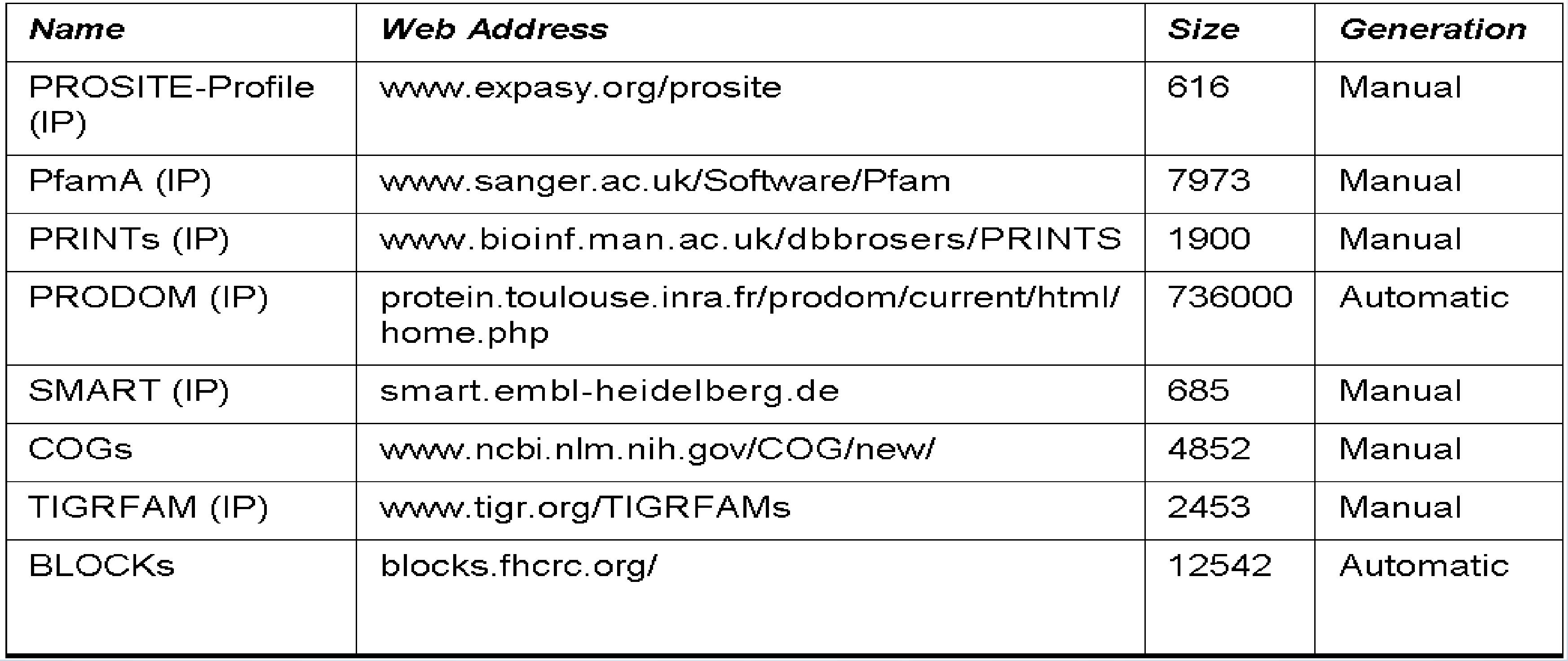
Domain Collections

- Scientists have been discovering and characterizing protein domains for many years

- 8 collections of domains have been established
  - Manual collections are very precise but small
  - Automatic collections are very extensive but less informative

- These collections
  - Overlap
  - Have been assembled by different scientists
  - Have different strengths and weaknesses

8 Domain Collections



Name	*Web Address*	*Size*	*Generation*
PROSITE-Profile (IP)	www.expasy.org/prosite	616	Manual
PfamA (IP)	www.sanger.ac.uk/Software/Pfam	7973	Manual
PRINTs (IP)	www.bioinf.man.ac.uk/dbbrosers/PRINTS	1900	Manual
PRODOM (IP)	protein.toulouse.inra.fr/prodom/current/html/home.php	736000	Automatic
SMART (IP)	smart.embl-heidelberg.de	685	Manual
COGs	www.ncbi.nlm.nih.gov/COG/new/	4852	Manual
TIGRFAM (IP)	www.tigr.org/TIGRFAMs	2453	Manual
BLOCKs	blocks.fhcrc.org/	12542	Automatic

- Pfam is the most extensive manual collection and is often used as a reference
- ***Note:*** Some addresses may not work. Please search via the internet for new ones.

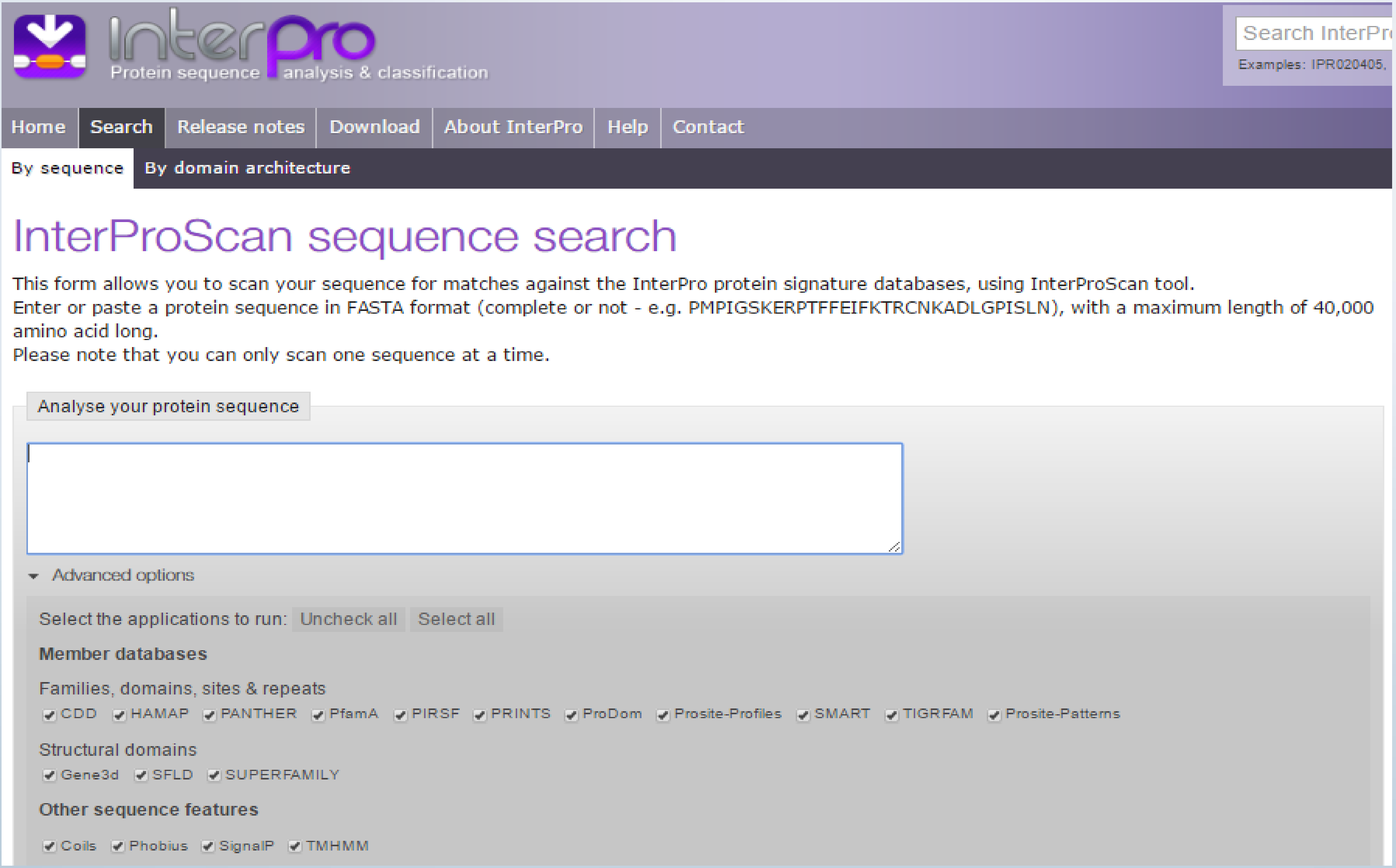
Searching Domain Collections

- Domains in Pfam often include known functions

- A match between your protein and a domain is desirable
  - A match is a potential indication of a function
  - This is **VERY** informative for further research!

- Three servers exist to compare proteins and domain collections:
  - InterProScan
  - CD-Search
  - Motif Scan

InterProScan



InterProScan sequence search

This form allows you to scan your sequence for matches against the InterPro protein signature databases, using InterProScan tool.
Enter or paste a protein sequence in FASTA format (complete or not - e.g. PMPIGSKERPTFFEIFKTRCNKADLGPISLN), with a maximum length of 40,000 amino acid long.
Please note that you can only scan one sequence at a time.

Analyse your protein sequence

Advanced options

Select the applications to run: Uncheck all Select all

Member databases

Families, domains, sites & repeats

☑ CDD ☑ HAMAP ☑ PANTHER ☑ PfamA ☑ PIRSF ☑ PRINTS ☑ ProDom ☑ Prosite-Profiles ☑ SMART ☑ TIGRFAM ☑ Prosite-Patterns

Structural domains

☑ Gene3d ☑ SFLD ☑ SUPERFAMILY

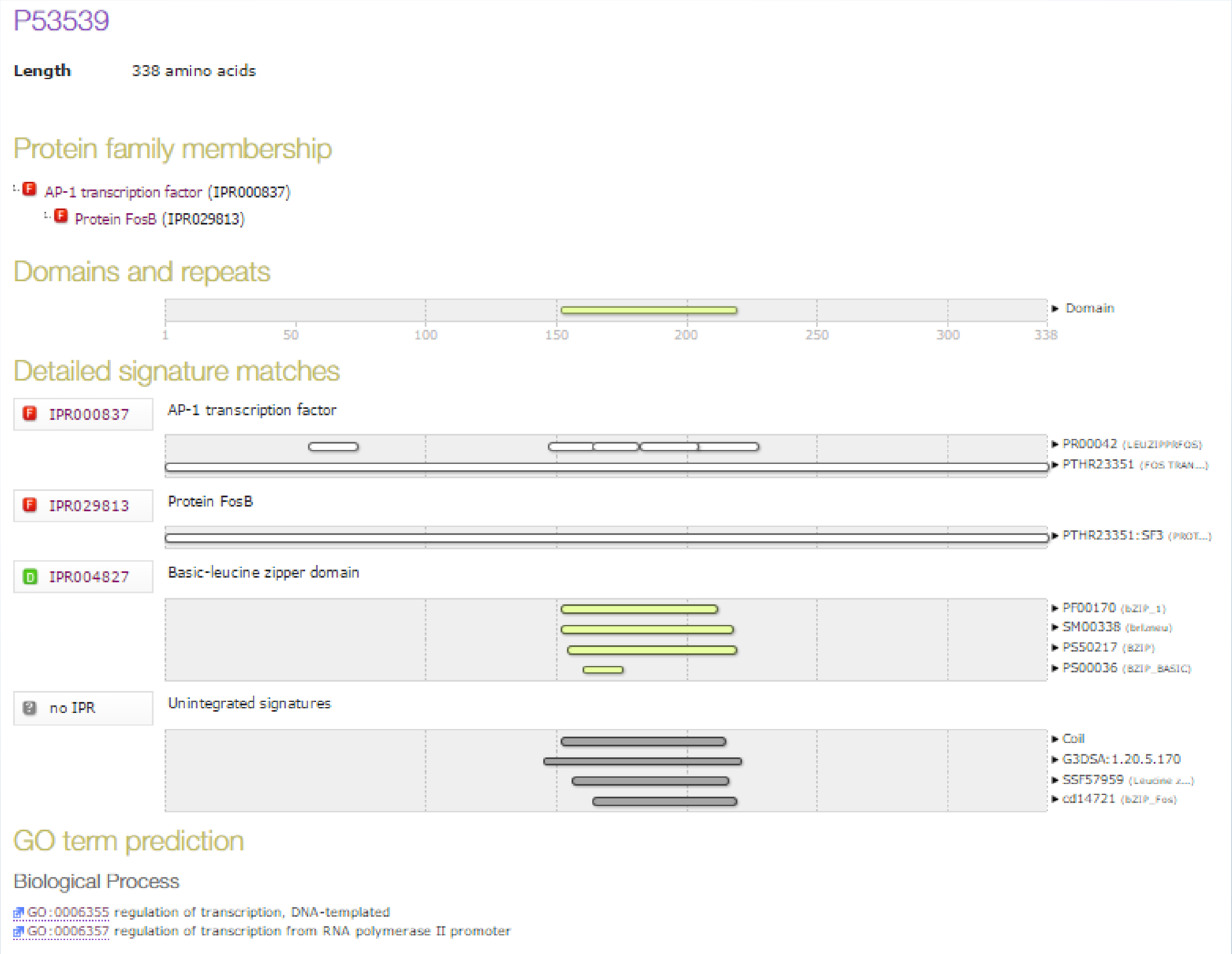
Other sequence features

☑ Coils ☑ Phobius ☑ SignalP ☑ TMHMM

https://www.ebi.ac.uk/interpro/search/sequence-search

InterProScan – Example (P53539)

- InterProScan is the most comprehensive search engine for domain databases
- Makes it possible to compare alternative results on most collections
- Does not provide a statistical score



P53539

Length 338 amino acids

Protein family membership

- AP-1 transcription factor (IPR000837)
  - Protein FosB (IPR029813)

Domains and repeats

Detailed signature matches

IPR000837 AP-1 transcription factor

IPR029813 Protein FosB

IPR004827 Basic-leucine zipper domain

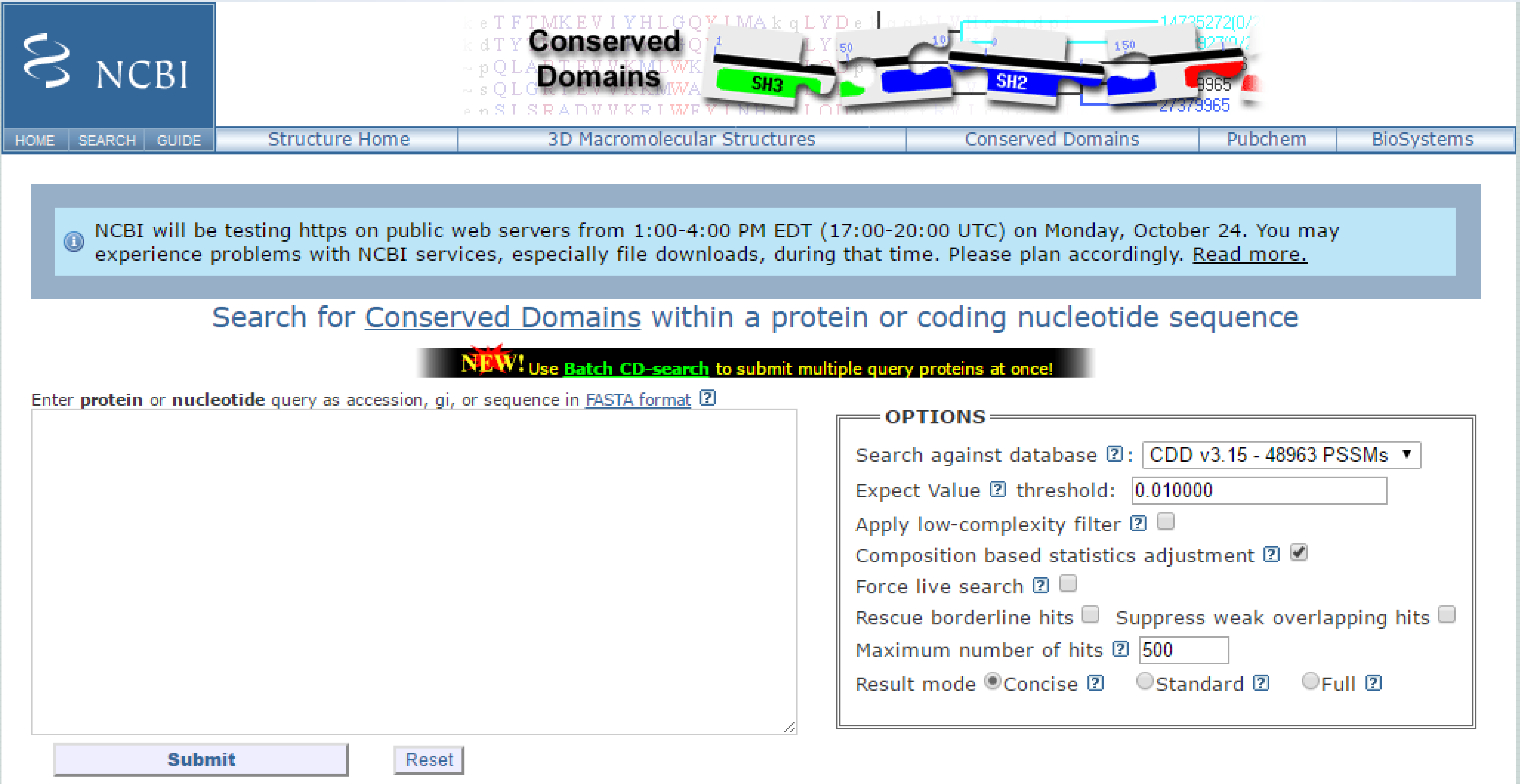
no IPR Unintegrated signatures

GO term prediction

Biological Process

GO:0006355 regulation of transcription, DNA-templated
GO:0006357 regulation of transcription from RNA polymerase II promoter

CD-Search



NCBI will be testing https on public web servers from 1:00-4:00 PM EDT (17:00-20:00 UTC) on Monday, October 24. You may experience problems with NCBI services, especially file downloads, during that time. Please plan accordingly. Read more.

Search for Conserved Domains within a protein or coding nucleotide sequence

NEW! Use Batch CD-search to submit multiple query proteins at once!

Enter **protein** or **nucleotide** query as accession, gi, or sequence in FASTA format

OPTIONS

Search against database: CDD v3.15 - 48963 PSSMs

Expect Value threshold: 0.010000

Apply low-complexity filter □

Composition based statistics adjustment ☑

Force live search □

Rescue borderline hits □ Suppress weak overlapping hits □

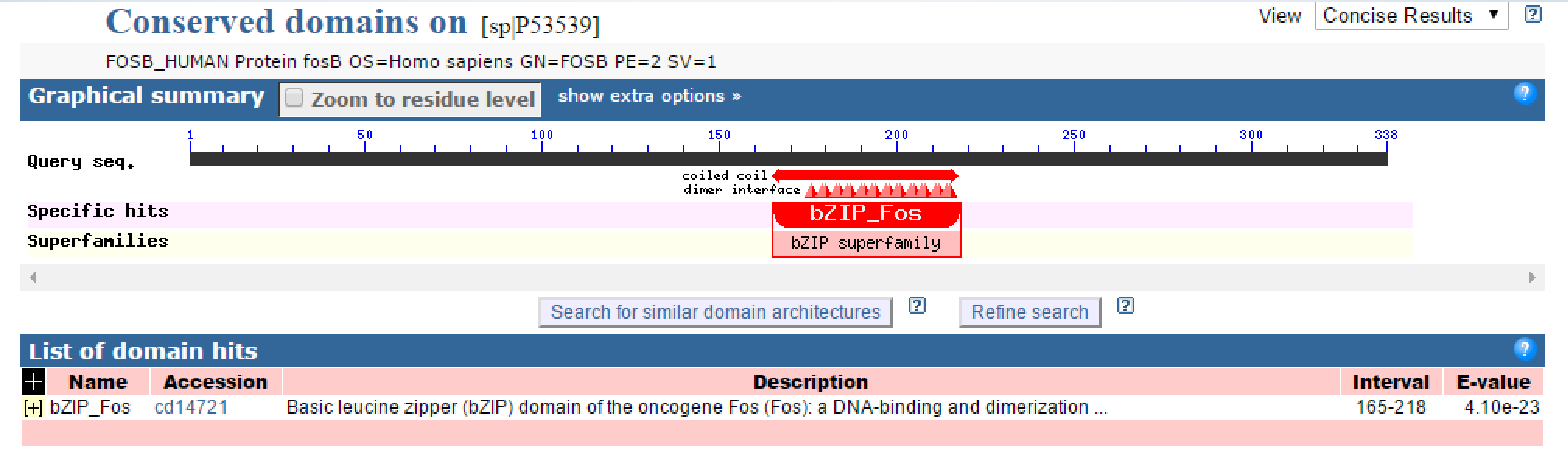
Maximum number of hits 500

Result mode ● Concise ○ Standard ○ Full

https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

CD-Search - Example (P53539)

- CD search is less extensive than that of InterProScan
- Results come with a statistical evaluation (E-value)
  - 10^{e-23} **Low E-value** Good match



Conserved domains on [sp|P53539]

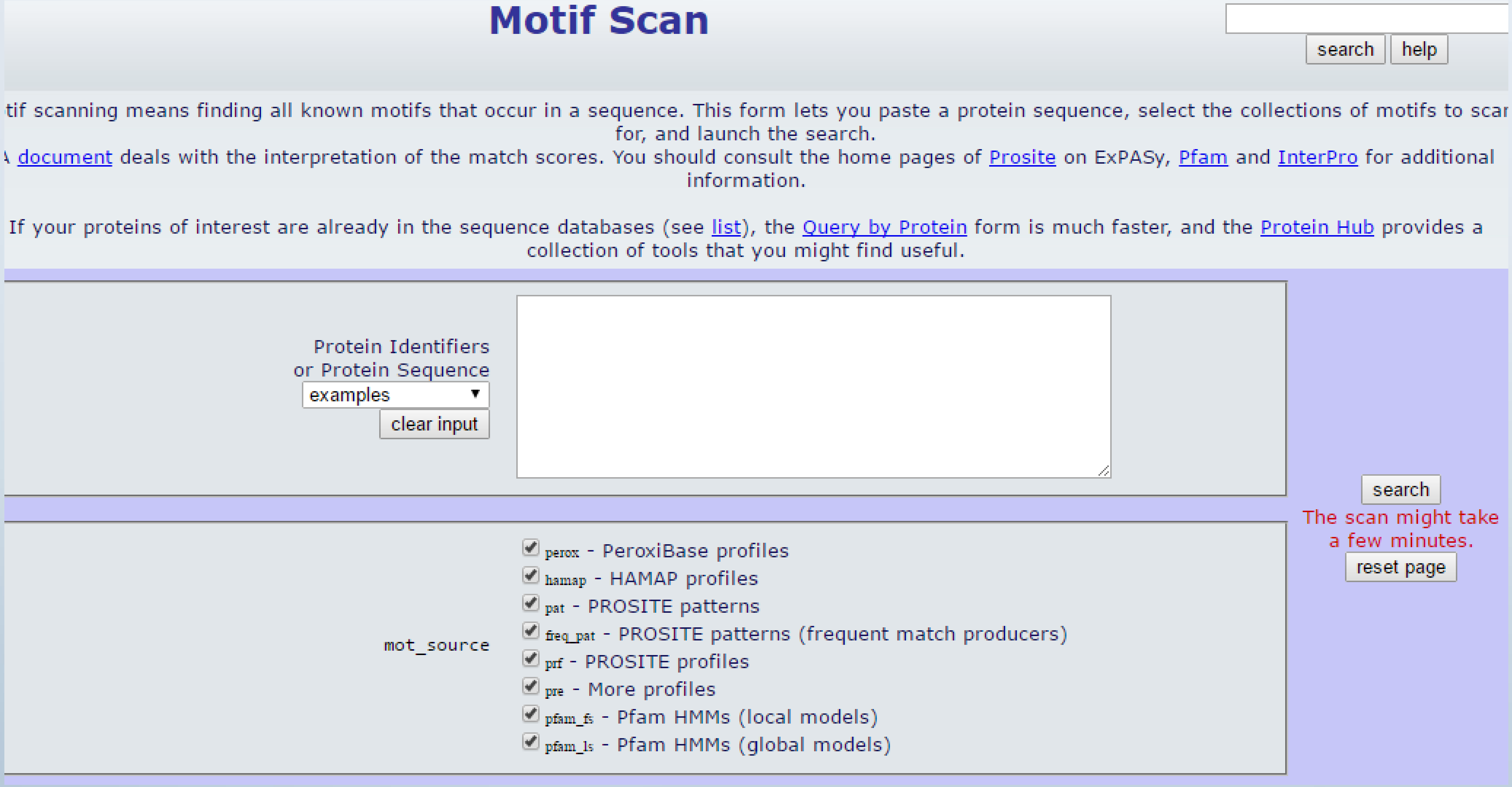
FOSB_HUMAN Protein fosB OS=Homo sapiens GN=FOSB PE=2 SV=1

View Concise Results ▼

List of domain hits

+	Name	Accession	Description	Interval	E-value
[+]	bZIP_Fos	cd14721	Basic leucine zipper (bZIP) domain of the oncogene Fos (Fos): a DNA-binding and dimerization ...	165-218	4.10e-23

Motif Scan



Motif Scan

tif scanning means finding all known motifs that occur in a sequence. This form lets you paste a protein sequence, select the collections of motifs to scan for, and launch the search.

A document deals with the interpretation of the match scores. You should consult the home pages of Prosite on ExPASy, Pfam and InterPro for additional information.

If your proteins of interest are already in the sequence databases (see list), the Query by Protein form is much faster, and the Protein Hub provides a collection of tools that you might find useful.

Protein Identifiers or Protein Sequence

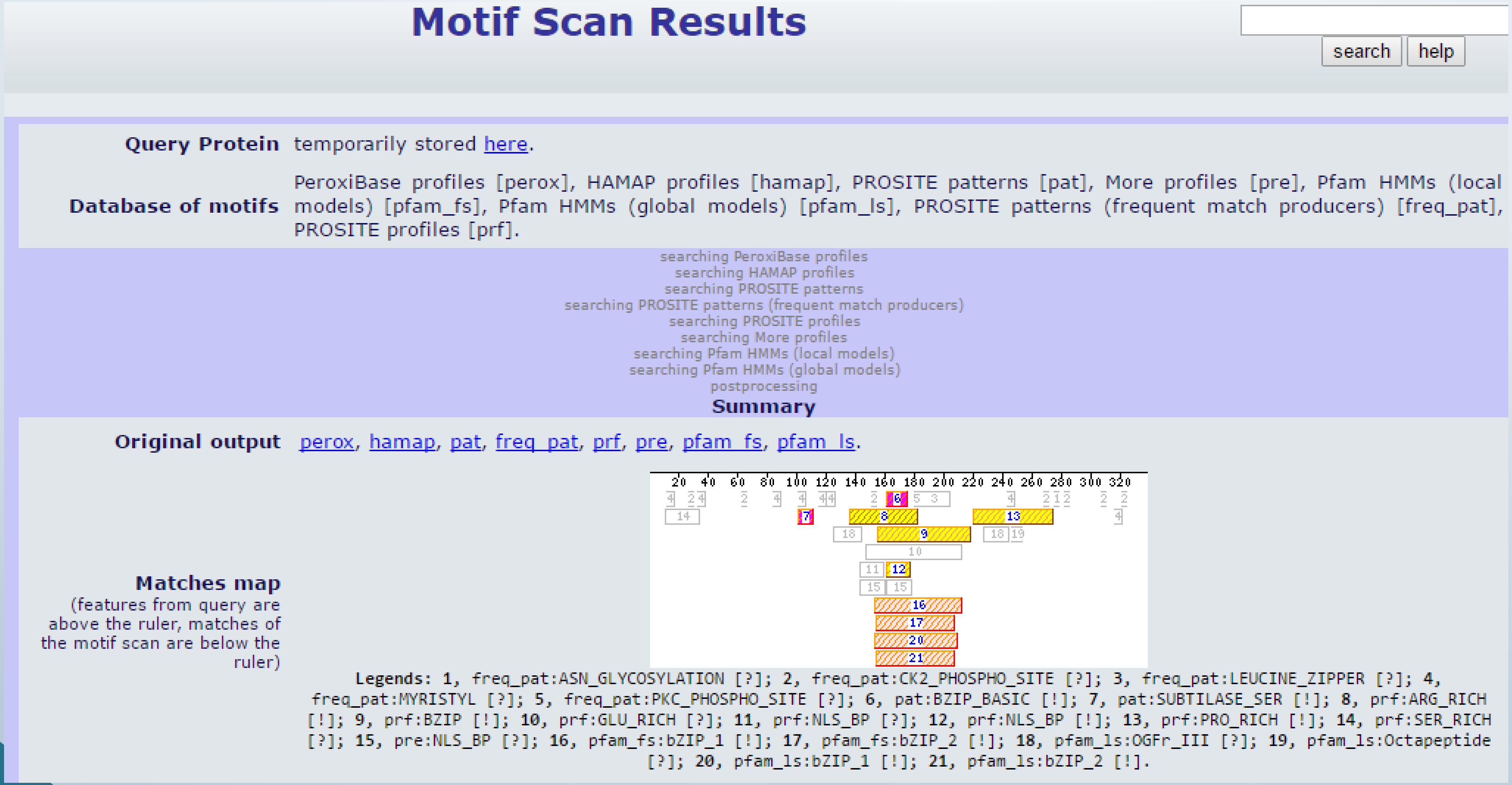
The scan might take a few minutes.

mot_source

- ☑ perox - PeroxiBase profiles
- ☑ hamap - HAMAP profiles
- ☑ pat - PROSITE patterns
- ☑ freq_pat - PROSITE patterns (frequent match producers)
- ☑ prf - PROSITE profiles
- ☑ pre - More profiles
- ☑ pfam_fs - Pfam HMMs (local models)
- ☑ pfam_ls - Pfam HMMs (global models)

http://myhits.isb-sib.ch/cgi-bin/motif_scan

Motif Scan - Example (P53539)



Motif Scan Results

Query Protein temporarily stored here.

Database of motifs PeroxiBase profiles [perox], HAMAP profiles [hamap], PROSITE patterns [pat], More profiles [pre], Pfam HMMs (local models) [pfam_fs], Pfam HMMs (global models) [pfam_ls], PROSITE patterns (frequent match producers) [freq_pat], PROSITE profiles [prf].

searching PeroxiBase profiles
searching HAMAP profiles
searching PROSITE patterns
searching PROSITE patterns (frequent match producers)
searching PROSITE profiles
searching More profiles
searching Pfam HMMs (local models)
searching Pfam HMMs (global models)
postprocessing

Summary

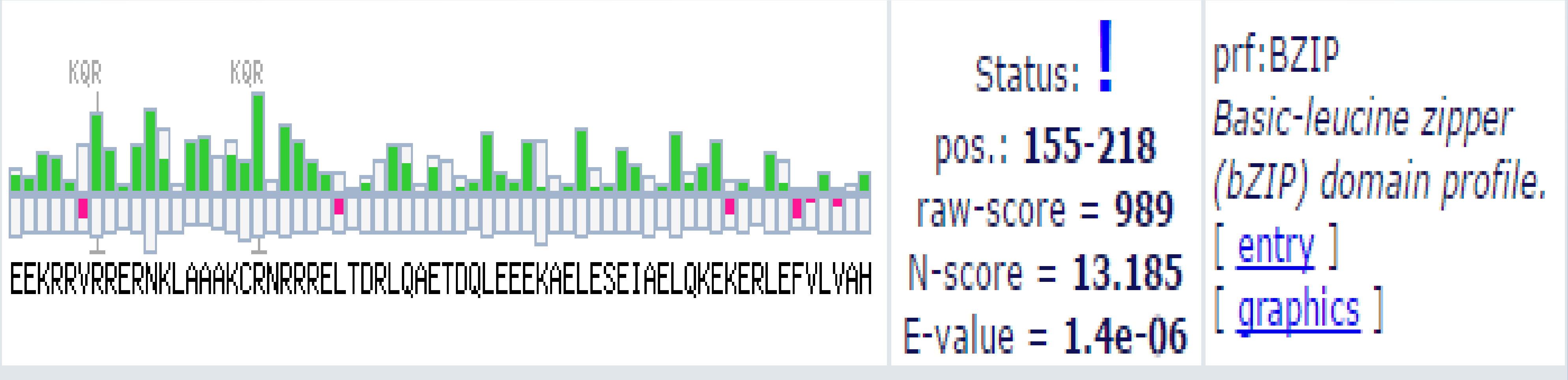
Original output perox, hamap, pat, freq_pat, prf, pre, pfam_fs, pfam_ls.

Matches map (features from query are above the ruler, matches of the motif scan are below the ruler)

Legends: 1, freq_pat:ASN_GLYCOSYLATION [?]; 2, freq_pat:CK2_PHOSPHO_SITE [?]; 3, freq_pat:LEUCINE_ZIPPER [?]; 4, freq_pat:MYRISTYL [?]; 5, freq_pat:PKC_PHOSPHO_SITE [?]; 6, pat:BZIP_BASIC [!]; 7, pat:SUBTILASE_SER [!]; 8, prf:ARG_RICH [!]; 9, prf:BZIP [!]; 10, prf:GLU_RICH [?]; 11, prf:NLS_BP [?]; 12, prf:NLS_BP [!]; 13, prf:PRO_RICH [!]; 14, prf:SER_RICH [?]; 15, pre:NLS_BP [?]; 16, pfam_fs:bZIP_1 [!]; 17, pfam_fs:bZIP_2 [!]; 18, pfam_ls:OGFr_III [?]; 19, pfam_ls:Octapeptide [?]; 20, pfam_ls:bZIP_1 [!]; 21, pfam_ls:bZIP_2 [!].

Looking into the Details

- Catalytic residues are normally highly conserved in domains
- Motif Scan makes it possible to check whether these important residues are conserved in your sequence
  - **High bar above 0** = Highly conserved residues
  - Green = Your sequence has an expected residue
  - Red = Your sequence has an unexpected residue


KQR
KQR
EEKRRVRRERNKLAAAKCRNRRRELTDRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAH

Status: !

pos.: 155-218

raw-score = 989

N-score = 13.185

E-value = 1.4e-06

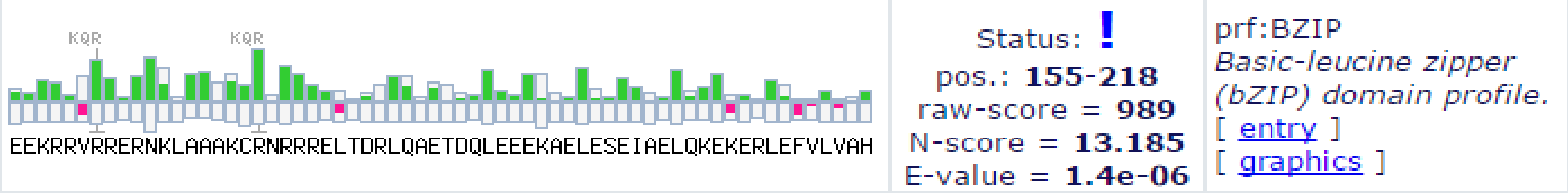
prf:BZIP

Basic-leucine zipper (bZIP) domain profile.

[entry]

[graphics]

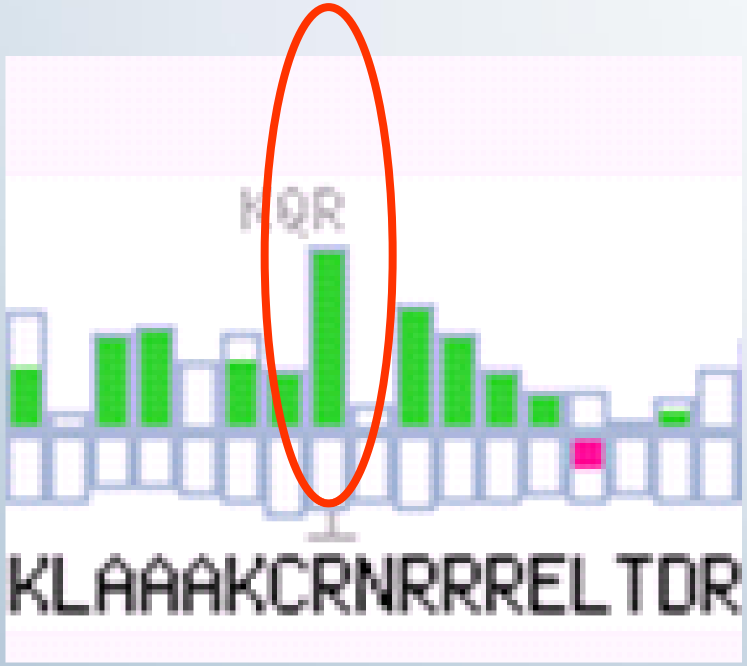
Looking into the Details



EEKRRVRRERNKLAAAKCRNRRRELTDRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAH

Status: **!**
pos.: **155-218**
raw-score = **989**
N-score = **13.185**
E-value = **1.4e-06**

prf:BZIP
Basic-leucine zipper (bZIP) domain profile.
[entry]
[graphics]



- ✓ R (Arginine) is highly expected at this position
  - High bar
  - Potential active site
- ✓ If your protein has an arginine on this position . . .
  - Bar is filled with green
  - Your protein could be active

Predicting Post-translational Modifications

- Post-translational modifications often occur on similar motifs in different proteins
- PROSITE is a database containing a list of known motifs, each associated with a function or a post-translational modification
- You can search PROSITE by looking for each motif it contains in your protein
- PROSITE entries come with an extensive documentation on each function of the motif

Predicting Functions with Domains

- Finding a match with a domain having a catalytic function is good news . . . but what, exactly, does it mean?

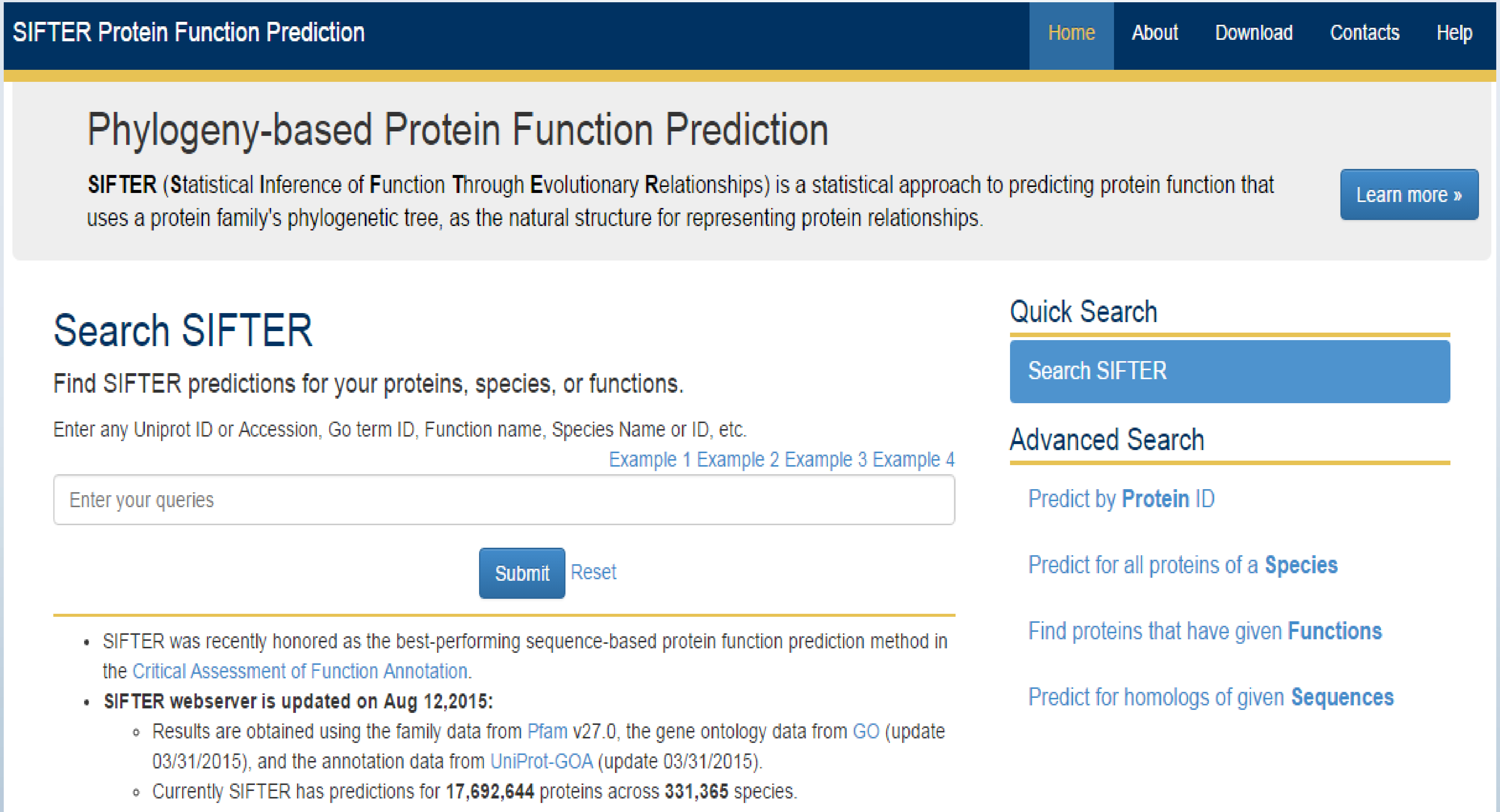
- A match indicates that your sequence has the domain structure . . . but does it also have the function?

- You cannot say before looking into these details:
  - Where are the catalytic residues on the domain?
  - Does your sequence have the right residues at these positions?

Function Prediction Methods

- Homology-based methods
- Sequence motif/domain-based methods
- Structure-based methods
- Genomic context-based methods
  - Gene fusion
  - Co-location/co-expression
- Computational Solvent Mapping
- Network-based methods

SIFTER – Example (P0C871)



Phylogeny-based Protein Function Prediction

SIFTER (**S**tatistical **I**nference of **F**unction **T**hrough **E**volutionary **R**elationships) is a statistical approach to predicting protein function that uses a protein family's phylogenetic tree, as the natural structure for representing protein relationships.

Search SIFTER

Find SIFTER predictions for your proteins, species, or functions.

Enter any Uniprot ID or Accession, Go term ID, Function name, Species Name or ID, etc.

- SIFTER was recently honored as the best-performing sequence-based protein function prediction method in the Critical Assessment of Function Annotation.
- **SIFTER webserver is updated on Aug 12,2015:**
  - Results are obtained using the family data from Pfam v27.0, the gene ontology data from GO (update 03/31/2015), and the annotation data from UniProt-GOA (update 03/31/2015).
  - Currently SIFTER has predictions for **17,692,644** proteins across **331,365** species.

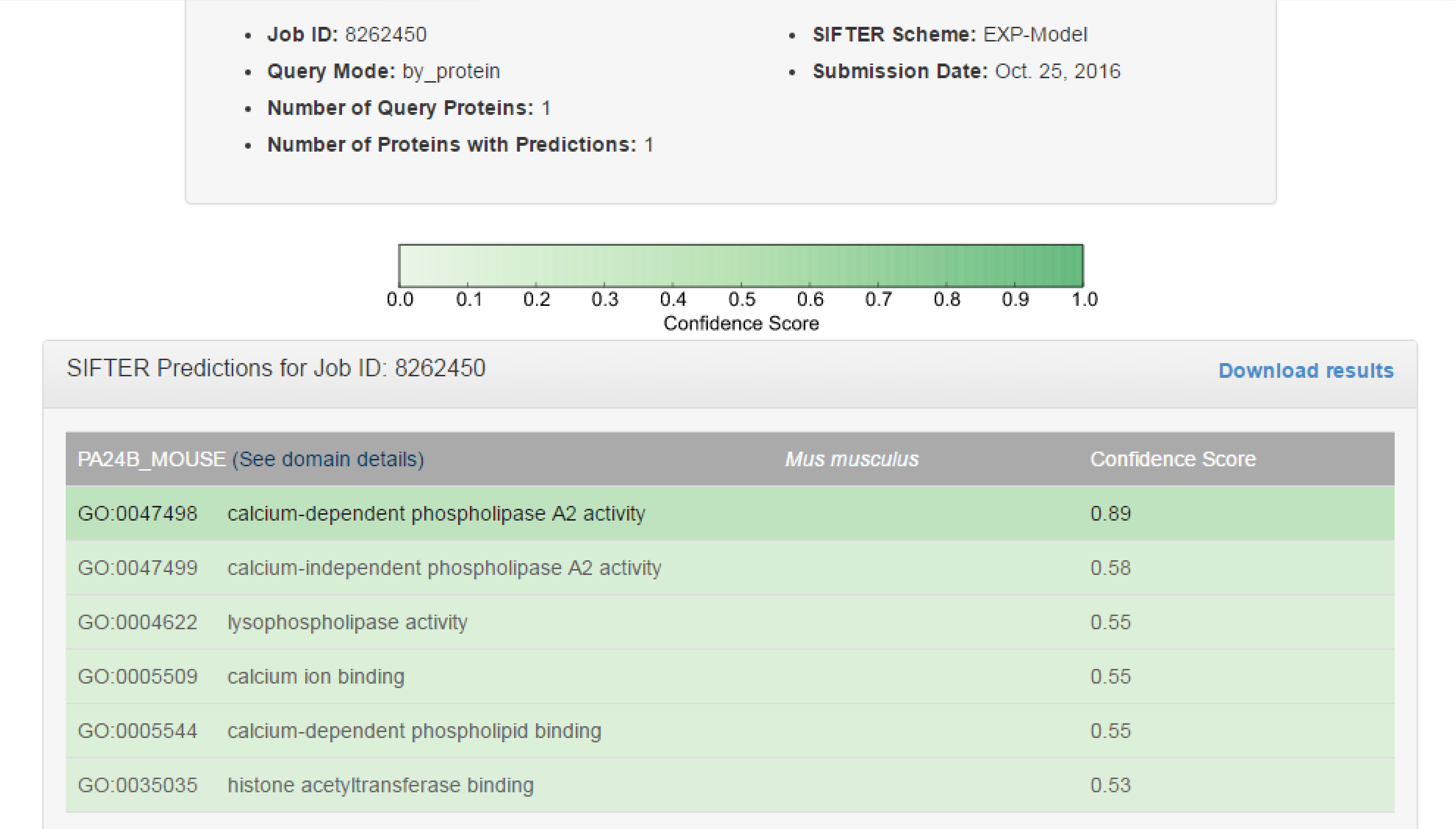
Quick Search

Advanced Search

Predict by **Protein** ID
Predict for all proteins of a **Species**
Find proteins that have given **Functions**
Predict for homologs of given **Sequences**

http://sifter.berkeley.edu/

Example (PA24B_MOUSE - P0C871)



- **Job ID:** 8262450
- **Query Mode:** by_protein
- **Number of Query Proteins:** 1
- **Number of Proteins with Predictions:** 1
- **SIFTER Scheme:** EXP-Model
- **Submission Date:** Oct. 25, 2016

SIFTER Predictions for Job ID: 8262450

PA24B_MOUSE (See domain details)		*Mus musculus*	Confidence Score
GO:0047498	calcium-dependent phospholipase A2 activity		0.89
GO:0047499	calcium-independent phospholipase A2 activity		0.58
GO:0004622	lysophospholipase activity		0.55
GO:0005509	calcium ion binding		0.55
GO:0005544	calcium-dependent phospholipid binding		0.55
GO:0035035	histone acetyltransferase binding		0.53

Protein Function Prediction – General Protocol

1. Similarity search: First start with Blastp, if your sequence is less than 40% identity go for PSI-Blast

2. Domain search: Do domain search using Interproscan, Pfam or CDART

3. Search for signal peptide and transmembrane (TM): search for signal peptide using signalp and TM using TMHMM, phobius

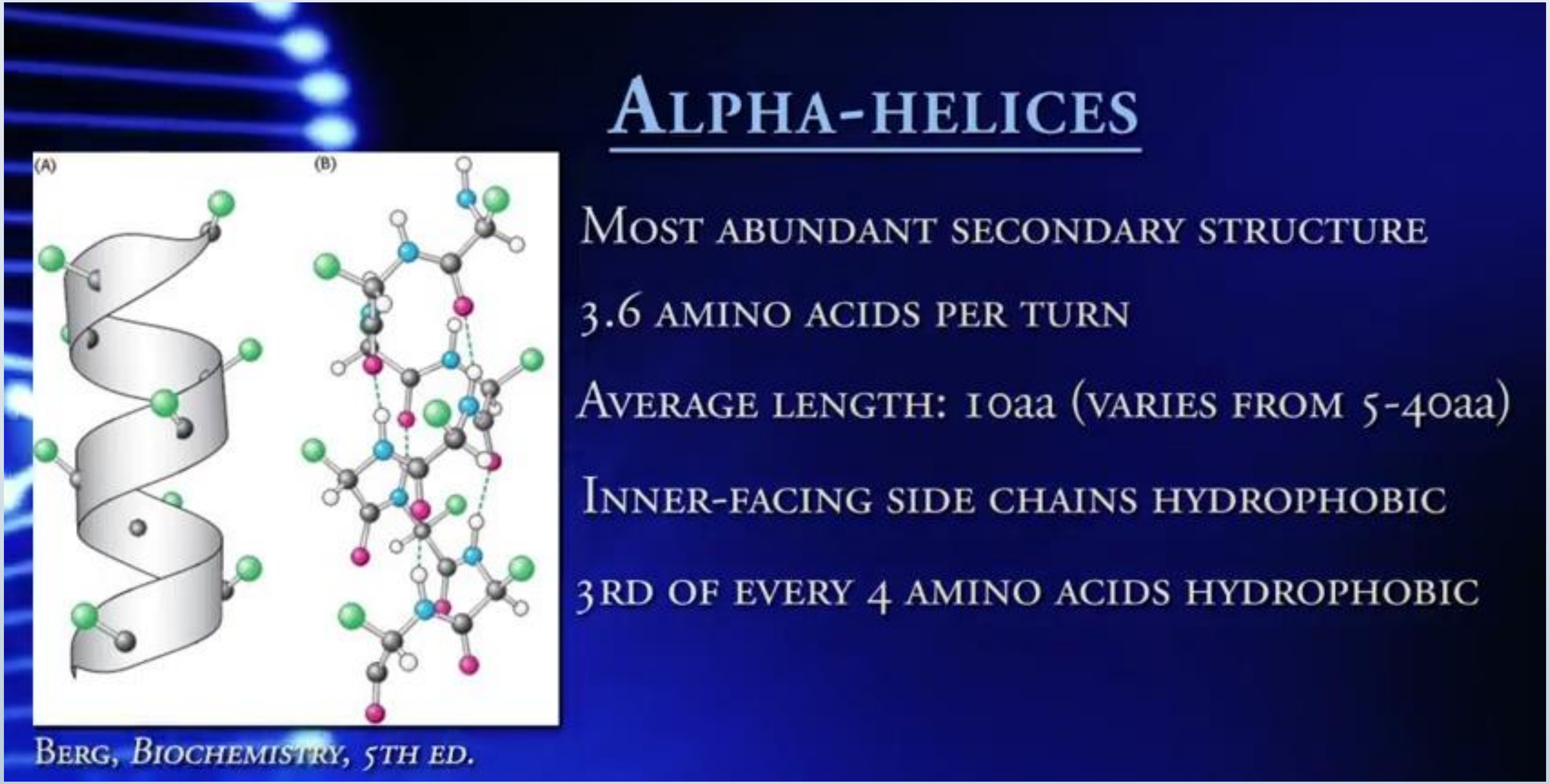
4. Comparative modelling: Do homology modelling using swiss model, if your sequence less than 40% identity from blast result go for ab-intio modelling using I-Tasser

5. Gene ontology classification: You can search sequence for GO classification using blast2go or STRAP

6. Functional association prediction: Try searching sequence using STRING search

For further reading: Predicting protein function from sequence and structure, D. Lee, O. Redfern & C. Orengo, *Nature Reviews Molecular Cell Biology*, 8, 995-1005, 2007, doi:10.1038/nrm2281

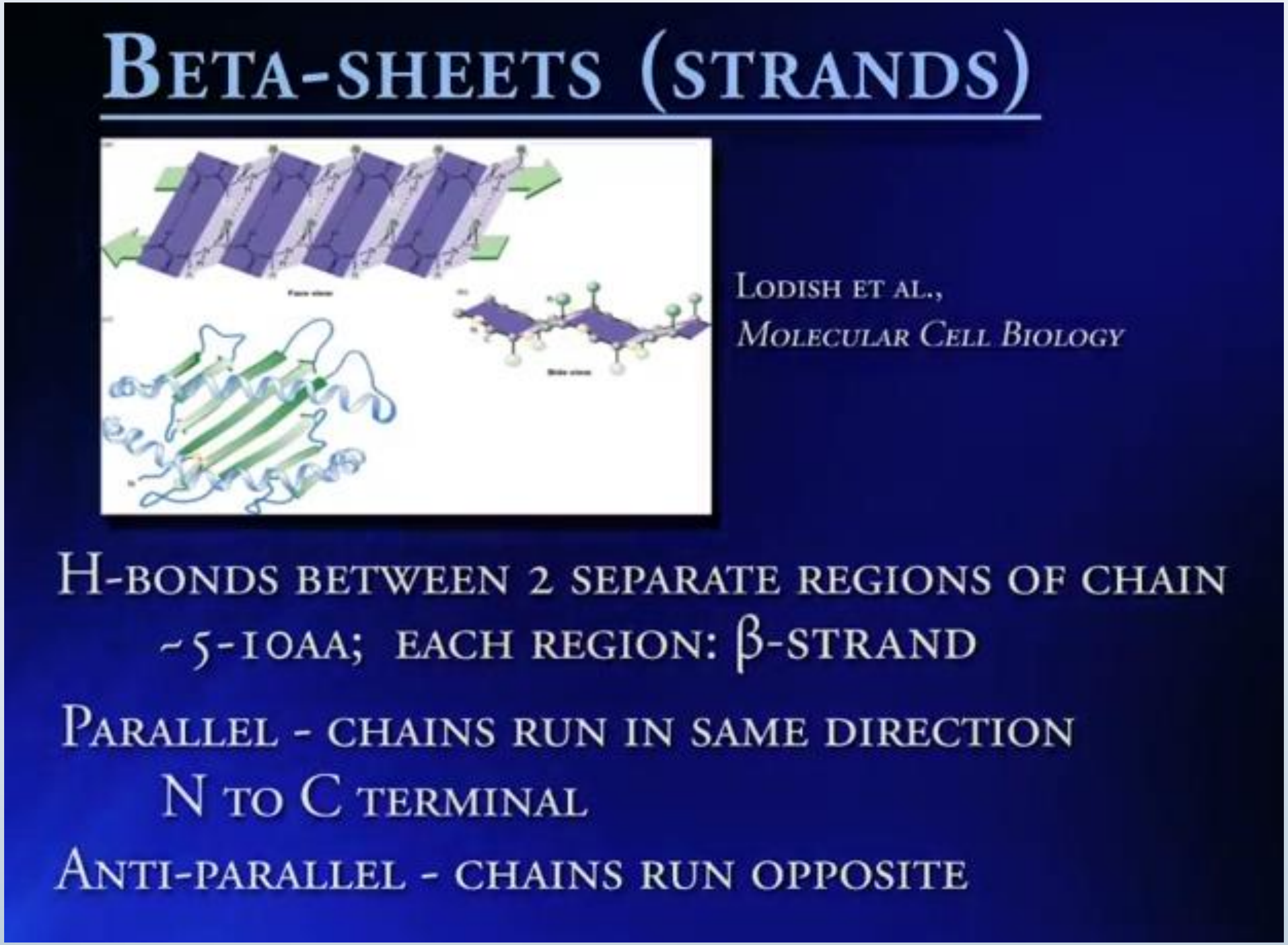
Secondary Structure Prediction



Alpha-helices

- Most abundant secondary structure
- 3.6 amino acids per turn
- Average length: 10aa (varies from 5-40aa)
- Inner-facing side chains hydrophobic
- 3rd of every 4 amino acids hydrophobic

Berg, *Biochemistry*, 5th ed.



Beta-sheets (strands)

Lodish et al.,
Molecular Cell Biology

H-bonds between 2 separate regions of chain
~5-10aa; each region: β-strand

Parallel - chains run in same direction
N to C terminal

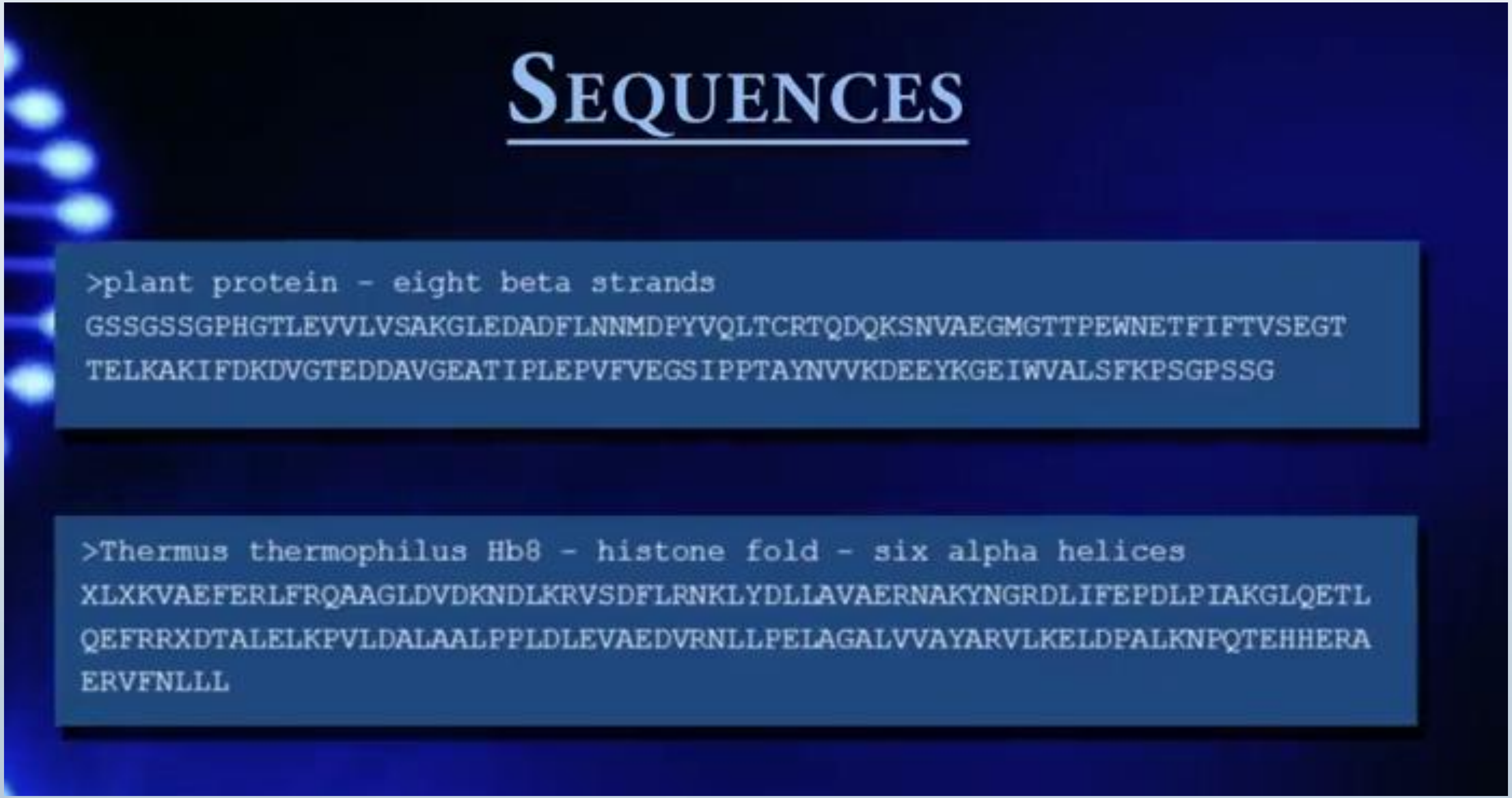
Anti-parallel - chains run opposite

Secondary Structure Prediction Method

- **HNN** (Hierarchical Neural Network)
  https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_hnn.html

- **PHD** (Profile network from Heidelburg)
  https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_phd.html

Example

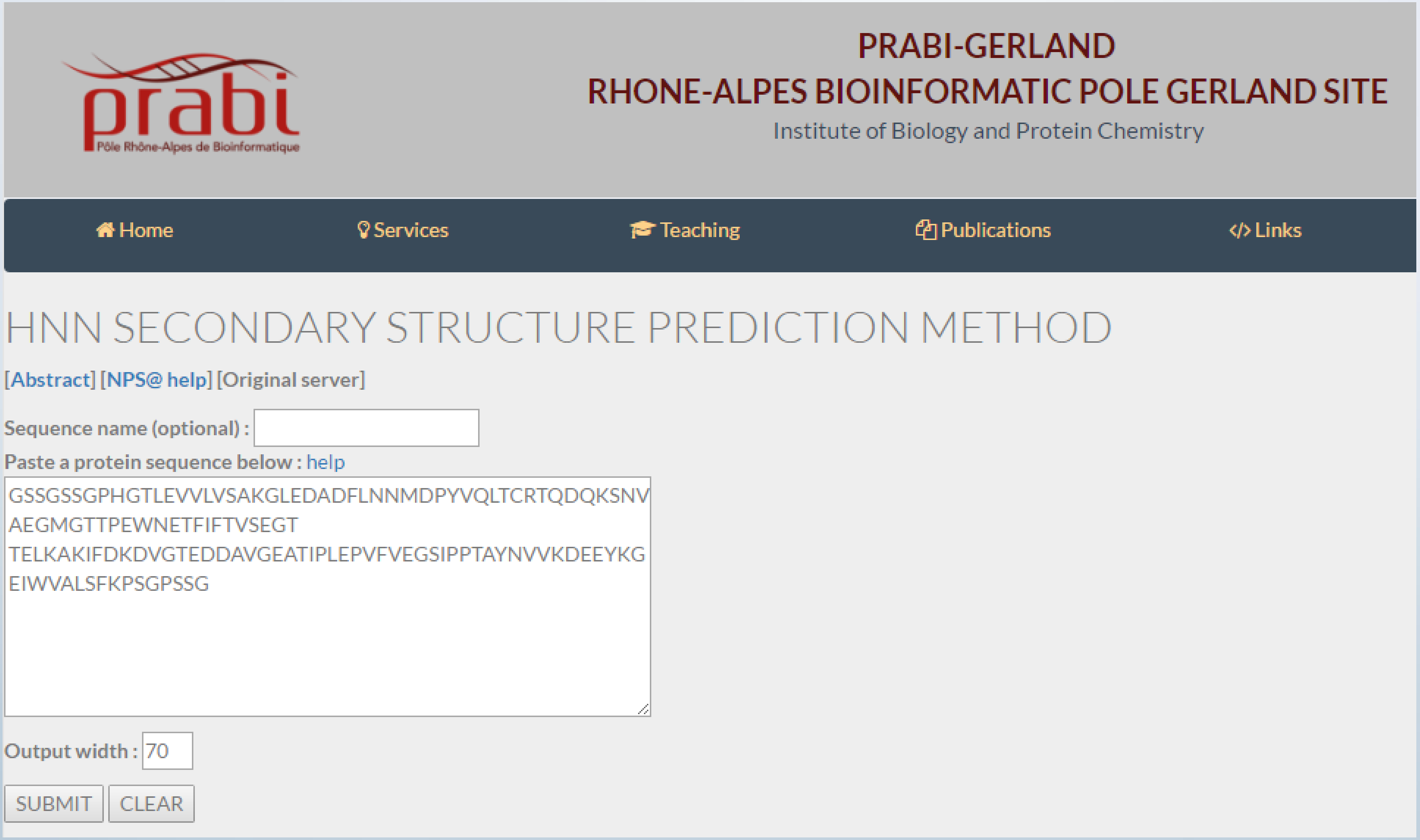


SEQUENCES

```
>plant protein - eight beta strands
GSSGSSGPHGTLEVVLVSAKGLEDADFLNNMDPYVQLTCRTQDQKSNVAEGMGTTPEWNETFIFTVSEGT
TELKAKIFDKDVGTEDDAVGEATIPLEPVFVEGSIPPTAYNVVKDEEYKGEIWVALSFKPSGPSSG
```

```
>Thermus thermophilus Hb8 - histone fold - six alpha helices
XLXKVAEFERLFRQAAGLDVDKNDLKRVSDFLRNKLYDLLAVAERNAKYNGRDLIFEPDLPIAKGLQETL
QEFRRXDTALELKPVLDALAALPPLDLEVAEDVRNLLPELAGALVVAYARVLKELDPALKNPQTEHHERA
ERVFNLLL
```

HNN Algorithm



PRABI-GERLAND

RHONE-ALPES BIOINFORMATIC POLE GERLAND SITE

Institute of Biology and Protein Chemistry

HNN SECONDARY STRUCTURE PREDICTION METHOD

[Abstract] [NPS@ help] [Original server]

Sequence name (optional) :

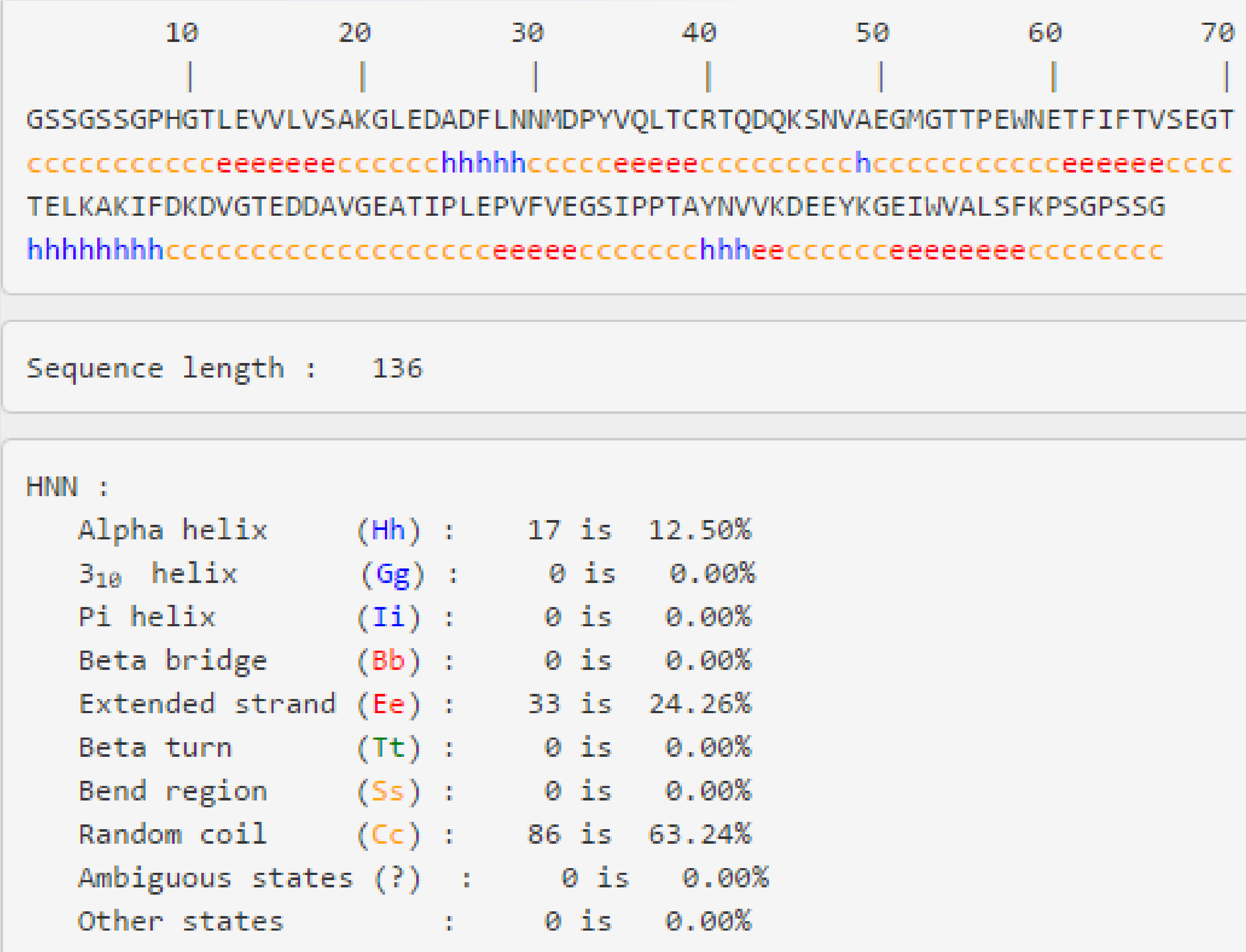
Paste a protein sequence below : help

```
GSSGSSGPHGTLEVVLVSAKGLEDADFLNNMDPYVQLTCRTQDQKSNV
AEGMGTTPEWNETFIFTVSEGT
TELKAKIFDKDVGTEDDAVGEATIPLEPVFVEGSIPPTAYNVVKDEEYKG
EIWVALSFKPSGPSSG
```

Output width : 70

https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_hnn.html

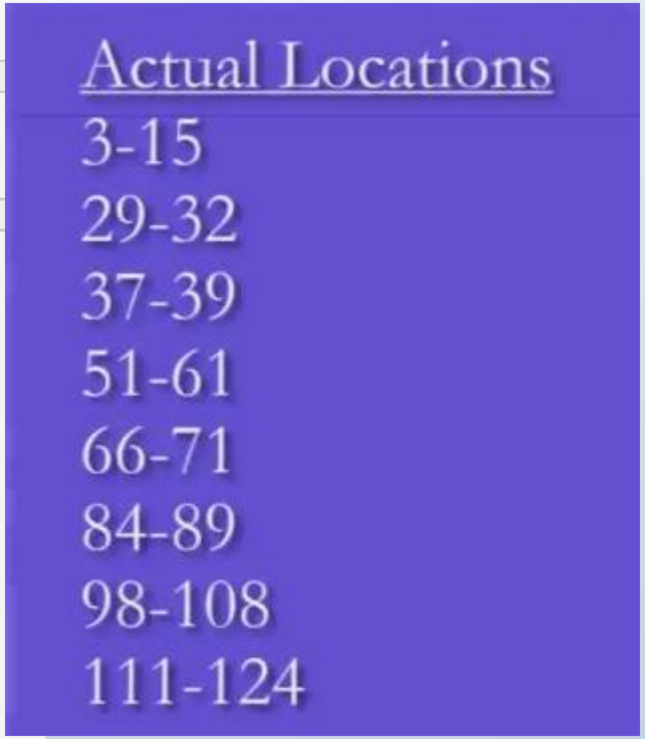
HNN Algorithm



```
         10        20        30        40        50        60        70
          |         |         |         |         |         |         |
GSSGSSGPHGTLEVVLVSAKGLEDADFLNNMDPYVQLTCRTQDQKSNVAEGMGTTPEWNETFIFTVSEGT
ccccccccccceeeeeeecccccchhhhhccccceeeeecccccccccchccccccccccceeeeeecccc
TELKAKIFDKDVGTEDDAVGEATIPLEPVFVEGSIPPTAYNVVKDEEYKGEIWVALSFKPSGPSSG
hhhhhhhhccccccccccccccccccceeeeeccccccchhheecccccceeeeeeeecccccccc
```

```
Sequence length :    136
```

```
HNN :
   Alpha helix     (Hh) :    17 is  12.50%
   $3_{10}$  helix       (Gg) :     0 is   0.00%
   Pi helix        (Ii) :     0 is   0.00%
   Beta bridge     (Bb) :     0 is   0.00%
   Extended strand (Ee) :    33 is  24.26%
   Beta turn       (Tt) :     0 is   0.00%
   Bend region     (Ss) :     0 is   0.00%
   Random coil     (Cc) :    86 is  63.24%
   Ambiguous states (?)  :     0 is   0.00%
   Other states         :     0 is   0.00%
```



Actual Locations

3-15

29-32

37-39

51-61

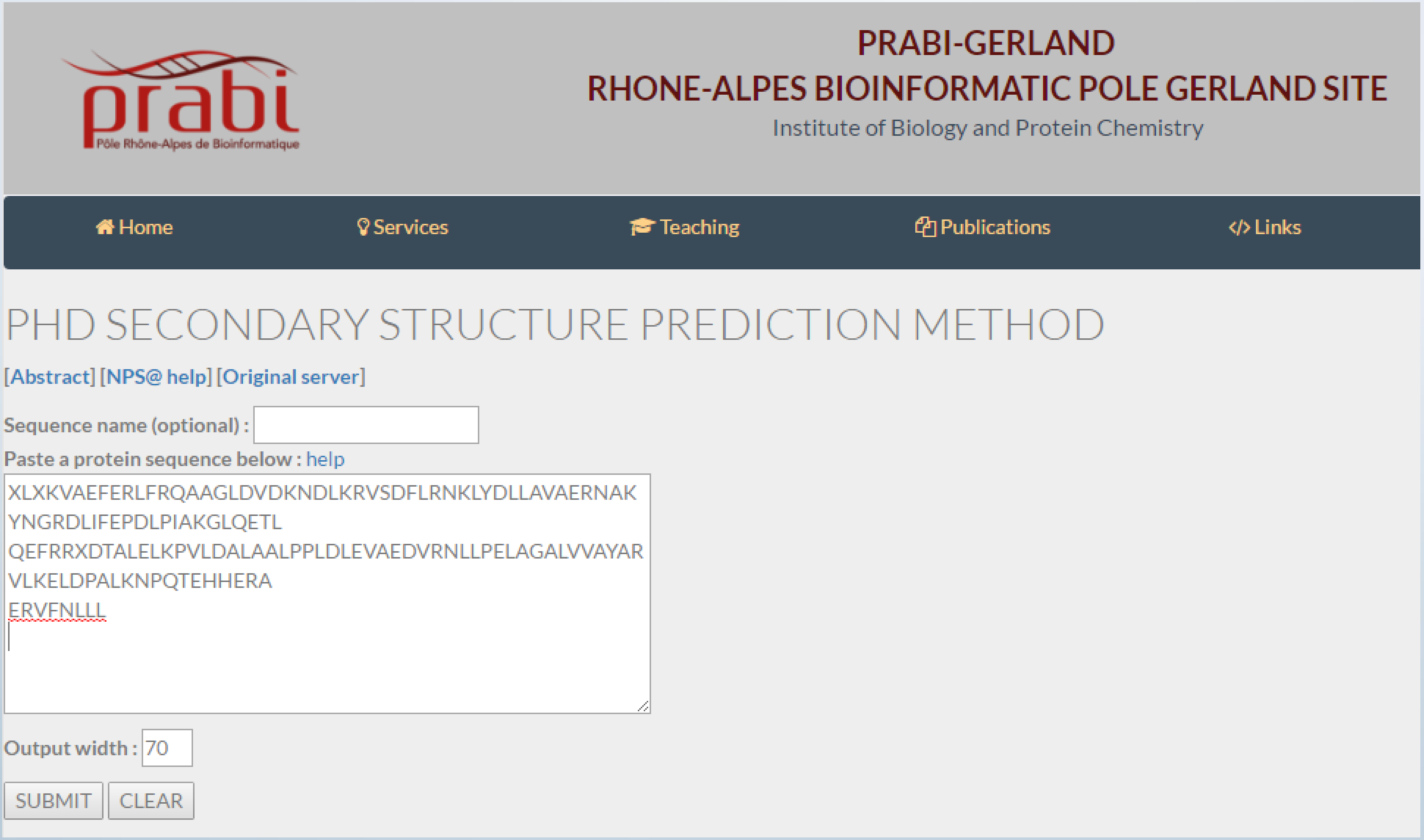
66-71

84-89

98-108

111-124

PHD Algorithm


prabi
Pôle Rhône-Alpes de Bioinformatique
PRABI-GERLAND
RHONE-ALPES BIOINFORMATIC POLE GERLAND SITE
Institute of Biology and Protein Chemistry
Home
Services
Teaching
Publications
Links

PHD SECONDARY STRUCTURE PREDICTION METHOD

[Abstract] [NPS@ help] [Original server]

Sequence name (optional) :

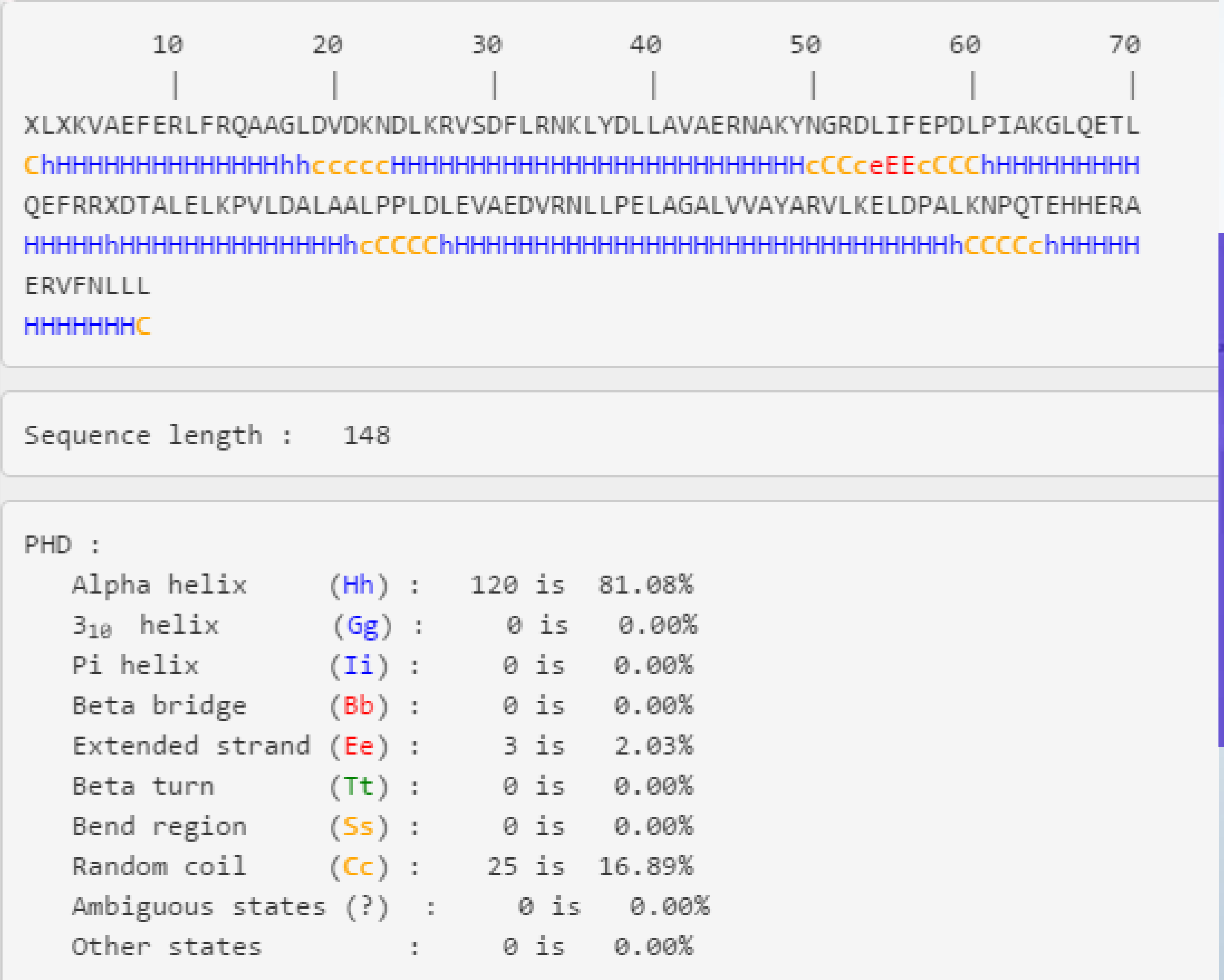
Paste a protein sequence below : help

```
XLXKVAEFERLFRQAAGLDVDKNDLKRVSDFLRNKLYDLLAVAERNAK
YNGRDLIFEPDLPIAKGLQETL
QEFRRXDTALELKPVLDALAALPPLDLEVAEDVRNLLPELAGALVVAYAR
VLKELDPALKNPQTEHHERA
ERVFNLLL
```

Output width : 70

https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_phd.html

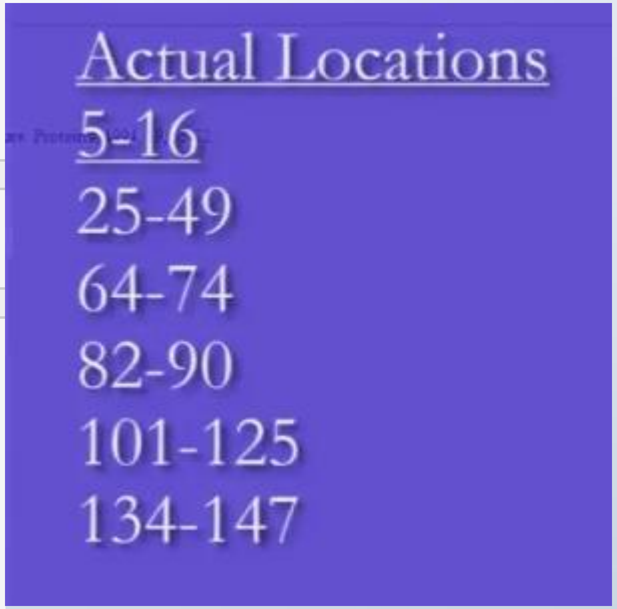
PHD Algorithm



```
        10        20        30        40        50        60        70
         |         |         |         |         |         |         |
XLXKVAEFERLFRQAAGLDVDKNDLKRVSDFLRNKLYDLLAVAERNAKYNGRDLIFEPDLPIAKGLQETL
ChHHHHHHHHHHHHHHhhcccccHHHHHHHHHHHHHHHHHHHHHHHHHHcCCceEEcCCChHHHHHHHHH
QEFRRXDTALELKPVLDALAALPPLDLEVAEDVRNLLPELAGALVVAYARVLKELDPALKNPQTEHHERA
HHHHHhHHHHHHHHHHHHHHhcCCCChHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHhCCCCchHHHHH
ERVFNLLL
HHHHHHHC
```

```
Sequence length :   148
```

```
PHD :
   Alpha helix      (Hh) :   120 is  81.08%
   3₁₀  helix       (Gg) :     0 is   0.00%
   Pi helix         (Ii) :     0 is   0.00%
   Beta bridge      (Bb) :     0 is   0.00%
   Extended strand  (Ee) :     3 is   2.03%
   Beta turn        (Tt) :     0 is   0.00%
   Bend region      (Ss) :     0 is   0.00%
   Random coil      (Cc) :    25 is  16.89%
   Ambiguous states (?)  :     0 is   0.00%
   Other states          :     0 is   0.00%
```



Actual Locations

5-16

25-49

64-74

82-90

101-125

134-147