Non-globular proteins
A case for protein “dark matter”??
The challenge of increasing Pfam coverage of the human proteome

Jaina Mistry$^{1,2}$, Penny Coggill$^{1,2}$, Ruth Y. Eberhardt$^{1,2}$, Antonio Deiana$^3$, Andrea Giansanti$^{3,4}$, Robert D. Finn$^5$, Alex Bateman$^1$ and Marco Punta$^{1,2,*}$

Figure 1. Pfam-A coverage of *H. sapiens*, *S. cerevisiae* and *E. coli*. Sequence coverage (blue) is calculated as the percentage of the proteome (Methods) that matches at least one Pfam-A family. Residue coverage (red) is the percentage of amino acids in the proteome that are covered by a Pfam-A family.
The challenge of increasing Pfam coverage of the human proteome

Jaina Mistry\textsuperscript{1,2}, Penny Coggill\textsuperscript{1,2}, Ruth Y. Eberhardt\textsuperscript{1,2}, Antonio Deiana\textsuperscript{3}, Andrea Giansanti\textsuperscript{3,4}, Robert D. Finn\textsuperscript{5}, Alex Bateman\textsuperscript{1} and Marco Punta\textsuperscript{1,2,*}

Table 2. Percentage of residues predicted to be compositionally biased in Pfam-A families, Pfam-B families and in regions that are not Pfam-A, not Pfam-B, not predicted to be signal peptides and of at least 50 consecutive amino acids in length.

<table>
<thead>
<tr>
<th></th>
<th>Pfam-A</th>
<th>Pfam-B</th>
<th>Not (Pfam-A, Pfam-B, signal peptide), ≥50aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coiled-coil</td>
<td>2.1</td>
<td>4.9</td>
<td>3.8</td>
</tr>
<tr>
<td>Disorder</td>
<td>9.3</td>
<td>42.0</td>
<td>38.5</td>
</tr>
<tr>
<td>Low complexity</td>
<td>5.1</td>
<td>13.8</td>
<td>13.2</td>
</tr>
<tr>
<td>Signal peptide</td>
<td>0.2</td>
<td>1.2</td>
<td>0</td>
</tr>
<tr>
<td>Transmembrane</td>
<td>6.2</td>
<td>1.9</td>
<td>2.5</td>
</tr>
</tbody>
</table>
The challenge of increasing Pfam coverage of the human proteome

Jaina Mistry¹,², Penny Coggill¹,², Ruth Y. Eberhardt¹,², Antonio Deiana³, Andrea Giansanti³,⁴, Robert D. Finn⁵, Alex Bateman¹ and Marco Punta¹,²,*

Table 1. Top ten largest clusters of human regions not covered by Pfam

<table>
<thead>
<tr>
<th>Cluster number</th>
<th>Number of regions in the cluster</th>
<th>Region length in amino acids (mean)</th>
<th>Phmmr UniProtKB matches (mean)</th>
<th>Number of phmmr matches with overlaps to Pfam-A families (mean)</th>
<th>Likely annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>395</td>
<td>138</td>
<td>27,337</td>
<td>4,023</td>
<td>Olfactory receptors</td>
</tr>
<tr>
<td>2</td>
<td>154</td>
<td>121</td>
<td>58,170</td>
<td>57,777</td>
<td>Zinc fingers</td>
</tr>
<tr>
<td>3</td>
<td>104</td>
<td>121</td>
<td>20,299</td>
<td>19,748</td>
<td>Zinc fingers</td>
</tr>
<tr>
<td>4</td>
<td>85</td>
<td>151</td>
<td>17,578</td>
<td>14,380</td>
<td>Collagen repeat</td>
</tr>
<tr>
<td>5</td>
<td>76</td>
<td>127</td>
<td>3836</td>
<td>3,033</td>
<td>YWTD motifs</td>
</tr>
<tr>
<td>6</td>
<td>62</td>
<td>130</td>
<td>2190</td>
<td>1,958</td>
<td>Leucine-rich repeats</td>
</tr>
<tr>
<td>7</td>
<td>56</td>
<td>123</td>
<td>9662</td>
<td>8,058</td>
<td>EGF domains</td>
</tr>
<tr>
<td>8</td>
<td>46</td>
<td>158</td>
<td>23,652</td>
<td>23,285</td>
<td>Zinc fingers</td>
</tr>
<tr>
<td>9</td>
<td>41</td>
<td>203</td>
<td>701</td>
<td>297</td>
<td>PRY domain</td>
</tr>
<tr>
<td>10</td>
<td>40</td>
<td>178</td>
<td>677</td>
<td>2</td>
<td>Cadherin cytoplasmic domains</td>
</tr>
</tbody>
</table>

Mean number of UniProtKB matches is based on running each region in the cluster against UniProtKB with phmmr. The number of matches with E-value <10⁻³ is collected, and the average is taken over all regions in a cluster. Overlaps with existing Pfam-A families are calculated based on sequences that match simultaneously a cluster member (according to alignment co-ordinates in phmmr output) and a Pfam-A family (according to alignment co-ordinates in Pfam 27.0). ‘Likely annotation’ is assigned based on analysis of overlapping Pfam clans (when a family is not in a clan, it is considered as being in a clan by itself) and on manual inspection of region annotation in UniProtKB, InterPro (18) and Pfam.
A case for protein “dark matter” ??

Unexpected features of the dark proteome

Nelson Perdigão, Julian Heinrich, Christian Stolte, Kenneth S. Sabidó, Michael J. Buckley, Bruce Tabor, Beth Signall, Brian S. Gloss, Christopher J. Hammang, Burkhard Rost, Andrea Schafferhans, and Sean L. O’Donoghue

Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisbon, Portugal; Instituto de Sistemas e Robótica, 1049-001 Lisbon, Portugal; Commonwealth Scientific and Industrial Research Organisation (CSIRO), Sydney, NSW 1670, Australia; Genomics and Epigenetics Division, Garvan Institute of Medical Research, Sydney, NSW 2010, Australia; School of Information Technology, The University of Sydney, Sydney, NSW 2006, Australia; Department for Bioinformatics and Computational Biology, Technische Universität München, 80333 Munich, Germany; and School of Molecular Bioscience, The University of Sydney, Sydney, NSW 2006, Australia

Edited by Alan R. Fersht, Medical Research Council Laboratory of Molecular Biology, Cambridge, United Kingdom, and approved October 13, 2015 (received for review April 29, 2015)

We surveyed the “dark” proteome—that is, regions of proteins never observed by experimental structure determination and inaccessible to homology modeling. For 546,000 Swiss-Prot proteins, we found that 44–51% of the proteome in eukaryotes and viruses was dark, compared with only ~14% in archaea and bacteria. Surprisingly, most of the dark proteome could not be accounted for by conventional explanations, such as intrinsic disorder or transmembrane regions. Nearly half of the dark proteome comprised dark proteins, in which the entire sequence lacked similarity to any known structure. Dark proteins fulfill a wide variety of functions, but a subset showed distinct and largely unexpected features, such as association with secretion, specific tissues, the endoplasmic reticulum, disulfide bonding, and proteolytic cleavage. Dark proteins also had short sequence length, low evolutionary reuse, and few known interactions with other proteins. These results suggest new research directions in structural and computational biology.

In fact, discoveries have already resulted from studying regions of unknown structure, namely, intrinsically disordered regions. Long known to confound structure determination (15)—thus forming part of the dark proteome—disorder was largely ignored until recently (16) and yet is now known to play key functional roles, especially in eukaryotes (17). In addition, there is a second type of region that often has unknown structure and is associated with specific biological functions, namely, transmembrane segments (18). Thus, both disorder and transmembrane regions are “known unknowns” (i.e., we know that they are often “dark”). Could the dark proteome contain “unknown unknowns” (i.e., regions with specific functions that confound structure determination and that we are unaware of)?

To address this question, we need to map the dark proteome (i.e., determine all protein regions that cannot be modeled onto any PDB structure). Most available modeling datasets—collected in the Protein Model Portal (PMP) (2)—are not well suited because they aim for breadth of coverage, typically providing only a few PDB

Significance

A key remaining frontier in our understanding of biological systems is the “dark proteome”—that is, the regions of proteins where molecular conformation is completely unknown. We systematically surveyed those regions, finding that nearly half of the proteome in eukaryotes is dark and that, surprisingly, most of the darkness cannot be accounted for. We also found that the dark proteome has unexpected features, including an association with secretory tissues, disulfide bonding, low evolutionary conservation, and very few known interactions with other proteins. This work will help future research shed light on the remaining dark proteome, thus revealing molecular processes of life that are currently unknown.
**Unexpected features of the dark proteome**

Nelson Perdigão⁴, Julian Heinrich⁵, Christian Stolte⁶, Kenneth S. Sabir⁷, Michael J. Buckley⁵, Bruce Tabor⁶, Beth Signal⁷, Brian S. Gloss⁷, Christopher J. Hammang⁷, Burkhard Rost⁶, Andrea Schafferhans⁷, and Seán I. O’Donoghue⁶,⁷,⁸,⁹,¹

---

**A**

---

**B**

---

**Fig. 1.** Mapping the dark proteome. (A) For all proteins in Swiss-Prot, each residue was classified into one of four categories: (i) PDB regions—residues exactly matched to a PDB entry in Aquaria; (ii) gray regions—residues aligned to at least one PDB entry in Aquaria but always with amino acid substitutions (dark gray); (iii) dark regions—residues with no matching PDB entry in Aquaria; and (iv) dark proteins, where a single dark region spans the entire sequence. (B) We then calculated the total fraction of residues in each of the above four categories for all proteins in eukaryotes, bacteria, archaea, and viruses. The dark proteome (i.e., the fraction of residues in dark proteins or dark regions) varies from 13% (bacteria) to 54% (viruses).
Unexpected features of the dark proteome

Nelson Perdigão, Julian Heinrich, Christian Stolte, Kenneth S. Sabir, Michael J. Buckley, Bruce Tabor, Beth Signal, Brian S. Gloss, Christopher J. Hamman, Burkhard Rost, Andrea Schafferhans, and Seán I. O’Donoghue

Fig. 3. Known vs. unknown dark proteins. Each linear diagram (38) shows known dark proteins [i.e., those with $\geq25\%$ of residues disordered (magenta), compositionally biased (blue), transmembrane (green), or both disordered and compositionally biased (stripes)]. The remaining fraction (gray) are unknown unknowns (i.e., dark proteins predominately ordered, globular, and low in compositional bias). (A) In eukaryotes, high disorder accounted for most of the known dark proteins. Most dark proteins with high compositional bias were also highly disordered. (B and C) In bacteria and archaea, highly transmembrane proteins accounted for most of the known dark proteins (consistent with Figs. 54 and 55). (D) Viruses had the largest unknown unknown fraction and, like eukaryotes, had a large fraction of highly disordered dark proteins.
A case for protein “dark matter” ??
Non-globular proteins (NGPs)

Unlike their globular counterparts, non-globular proteins do not have the typical globular shape. They tend to be elongated

- Globular proteins
- Non-globular proteins
  - Intrinsically disordered (IDPs)
  - Tandem repeated
  - Aggregating

Open questions
- Determinants?
- Structure?
- Interactions?
- Function?
- Classification?
Non-globular proteins
Towards an understanding of the "dark matter" in the protein universe
Secondary structure
Secondary structure of proteins

- Predicting the secondary structure of a protein is an important first step commonly used for classification and modelling.

- Definitions of secondary structure through the spatial coordinates of amino acids and their torsion angles in PDB files.
  - Prediction methods based on AA propensity to certain secondary structures (Chou & Fasman).
  - Advanced prediction methods based on neural networks.
**Chou & Fasman method (1974)**

Amino acids have different propensities for α-helix and β–strand structures. E.g. **proline** interrupts α-helices. Prediction rules exploiting conformational propensity parameters are obtained from a statistical analysis of crystallographic data.

<table>
<thead>
<tr>
<th>Residue</th>
<th>$P_\alpha$</th>
<th>α-Type</th>
<th>Residue</th>
<th>$P_\beta$</th>
<th>β-Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glu</td>
<td>1.51</td>
<td></td>
<td>Val</td>
<td>1.70</td>
<td></td>
</tr>
<tr>
<td>Met</td>
<td>1.45</td>
<td></td>
<td>Ile</td>
<td>1.60</td>
<td></td>
</tr>
<tr>
<td>Ala</td>
<td>1.42</td>
<td></td>
<td>Tyr</td>
<td>1.47</td>
<td></td>
</tr>
<tr>
<td>Leu</td>
<td>1.21</td>
<td></td>
<td>Phe</td>
<td>1.38</td>
<td></td>
</tr>
<tr>
<td>LYS</td>
<td>1.16</td>
<td></td>
<td>Thr</td>
<td>1.37</td>
<td></td>
</tr>
<tr>
<td>Phe</td>
<td>1.13</td>
<td></td>
<td>Leu</td>
<td>1.30</td>
<td></td>
</tr>
<tr>
<td>Gln</td>
<td>1.11</td>
<td></td>
<td>Cys</td>
<td>1.19</td>
<td></td>
</tr>
<tr>
<td>Trp</td>
<td>1.08</td>
<td></td>
<td>Thr</td>
<td>1.19</td>
<td></td>
</tr>
<tr>
<td>Ile</td>
<td>1.08</td>
<td></td>
<td>Gln</td>
<td>1.10</td>
<td></td>
</tr>
<tr>
<td>Val</td>
<td>1.06</td>
<td></td>
<td>Met</td>
<td>1.05</td>
<td></td>
</tr>
<tr>
<td>Asp</td>
<td>1.01</td>
<td></td>
<td>Arg</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td>His</td>
<td>1.00</td>
<td></td>
<td>Asn</td>
<td>0.89</td>
<td></td>
</tr>
<tr>
<td>Arg</td>
<td>0.98</td>
<td></td>
<td>His</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>Thr</td>
<td>0.83</td>
<td></td>
<td>Ala</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>Ser</td>
<td>0.77</td>
<td></td>
<td>Ser</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>Cys</td>
<td>0.70</td>
<td></td>
<td>Gly</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>Tyr</td>
<td>0.69</td>
<td></td>
<td>Lys</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>Asn</td>
<td>0.67</td>
<td></td>
<td>Pro</td>
<td>0.55</td>
<td></td>
</tr>
<tr>
<td>Pro</td>
<td>0.57</td>
<td></td>
<td>Asp</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>Gly</td>
<td>0.57</td>
<td></td>
<td>Glu</td>
<td>0.37</td>
<td></td>
</tr>
</tbody>
</table>

- $N_j$ = total number of residues in state $j$ (i.e. α-helix, β–strand, coil)
- $N_T$ = total number of residues
- $n_{ij}$ = frequency of amino acid $i$ in state $j$
- $n_i$ = frequency of amino acid $i$ (e.g. GLY)

The propensity ($P_{ij}$) of amino acid $i$ to be in state $j$ is

$$P_{ij} = \frac{n_{ij}/n_i}{N_j/N_T}$$
Accuracy

- The most common and intuitive measure is $Q_3$ that is the percentage of correctly predicted residues
  \[
  Q_3 = \left( \frac{\sum_{i=a, \beta, \text{loop}} M_i}{N} \right) \times 100
  \]
  - $N$ is the total number of residues,
  - $M_i$ are correct predictions ($\alpha$, $\beta$, loop)

- Another measure is the SOV (segment overlap) that tends to further penalize presence and/or absence of whole secondary structure elements

- Commonly, SOV values are around 5-6% below $Q_3$ values
Other single sequence-based methods

- I GENERATION methods - Compute a propensity for secondary structures on single sequences with a very narrow local context, around 50-60% $Q_3$ (1970s-80s)
  - GOR method (Garnier-Osguthorpe-Robson, 1978) is a modification of the previous method with a wider local context window ($Q_3$ 50-60%)

- II GENERATION methods - Used until the 1990s these methods use wider local contexts (up to 51 residues) and multiple alignment. These methods are prone to some problems and have $Q_3$ just above 60%
  - Predicted secondary structure segments shorter than in reality
  - Beta sheets are inefficiently predicted (random) since H-bonds form among residues distant in the sequence
Machine learning

**Learning phase**

- **Input**
- **Black box** *(neural network)*
- **Output (known)*

**Prediction phase**

- **Prediction**
- **Black box** *(neural network)*
- **Input**

**Example:** Prediction of secondary structure. Sequence AA → {α-helix, β-strand, coil}
Neural Network (neural net – NN)

Output (CCHHHH)

Input (ABCDEFG)

Prediction phase

Prediction

Black box (neural network)

Input
Neural network for secondary structure prediction

- Input sequence is a window of 13 positions centered around a residue (leucine)
- Each residue of the window is a group of 21 units, one for each type of residue plus one for gaps. For 13 residues, there will be 13x21 input units
- Each group is connected to the output layer by variable weights, defined during the learning phase.
- The secondary structure for the central input residue is the output with the highest value
Profile HeiDelberg (**PHD**)

- First “third generation” method (**Rost & Schneider, 1993**), $Q_3$ around 72%
- Two important innovations
  - Information on the homologous sequences (extracted from HSSP)
  - Three levels of prediction to refine the predictions and reduce local errors
PHD

- Input: profile from multiple sequence alignment (BLAST or ad hoc)
- The novelty is to use more than one neural network, in this way you can remove small local errors, e.g. α-helices less than 4 residues long.
Beyond PHD: **PSIPRED** (Jones, 2000)

PSI-BLAST profiles and (much) larger training set than used in PHD

15 residues window chosen after various trainings on different windows

Two neural networks

- The second, decides the propensity for secondary structure of the residue on the basis of the first input
- $Q_3$ above 75-76%

![Diagram of PSIPRED method](image)

*Figure 1. An outline of the PSIPRED method, which shows how the PSI-BLAST score matrices are processed.*
Secondary Structure in Practice

- The best methods (e.g., PSIPRED) reach a prediction accuracy around 80%
  - This means the output is wrong for 20% of residues
  - In particular cases the accuracy may be much higher or much lower!
- This depends on how the methods have been trained, on the accuracy of the test set, and if the protein of interest follows general rules or is new
- Evaluating the reliability score of the prediction for each amino acid or using multiple methods and comparing them can help
- Conclusions can be drawn based on their agreement
- In case of discrepancy
  - The prediction may not be reliable
  - It could be symptomatic of a flexible region, which undergoes a conformational switch
Transmembrane proteins
Transmembrane proteins
Transmembrane proteins

- Experimental determination of structures is difficult (NMR is hard, X-ray do not crystallize). Due to this, few crystallized structures exist.

- **20% of proteins coded by a genome generally are transmembrane**

- Prediction is made easier because membrane constraints reduce degrees of freedom.

- Two kinds of proteins:
  - **Alpha-helices**: well-defined properties. Being inserted into the membrane limits the possible conformations. Allows structural and topological analysis.
  - **Beta-strands**: porin conformation. Little experimental data. Accuracy prediction hard to estimate (no baseline).
Transmembrane proteins (TM)

Out of ca. 200K PDB structures about 7K (3.7%) are TM

- 6,5 K alpha TM
- 0.5 K beta TM
Transmembrane proteins

Distinction can be made between membrane segment prediction, which requires only the prediction of local context, and topology prediction.

Correct prediction of the topology means:

1. Identify correctly all the membrane segments
2. Predict loop (inside or outside) orientation
Transmembrane proteins

Alpha helices have an orientation or topology based on N-terminus. **IN** if N-terminus is intracellular, **OUT** otherwise.

The following characteristics can be distinguished:

1. Long fragments (12-35 residues) and typically very hydrophobic, thus easy to predict
2. Outside loops are typically shorter than 60 AA
3. "Positive-inside-rule", connection loops are more positively charged if inside. Enrichment of ARG and LYS (von Heijne, 1986)
4. Globular regions > 60 AA do not satisfy the positive-inside-rule

**Kyte e Doolittle (1982)**
One of the earliest accurate methods for protein transmembrane prediction is **HMMTOP** (Tusnády & Simon, 1998)

- TM segments are correctly predicted in 98% of cases, topology in 78%
- Uses an **HMM (hidden Markov model)** to learn the “reality” of transmembrane proteins and deduce rules
- The varying amino acid distribution on the different parts of transmembrane proteins is used to train the HMM
- Parameters
  - Membrane
    - $MIN_L = 17$
    - $MAX_L = 25$
  - Tail
    - $MIN_L = 1$
    - $MAX_L = 15$

- Alternative $\rightarrow$ TMHMM (Krogh et al. 2001)
- Similar accuracy, about 80% for the topology
## TM predictors

<table>
<thead>
<tr>
<th>Method</th>
<th>Release year</th>
<th>Algorithm</th>
<th>URL</th>
<th>MSA</th>
<th>SP</th>
<th>Constrained</th>
<th>Training data</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMMTOP</td>
<td>2001</td>
<td>HMM</td>
<td><a href="http://www.molbio.tau.ac.il/HMMTOP/">http://www.molbio.tau.ac.il/HMMTOP/</a></td>
<td>*</td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>MEMSAT1</td>
<td>1994</td>
<td>ANN + Grammar</td>
<td><a href="http://single.topcons.net/">http://single.topcons.net/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEMSAT3</td>
<td>2007</td>
<td>ANN + Grammar</td>
<td><a href="http://bioinf.cs.ucl.ac.uk/psipred/">http://bioinf.cs.ucl.ac.uk/psipred/</a></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEMSAT-SVM</td>
<td>2009</td>
<td>SVM + Model</td>
<td><a href="http://bioinf.cs.ucl.ac.uk/psipred/">http://bioinf.cs.ucl.ac.uk/psipred/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OCTOPUS</td>
<td>2003</td>
<td>ANN + HMM</td>
<td><a href="http://octopus.cbr.su.se/">http://octopus.cbr.su.se/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Philius</td>
<td>2008</td>
<td>DBN</td>
<td><a href="http://yeast-cterm.cse.ucsc.edu/philius/">http://yeast-cterm.cse.ucsc.edu/philius/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phobius</td>
<td>2004</td>
<td>HMM</td>
<td><a href="http://phobius.sbc.su.se/">http://phobius.sbc.su.se/</a></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRO</td>
<td>2004</td>
<td>HMM</td>
<td><a href="http://topcons.net/">http://topcons.net/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRODIV</td>
<td>2004</td>
<td>HMM</td>
<td><a href="http://topcons.net/">http://topcons.net/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCAMP1</td>
<td>2008</td>
<td>Hydrophobicity + Model</td>
<td><a href="http://scamp1.cbr.su.se/">http://scamp1.cbr.su.se/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCAMPi – single</td>
<td>2008</td>
<td>Hydrophobicity + Model</td>
<td><a href="http://scamp1.cbr.su.se/">http://scamp1.cbr.su.se/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPOCTOPUS</td>
<td>2008</td>
<td>ANN + HMM</td>
<td><a href="http://octopus.sbc.su.se/">http://octopus.sbc.su.se/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TM-HMM</td>
<td>2001</td>
<td>HMM</td>
<td><a href="http://www.cbs.dtu.dk/services/TMHMM/">http://www.cbs.dtu.dk/services/TMHMM/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TopPred2</td>
<td>1994</td>
<td>Hydrophobicity</td>
<td><a href="http://mobyle.pasteur.fr/">http://mobyle.pasteur.fr/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**MSA:** the predictor uses multiple sequence alignments as input; **+** the original version of HMMTOP is capable of utilizing homologous sequences to improve predictions, however we benchmark the single-sequence version; **SP:** the predictor also predicts signal peptides; **Constrained:** the predictor allows constrained predictions (i.e. it allows using (experimentally derived) topology information as input); **Training data:** the predictor is trained on high-resolution (highres) or low-resolution (lowres) topology information; **Filter:** the predictor incorporates a filter for non-TM (globular) proteins; **+:** The method is used as input to TOPCONS-single, and is prefiltered with the SCAMPi-single filter for non-TM (globular) proteins.
Transmembrane Proteins

Two chains from aquaporin-1 (PDB code 1h6i) with re-entrant helices shown in orange.

Each chain of the homotetrameric channel protein contains six transmembrane and two re-entrant helices.

Rhodopsin structure (PDB code 1gzm) with an interfacial helix shown in red.

Evidence suggests that, in addition to its role in binding the G protein transducin, the interfacial helix acts as a membrane-dependent conformational switch domain which adopts a helical structure only in the presence of membranes (Krishna et al., 2002).
TM Structure in Practice

- TM methods reliability is difficult to evaluate because few experimentally resolved TM structures exist. Thus, if on one hand TM structures are subject to strict rules, on the other hand the few experimental cases might not be enough to create a valid predictive method.

- Since methods as TMHMM and HMMTOP are based on a reduced number of sequences as training-set, they might be over-fitted. They predict well only those cases with very high accuracy. *In practice, their reliability might be overestimated*

- The distinction of Globular vs TM proteins seems to work in any case

- A *consensus* (majority vote) from more than one method improve results (5 methods, *Nilsson et al.*, 2000)
Low complexity proteins
Drosophila mastermind
Drosophila mastermind
SEG
(Wootton 1994)

- A tool that identifies low-complexity sequence regions from entropy in a sequence window
- Globular proteins, typically, have high complexity values, while atypical regions have lower values (e.g. QQQQQQ)
- SEG is used by BLAST as filter to mask low complexity region in sequences to lower the false positive rate (FPR)
## Low complexity regions

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition based on amino acid composition</strong></td>
<td></td>
</tr>
<tr>
<td>LCP</td>
<td>Regions with a skewed amino acid composition</td>
</tr>
<tr>
<td>Compositonally biased region</td>
<td>Region with a high proportion of a specific amino acid, where X is the abundant residue</td>
</tr>
<tr>
<td>X-rich region</td>
<td></td>
</tr>
<tr>
<td><strong>Definition based on amino acid periodicity</strong></td>
<td></td>
</tr>
<tr>
<td>Repeat motif</td>
<td>Reiteration of residues: (...)(n)</td>
</tr>
<tr>
<td>Homorepeat (polyX)</td>
<td>Consecutive runs of a single residue: (X)n</td>
</tr>
<tr>
<td>Disrepeat</td>
<td>Consecutive runs of two ordered different residues: (XY)^n</td>
</tr>
<tr>
<td>Tandem repeat</td>
<td>Pattern of residues which are directly adjacent to each other: (XYZ ...)(n)</td>
</tr>
<tr>
<td>Cryptic repeat</td>
<td>Scrumbled arrangements of repetitive motifs</td>
</tr>
<tr>
<td>Imperfect repeat</td>
<td>Regions in which the repeat units are not the same</td>
</tr>
<tr>
<td><strong>Definition based on structure</strong></td>
<td></td>
</tr>
<tr>
<td>Intrinsically disordered protein</td>
<td>Protein that lacks a fixed or ordered 3D-structure</td>
</tr>
<tr>
<td>Coiled coil</td>
<td>Structural motif characterised by a seven-residue sequence repeat in which alpha-helices are coiled together to form an extended rope-like structure: (α-b-c-d-e-α-g)</td>
</tr>
<tr>
<td>(Charged) single alpha-helix</td>
<td>A segment forming stable monomeric alpha-helix in aqueous solution, typically rich in Arg/Lys/Glu forming an alternating pattern of short runs of oppositely charged residues</td>
</tr>
<tr>
<td>Protein flexibility</td>
<td>Ability of a protein to fold into multiple stable 3D-structures</td>
</tr>
<tr>
<td>Amyloid filaments</td>
<td>Stable insoluble protein assemblies composed predominantly of β-sheet structures in a cross-β conformation</td>
</tr>
</tbody>
</table>

**Figure 1.** The LC diagram: sequence complexity composition versus periodicity. The diagram illustrates where several types of sequences would be placed in relation to two measures related to sequence complexity.

*Disentangling the complexity of low complexity proteins*. Mier et al., Briefings in Bioinformatics 2020