Prediction of Intrinsically Disordered Proteins (IDPs)

Classical dogma: a well defined 3D structure is needed for protein function.
Some findings challenge this dogma:

- There are lots of genes coding for IDPs in higher eukaryotic genomes
- Those IDPs have specific and very important functions
- Their functionality is largely influenced by the fact that they are disordered/have disordered regions

How can disordered regions be useful in protein function?

Some examples:

- Entropic chains
- IDRs can act as flexible linkers between folded domains in multi-domain proteins
- IDRs can bind to DNA/RNA: "coupled folding and binding". This involves a "disordered → more ordered" transition. As the both processes (folding and binding) are coupled, the entropic penalty (from the folding) counterbalances the enthalpy gain (from the binding). This allows for weak transient but specific bindings, very needed in gene regulation, cell cycle control, etc.
- Disordered proteins/regions are very flexible. This is very useful in binding interfaces: different partners can be bound, different conformations can be adopted.
- They are involved in various post-translational modifications that control their functions, localization and turnover → they can integrate and mediate signals and act as central elements of regulatory networks. Proof:
    - Targeted attack of hubs causes network disruption.
    - They are related to several diseases:
        - p53 is involved in 50% of cancers
        - 79% of human cancer associated proteins are IDPs, while only 47% of all eukaryotic proteins in SwissProt are IDPs.
        - May be involved in diabetes, cardiovascular diseases and neurodegenerative diseases.

- IDPs are very attractive targets for drug discovery. This highlights the importance of finding a way to search for disordered binding regions (= functional sites in proteins that undergo disorder → order transitions).

Disordere binding sites

There are very few well-characterized disordered binding regions. Main differences between complexes formed by ordered proteins vs disordered proteins:

Globular proteins	Disordered proteins
They adopt rather compact conformations in the complex.	They adopt very extended conformations inside the complex, exposing the majority of their residues for interacting with their partner.
Interface not that rich in hydrophobic residues.	The interface is enriched with hydrophobic residues, both compared to the interface of ordered proteins and also to disordered regions in general.
Interface formed by very distant segments of the aa sequence, which are brought together by folding.	Interface formed by segments that are quite localized in the primary structure.

Conclusion: the underlying principles of molecular recognition of disordered binding sites are different from the complex formation of globular proteins.

How can we recognise disordered binding sites?
We want to be able to distinguish between disordered regions vs ordered regions. But we also want to be able to distinguish between disordered binding sites and general disordered regions not involved in binding. Disordered binding sites are expected to have some features that enable us to detect them.

Some difficulties:

- Protein disorder "comes in many flavours": specific prediction methods are needed for each type of disordered binding sites.
- We need larger datasets than the ones available today in order to train specific prediction methods.
- It has been suggested that specific patterns of disordered binding sites may be associated to regions undergoing disordered → ordered transitions. There is no clear recipe whether these regions should be considered ordered, disordered or borderline cases.

A recent analysis compared different methods for detection of disordered binding sites that have α-helical elements in their bound state (which are called "α-MoRFs"): a large variablity was found between methods. Nevertheless, those are quite specific methods, they only search for a specific type of disordered binding site.

A general method to identify specific binding regions undergoing disordered→ordered transitions:
Based on IUPred, which consideres that IDPs have a specific aa composition that does not allow them to form a stable well-defined structure.

- We take a dataset of globular proteins
- We calculate the pairwise interaction energy of proteins directly from their aa sequence (the algorithm uses statistical potentials that can be used to calculate the pairwise interaction energy from known coordinates)
- Now, we can detect disordered regions because they have unfavourable estimated pairwise energies.

The energy of each residue is based on the aa type and its neighbour aa-s (in the sequence). This enables IUPred to take into account that the disorder tendency of residues can be influenced by their environment.

We can search for the regions that are most likely to undergo disorder→order transitions by calculating their pairwise energies in different contexts. We estimate the energy in the free state and in the bound state and identify the segments that are potentially sensitive to these changes.

Previously: other methods were more focused on finding regions that were involver in binding externally rather than internally. This method searches for both (makes no distinction between different types of interactions).

Main features of ANCHOR:

- It considers the environment only at the level of aa composition.
- It captures the very essential properties of disordered binding regions.
- It is a robust prediction method.

Outline of the algorithm
Aim: to recognise proteins that undergo a disorder-to-order transition upon binding to a globular protein.
In the free state they are very disordered, and when they are bound to a globular protein they adopt a rigid conformation. The algorithm looks specifically for those proteins that cannot form enough favourable interchain interactions but have the capability to energetically gain by interacting with a globular protein.

How to search for the right residues? We follow three criteria:
1) We look for a residue that belongs to a long disordered region, we discard globular domains.
2) The residue is not able to form enough favourable interactions with its sequential neighbours (neighbouring residues in the sequence, not in the 3D structure!)
3) The residue can form enough favourable interaction with a globular protein. That is to say, there is an energy gain when interacting with globular proteins.

These three criteria are estimated individually and are then combined into a single predictor. How do we do that? Three different scores (that correspond to the three criteria mentioned above) are calculated for each residue and then the final prediction is obtained by calculating a linear combination of those three scores.

IUPred is a general disorder prediction method and it implements an energy estimation framework that we use here to calculate those three scores. The general IUPred formula is the following:

$$E_i^k = \sum_{j=1}^{20} P_{ij} f_j^k (w_0)$$

- P is the energy predictor matrix, which contains the elements $P_{i,j}$ that were calculated using a dataset of globular proteins. Each element of the matrix ($P_{i,j}$) estimates the pairwise energy of a residue type *i* in the presence of a residue type *j*.

- $f^k_j(w_0)$ is the fraction of residue type j in the sequential environment within the neighbourhood w_0.

1st criterion:
In order to estimate the tendency of the neighbourhood of a residue for being disordered, we calculate the score s_k:

$$S_k = \frac{1}{N} \sum_{k \neq j = b_{lower}}^{b_{upper}} s_j$$

- S_j is the IUPred score of the j-th residue in the sequence.
- S_k is the score of the k-th residue in the sequence. It is calculated by averaging the IUPred scores of all the j residues (for j between b_{lower} and b_{upper}).

2nd criterion:
To calculate the pairwise interaction energy gain if the residue interacted with its neighbours, we use again the IUPred formula but we take a neighbourhood size w_2 that is set as a parameter.

$$E_i^{\text{int},k} = \sum_{j=1}^{20} P_{ij} f_j^k (w_2)$$

3rd criterion:
To calculate the pairwise interaction energy gain if the residue interacted with a globular protein, we also use the IUPred formula and using the average amino acid composition of globular proteins.

$$E_i^{glob} = \sum_{j=1}^{20} P_{ij} \overline{f}_{glob,j}$$

- $\bar{f}_{glob,j}$ is the fraction of residue type j in the averaged reference amino acid composition of globular proteins.

In order to obtain the final prediction score, we calculate a linear combination of the score and the two energies calculated above:

$$I_k = p_1 S_k + p_2 E_i^{\mathrm{int},k} + p_3 E_i^{gain,k}$$

- p_1, p_2, and p_3 are coefficients whose values are calculated during the training of the predictor, together with the optimal values for w_1 and w_2. All those optimal values are calculated using a dataset of disordered protein complexes and ordered monomeric proteins using three-fold cross-validation. This means that the original sample is randomly partitioned into *3* equal sized subsamples, a single subsample is used as the validation data for testing the model, and the remaining two subsamples are used as training data.

One of the advantages of the model proposed in the article is that it only relies on disordered proteins datasets (which are scarce and incomplete) for the calculation of these five parameters and uses much bigger and complete globular protein datasets to calculate the rest of the parameters.

Example: p53 is a tumour suppressor gene that has its N-terminal region (residues 1-100) completely disordered. It is able to bind to at least three different globular proteins:

- MDM2 (residues 17-27)
- RPA 70N (33-56)
- RNA polymerase II (45-58)

They divided the globular protein dataset and the short disordered complex dataset in three groups. They used two groups for the training and one for the evaluation (in all three possible combinations).

They optimize the choice of the parameter values by maximising the amount of correctly predicted binding sites (true positives) and minimising the predicted binding sites in globular proteins (false positives). It the parameter values made the method too strict, then we would obtain very few “binding sites” in globular proteins (false positives) but we would also obtain very few true positives. If the parameters made the method two lax, then we would obtain too many false positives. Of course, the choice of parameters must ensure that we will not find binding sites in general disordered regions (disordered regions that cannot bind globular proteins).

How to assess the efficiency of the algorithm?

- True positive rate (TPR), false positive rate (FPR), percentage of binding sites identified.
- Receiver Operating Characteristic (ROC) curves. The relationship between the TPR and the FPR is mapped by scanning the interval between 0 and 1 with the score cutoff.
- The area under the curve (AUC). Random predictors give AUC = 0.5 and perfect ones give AUC = 1.