Working with a single protein sequence
Predicting the main physico-chemical properties of a proteins
1. go to **Primary structure analysis** section, click the on ProtParam
Type in the box `P32851` or alternatively find the `P32851` sequence and paste it in the box.

ProtParam tool

ProtParam (References / Documentation) is a tool which allows the computation of various physical and chemical parameters for a given protein: parameters include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, etc. (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`):

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

**RESET**  **Compute parameters**
1. Click on the blue number
2. Interpreting the ProtParam results

ProtParam
Selection of endpoints on the sequence

**STX1A_RAT** (P32851)

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

Please select one of the following features by clicking on a pair of endpoints, and the corresponding endpoints are highlighted.

- **Syntaxin-1A**.
- Cytoplasmic (Potential).
- Anchor for type IV membrane protein.
- t-SNARE coiled-coil homology.
- Potential.
- Asp-rich (acidic).

<table>
<thead>
<tr>
<th>Feature</th>
<th>Chain</th>
<th>Endpoints</th>
</tr>
</thead>
<tbody>
<tr>
<td>PT</td>
<td>TOP0_DOM</td>
<td>1-288</td>
</tr>
<tr>
<td>PT</td>
<td>TRANSMEM</td>
<td>266-288</td>
</tr>
<tr>
<td>PT</td>
<td>DOMAIN</td>
<td>192-254</td>
</tr>
<tr>
<td>PT</td>
<td>COILED</td>
<td>68-109</td>
</tr>
<tr>
<td>PT</td>
<td>COMPBIAS</td>
<td>13-19</td>
</tr>
<tr>
<td>PT</td>
<td>TURN</td>
<td>6-8</td>
</tr>
<tr>
<td>PT</td>
<td>HELIX</td>
<td>20-63</td>
</tr>
<tr>
<td>PT</td>
<td>HELIX</td>
<td>69-104</td>
</tr>
<tr>
<td>PT</td>
<td>HELIX</td>
<td>111-146</td>
</tr>
<tr>
<td>PT</td>
<td>HELIX</td>
<td>151-156</td>
</tr>
<tr>
<td>PT</td>
<td>HELIX</td>
<td>162-170</td>
</tr>
<tr>
<td>PT</td>
<td>TURN</td>
<td>178-180</td>
</tr>
<tr>
<td>PT</td>
<td>HELIX</td>
<td>192-254</td>
</tr>
</tbody>
</table>
Looking for transmembrane segments
ProtScale: uses sliding window
Sliding window
19–21 transmembrane regions
9–11 globular proteins

Window 1

Window 2

Window 3
Point your browser to: http://www.expasy.org/
go to primary structure analysis
Click on ProtScale
At bottom choose your window size (19)
Enter your sequence ID (P78588) or paste your sequence
Click on
FT CHAIN 1-669 Probable ferric reductase transmembrane

MIN: -1.695
MAX: 2.516
To be sure that the results is meaningful redo the computer analysis with another hydrophobicity scale (Eisenberg scale)
TMHMM: is a program for prediction transmembrane helices based on a hidden Markov model. Program reads a fasta formatted protein sequence and predicts locations of transmembrane, intracellular and extracellular regions.
Point your browser to: http://www.expasy.org/
go to topology prediction
Click on TMHMM
Paste in fasta format the sequence (P78588)
**TMHMM result**

- **sp_P78588_FREL_CANAL Length:** 669
- **Number of predicted TMHs:** 5
- **Exp number of AAs in TMHs:** 126.67
- **Exp number, first 60 AAs:** 0
- **Total prob of N-ini:** 0.892191

<table>
<thead>
<tr>
<th>sp_P78588_FREL_CANAL</th>
<th>TMHMM2.0</th>
<th>inside</th>
<th>1</th>
<th>116</th>
</tr>
</thead>
<tbody>
<tr>
<td>sp_P75858_FREL_CANAL</td>
<td>TMHMM2.0</td>
<td>TMHeli</td>
<td>117</td>
<td>139</td>
</tr>
<tr>
<td>sp_P75858_FREL_CANAL</td>
<td>TMHMM2.0</td>
<td>outside</td>
<td>140</td>
<td>195</td>
</tr>
<tr>
<td>sp_P75858_FREL_CANAL</td>
<td>TMHMM2.0</td>
<td>TMHeli</td>
<td>196</td>
<td>218</td>
</tr>
<tr>
<td>sp_P75858_FREL_CANAL</td>
<td>TMHMM2.0</td>
<td>inside</td>
<td>219</td>
<td>276</td>
</tr>
<tr>
<td>sp_P75858_FREL_CANAL</td>
<td>TMHMM2.0</td>
<td>TMHeli</td>
<td>277</td>
<td>296</td>
</tr>
<tr>
<td>sp_P75858_FREL_CANAL</td>
<td>TMHMM2.0</td>
<td>outside</td>
<td>297</td>
<td>310</td>
</tr>
<tr>
<td>sp_P75858_FREL_CANAL</td>
<td>TMHMM2.0</td>
<td>TMHeli</td>
<td>311</td>
<td>333</td>
</tr>
<tr>
<td>sp_P75858_FREL_CANAL</td>
<td>TMHMM2.0</td>
<td>inside</td>
<td>334</td>
<td>339</td>
</tr>
<tr>
<td>sp_P75858_FREL_CANAL</td>
<td>TMHMM2.0</td>
<td>TMHeli</td>
<td>340</td>
<td>359</td>
</tr>
<tr>
<td>sp_P75858_FREL_CANAL</td>
<td>TMHMM2.0</td>
<td>outside</td>
<td>363</td>
<td>669</td>
</tr>
</tbody>
</table>

**TMHMM posterior probabilities for sp_P78588_FREL_CANAL**

![Graph showing TMHMM posterior probabilities for sp_P78588_FREL_CANAL](image)
Predicting post-translational modification in your protein
Looking for PROSITE

Patterns: small well-conserved segment (signature, motive)
Profile: every position of an entire protein family not just a few highly conserved pattern
Residues and regions thought or proved to be important to the biological function of that group of proteins. These biologically significant regions or residues are generally:
- Enzyme catalytic sites.
- Prosthetic group attachment sites (heme, pyridoxal-phosphate, etc).
- Amino acids involved in binding a metal ion.
- Cysteines involved in disulfide bonds.
- Regions involved in binding a molecule (ADP/ATP, GDP/GTP, calcium, DNA, etc.) or another protein.

A-T-H-[D or E]
(too many false positive results)

[R, T or D]-[D, A or Q]-[F, E or A]-A-T-H-[D or E]
This pattern would probably only pick up the sequences which are in the alignment, but it would be biologically meaningless
Profiles are supposed to be more sensitive and more robust than patterns because they provide discriminatory weights not only for the residues already found at a given position of a motif but also for those not yet found. The weights for those not yet found are extrapolated from the observed amino acid compositions using empiric knowledge about amino acid substitutability. The effect of such a procedure is exemplified below.

Shown are a short alignment without gaps and the corresponding weighting table derived with our standard method.

<table>
<thead>
<tr>
<th>A</th>
<th>F</th>
<th>K</th>
<th>L</th>
<th>L</th>
<th>S</th>
<th>H</th>
<th>C</th>
<th>L</th>
<th>L</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>-18</td>
<td>-10</td>
<td>-1</td>
<td>-8</td>
<td>8</td>
<td>-3</td>
<td>3</td>
<td>-10</td>
<td>-2</td>
<td>-8</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-22</td>
<td>-33</td>
<td>-18</td>
<td>-18</td>
<td>-22</td>
<td>-26</td>
<td>22</td>
<td>-24</td>
<td>-19</td>
<td>-7</td>
</tr>
<tr>
<td>D</td>
<td>-35</td>
<td>0</td>
<td>-32</td>
<td>-33</td>
<td>-7</td>
<td>6</td>
<td>-17</td>
<td>-34</td>
<td>-31</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>-27</td>
<td>15</td>
<td>-25</td>
<td>-26</td>
<td>-9</td>
<td>23</td>
<td>-9</td>
<td>-24</td>
<td>-23</td>
<td>-1</td>
</tr>
<tr>
<td>F</td>
<td>60</td>
<td>-30</td>
<td>12</td>
<td>14</td>
<td>-26</td>
<td>-29</td>
<td>-15</td>
<td>4</td>
<td>12</td>
<td>-29</td>
</tr>
<tr>
<td>H</td>
<td>-13</td>
<td>-12</td>
<td>-25</td>
<td>-25</td>
<td>-16</td>
<td>14</td>
<td>-22</td>
<td>-22</td>
<td>-23</td>
<td>-10</td>
</tr>
<tr>
<td>I</td>
<td>3</td>
<td>-27</td>
<td>21</td>
<td>25</td>
<td>-29</td>
<td>-23</td>
<td>-8</td>
<td>33</td>
<td>19</td>
<td>-23</td>
</tr>
<tr>
<td>K</td>
<td>-26</td>
<td>25</td>
<td>-25</td>
<td>-27</td>
<td>-6</td>
<td>4</td>
<td>-15</td>
<td>-27</td>
<td>-26</td>
<td>0</td>
</tr>
<tr>
<td>L</td>
<td>14</td>
<td>-28</td>
<td>19</td>
<td>27</td>
<td>-27</td>
<td>-20</td>
<td>-9</td>
<td>33</td>
<td>26</td>
<td>-21</td>
</tr>
<tr>
<td>M</td>
<td>3</td>
<td>-15</td>
<td>10</td>
<td>14</td>
<td>-17</td>
<td>-10</td>
<td>-9</td>
<td>25</td>
<td>12</td>
<td>-11</td>
</tr>
<tr>
<td>N</td>
<td>-22</td>
<td>-6</td>
<td>-24</td>
<td>-27</td>
<td>1</td>
<td>8</td>
<td>-15</td>
<td>-24</td>
<td>-24</td>
<td>-4</td>
</tr>
<tr>
<td>P</td>
<td>-30</td>
<td>24</td>
<td>-26</td>
<td>-28</td>
<td>-14</td>
<td>-10</td>
<td>-22</td>
<td>-24</td>
<td>-26</td>
<td>-18</td>
</tr>
<tr>
<td>Q</td>
<td>-32</td>
<td>5</td>
<td>-25</td>
<td>-26</td>
<td>-9</td>
<td>24</td>
<td>-16</td>
<td>-17</td>
<td>-23</td>
<td>7</td>
</tr>
<tr>
<td>R</td>
<td>-18</td>
<td>9</td>
<td>-22</td>
<td>-22</td>
<td>-10</td>
<td>0</td>
<td>-18</td>
<td>-23</td>
<td>-22</td>
<td>-4</td>
</tr>
<tr>
<td>S</td>
<td>-22</td>
<td>-8</td>
<td>-16</td>
<td>-21</td>
<td>11</td>
<td>2</td>
<td>-1</td>
<td>-24</td>
<td>-19</td>
<td>-4</td>
</tr>
<tr>
<td>T</td>
<td>-10</td>
<td>-10</td>
<td>-6</td>
<td>-7</td>
<td>-5</td>
<td>-8</td>
<td>2</td>
<td>-10</td>
<td>-7</td>
<td>-11</td>
</tr>
<tr>
<td>V</td>
<td>0</td>
<td>-25</td>
<td>22</td>
<td>25</td>
<td>-19</td>
<td>-26</td>
<td>6</td>
<td>19</td>
<td>16</td>
<td>-16</td>
</tr>
<tr>
<td>Y</td>
<td>34</td>
<td>-18</td>
<td>-1</td>
<td>1</td>
<td>-23</td>
<td>-12</td>
<td>-19</td>
<td>0</td>
<td>0</td>
<td>-18</td>
</tr>
</tbody>
</table>
1. Point your browser to: http://www.expasy.org/
in the section Tools and software packages
2. Go to Pattern and profile searches (ScanProsite.)
3. Paste in fasta format the sequence (P12259) or type the ID
4. Uncheck the motifs with a high probability of occurrence
5. Check Do not scan profiles
6. Press the start button
## ScanProsite

### Sequence(s) to be scanned:
- UniProtKB(Swiss-Prot and TrEMBL) AC and/or ID (e.g. P00747, ENTK_HUMAN)
- PDB identifier(s)
- your own protein sequence(s)

### Motif(s) to scan for:
- PROSITE AC and/or ID (e.g. PS50808, CHEB)
- your own pattern(s)

### Protein database(s):
- UniProtKB/Swiss-Prot
- including splice variants
- UniProtKB/TrEMBL
- PDB
- randomize databases: no
- excluding fragments

### Filter(s):
- On taxonomy: (e.g. Eukaryota; Escherichia coli)
- On description: (e.g. protease)

### Pattern option(s):

---

**Note:** The image shows a webpage interface for a bioinformatics tool, ScanProsite, which allows for the scanning of protein sequences and motifs.
Interpreting ScanProsite results
1. Interpreting ScanProsite results

PS###: leads you to pattern documentation
The patterns name in bold

**hits by patterns**: [6 hits (by 3 distinct patterns) on 1 sequence]

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Description</th>
<th>Hits on PDB 3D structures:</th>
</tr>
</thead>
<tbody>
<tr>
<td>P12250</td>
<td>(FA5_HUMAN)</td>
<td>[1FV4-H, 1FV4-L, 1Y61-H, 1Y61-L]</td>
</tr>
<tr>
<td>PS00079</td>
<td>MULTICOPPER_OXIDASE1</td>
<td>Hits on PDB 3D structures: [1FV4-H, 1FV4-L, 1Y61-H, 1Y61-L]</td>
</tr>
<tr>
<td>304 - 324</td>
<td>GKNIIsSlTPkhLtqAMGqayI</td>
<td></td>
</tr>
<tr>
<td>1880 - 1900</td>
<td>GwNilnTeVGEnqRAGMqtpF</td>
<td></td>
</tr>
<tr>
<td>PS01265</td>
<td>FA58C_1 Coagulation factors 5/8 type C domain (FA58C) signature 1</td>
<td>Hits on PDB 3D structures: [1CZS-A, 1CZT-A, 1CZV-A, 1CZV-B, 1FV4-L, 1Y61-L]</td>
</tr>
<tr>
<td>1947 - 1980</td>
<td>AWSeklaaafskp...WIqVTmgevItgTqTQG</td>
<td></td>
</tr>
<tr>
<td>2111 - 2140</td>
<td>AWqakammkq......WEelDlkkktilTqTQG</td>
<td></td>
</tr>
<tr>
<td>PS01266</td>
<td>FA58C_2 Coagulation factors 5/8 type C domain (FA58C) signature 2</td>
<td>Hits on PDB 3D structures: [1CZS-A, 1CZT-A, 1CZV-A, 1CZV-B, 1FV4-L, 1Y61-L]</td>
</tr>
<tr>
<td>2045 - 2061</td>
<td>Ptrywrapt...LRLELqGC</td>
<td></td>
</tr>
<tr>
<td>2205 - 2221</td>
<td>Pktwqiai...LRLELqGC</td>
<td></td>
</tr>
</tbody>
</table>

**hits by patterns with a high probability of occurrence or by user-defined patterns**: [115 hits (by 7 distinct patterns) on 1 sequence]
1. Interpreting ScanProsite results

**PS###**: leads you to pattern documentation
The patterns name in bold

**Structure**: if exist you can see the 3D structure in PDB databases

---

**PDB structure viewer**

**PROSITE PS00079 matches on PDB 1FV4-H THREE DIMENSIONAL MODEL OF COAGULATION FACTOR VA**

Only chain H of the structure has been displayed (COAGULATION FACTOR VA HEAVY CHAIN). Click here to view the full structure (1FV4)

---

**Protein**

displayed as ribbons and colored in structure

**PROSITE entry PS00079 MULTICOPPER_OXIDASE1 (PATTERN)**

1 match displayed as ball-and-stick and colored in green

*Equivalent RasMol command: select *h and (276-296H)*
1. Interpreting ScanProsite results

PS#####: leads you to pattern documentation
The patterns name in bold
Structure: if exist you can see the 3D structure in PDB databases
The list: a list of all segments that contain patterns

Hits by patterns with a high probability of occurrence or by user-defined patterns: [115 hits (by 7 distinct patterns) on 1 sequence]
Interpreting ScanProsite results

1. PS###: leads you to pattern documentation
2. The patterns name in bold
3. Structure: if exist you can see the 3D structure in PDB databases
4. The list: a list of all segments that contain patterns
5. Being careful with short patterns (your output shows 19 myristillation position. It is true?
6. Using the species information: when you find a post-translation modification be sure that this modification occur in your species
7. Eliminating a weak patterns: make an alignment since a pattern that correspond to a genuine post-modification is normally well conserved
8. Everything is not in Prosite
Finding Known domains in your protein

Domains: independent globular folding units (is a portion of protein that can keeps its shape if you remove it from the rest of protein) A domain consist of at least 50 amino acids. A average protein consist of 2 or 3 domains.
Nell’ambito della stessa proteina strutture secondarie diverse possono organizzarsi in domini

gliceraldeide-3-fosfato deidrogenasi

in verde il dominio che lega il substrato

in rosso il dominio che lega il coenzima
Finding Known domains in your protein

In principle and in practice there is not a lot differences to search patterns or domains in a proteins. The trouble is that exist many sites each with a own collection of domains. If we wont perform a correct search of domains we have to try all sites.
Fortunately the InterPro project made the possibility to search many domain databases simultaneously.

Unfortunately, InterPro is not exhaustive. To be exhaustive you must search the three domain servers: 1 InterPro, 2 CD-Search, 3 Motif-Scan