Prediction of protein disorder

Supervisor: Michael Heinzinger
Students: Tba
Prediction of protein disorder [Recap]

- Task: Predict whether a residue in a protein is disordered or not using one-hot encoding (maybe using also other types of encoding, tbd). After predicting binary disorder, one student will focus on “functional disorder” (disorder induces function), while the other student will focus on “binding disorder” (disorder induces binding, e.g. to other proteins).

Prediction of protein disorder [Details]

Input: Protein sequence

Output: ordered (0) vs. disordered (1)

For example:

<table>
<thead>
<tr>
<th>In:</th>
<th>S</th>
<th>E</th>
<th>Q</th>
<th>W</th>
<th>E</th>
<th>N</th>
<th>C</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Out:</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

This is **only** an example. Intrinsically disordered stretches are usually a lot longer (e.g. 30 residues). Shorter stretches are usually linker-loops between e.g. helices (see next slide).
Intrinsic disorder vs. loops

Main criterion for differentiation between loop and disorder: length of the segment and function (loops connect sec. struct. elements, while disorder has function).

Intrinsic disorder

Input: Encoding of protein sequence

Machine learning algorithms work with numbers, not strings. Thus, we need to convert our protein sequence to numbers.

Easiest approach: One-hot-encoding; e.g. vocabulary of size 4 (A,T,C,G)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

You will need to map this DNA-vocabulary to the vocabulary of 20 amino acids (+1 for ‘X’ (=unknown)).
Output: Annotation for order/disorder

Database: DisProt (https://www.disprot.org/)

Remove ambiguous/obsolete entries (flag while downloading, no parsing required)

Create statistics on:

- Number of proteins
- Number of ordered/disordered residues
- Amino acid distribution for ordered vs. disordered (one plot showing the diff.)
- Average length of disordered stretches (loop vs. disorder...)
- Differentiate between different flavors of disorder: “Functional disorder” vs. “Binding disorder” annotation (will require parsing json from DisProt)

Literature


… to be extended if necessary. Write me an email if you should need more literature.