Predicting Secondary Structures of Protein and Global Optimization

Piotr Berman and Jieun Jeong

- A major goal of bioinformatics: find protein structure (shape) from the sequence data.

The partial task that we focus on:

- given sequence of residues (aminoacids) find secondary and tertiary structures.

Protein structure ≈ shape.

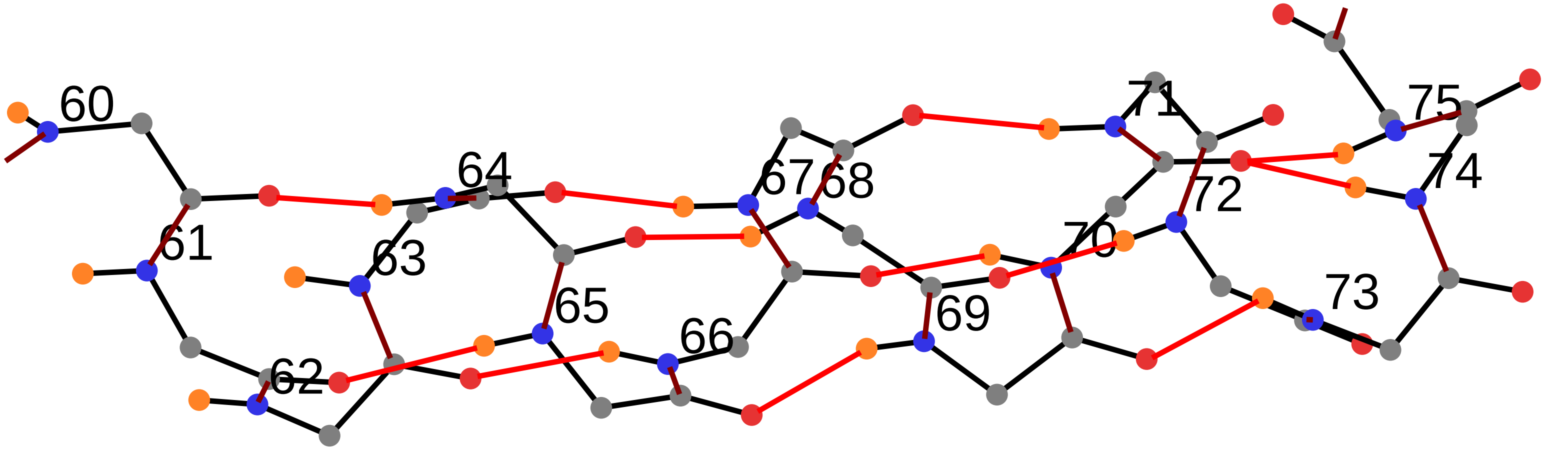
Proteins contain repeating substructures, predicting these substructures is a major part of predicting the shape.

We are interested in secondary structures that can be defined in terms of

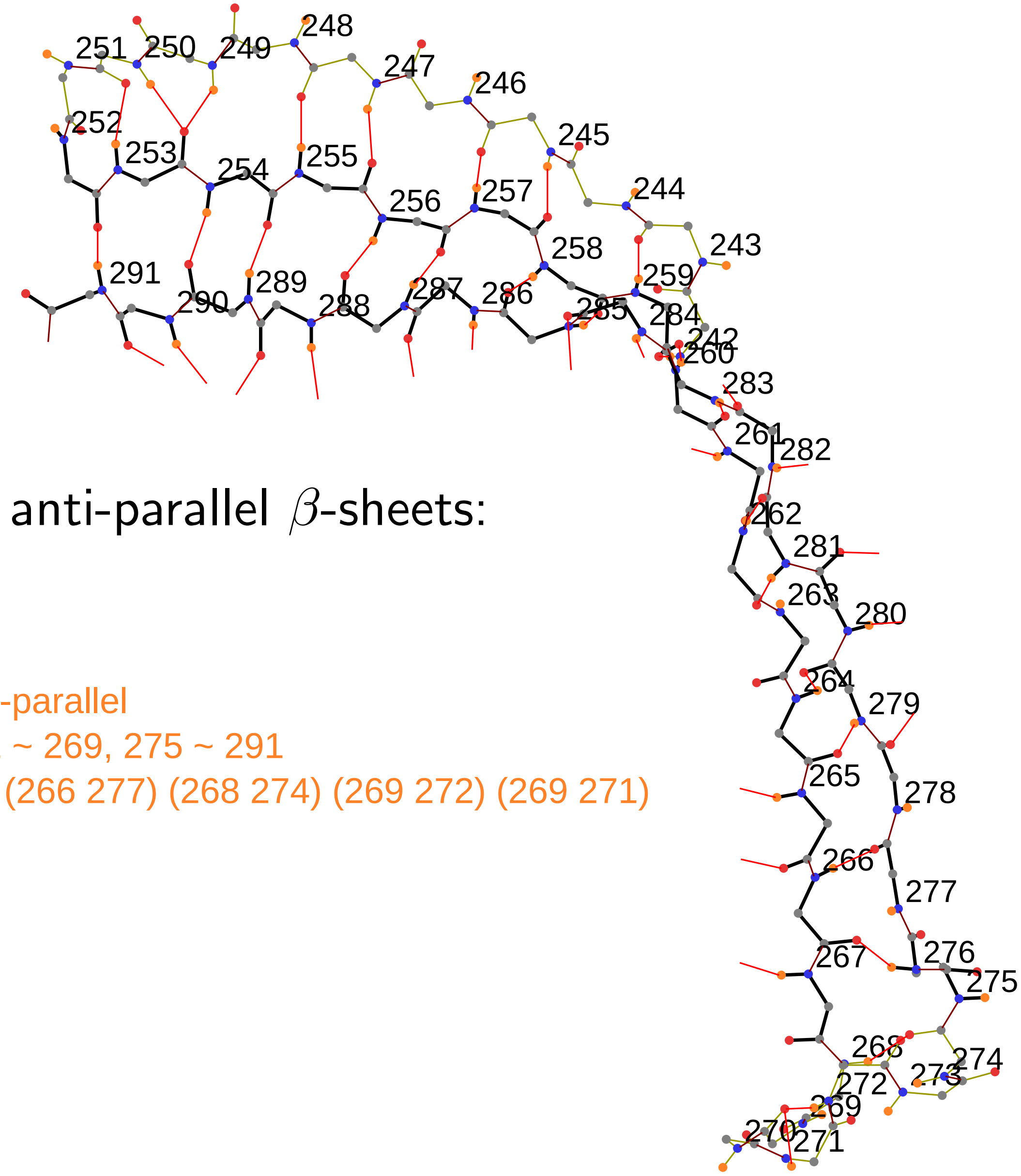
- dihedral angles defined by chemical bonds in the protein backbone, and
- hydrogen bonds between atoms that are directly attached to the backbone.

Such structures can be easily computed given crystallographic data about a protein. The most important secondary structures are α-helices and β-strands, the latter are paired into *parallel* and *anti-parallel* β-sheets.

An example of α-helix:

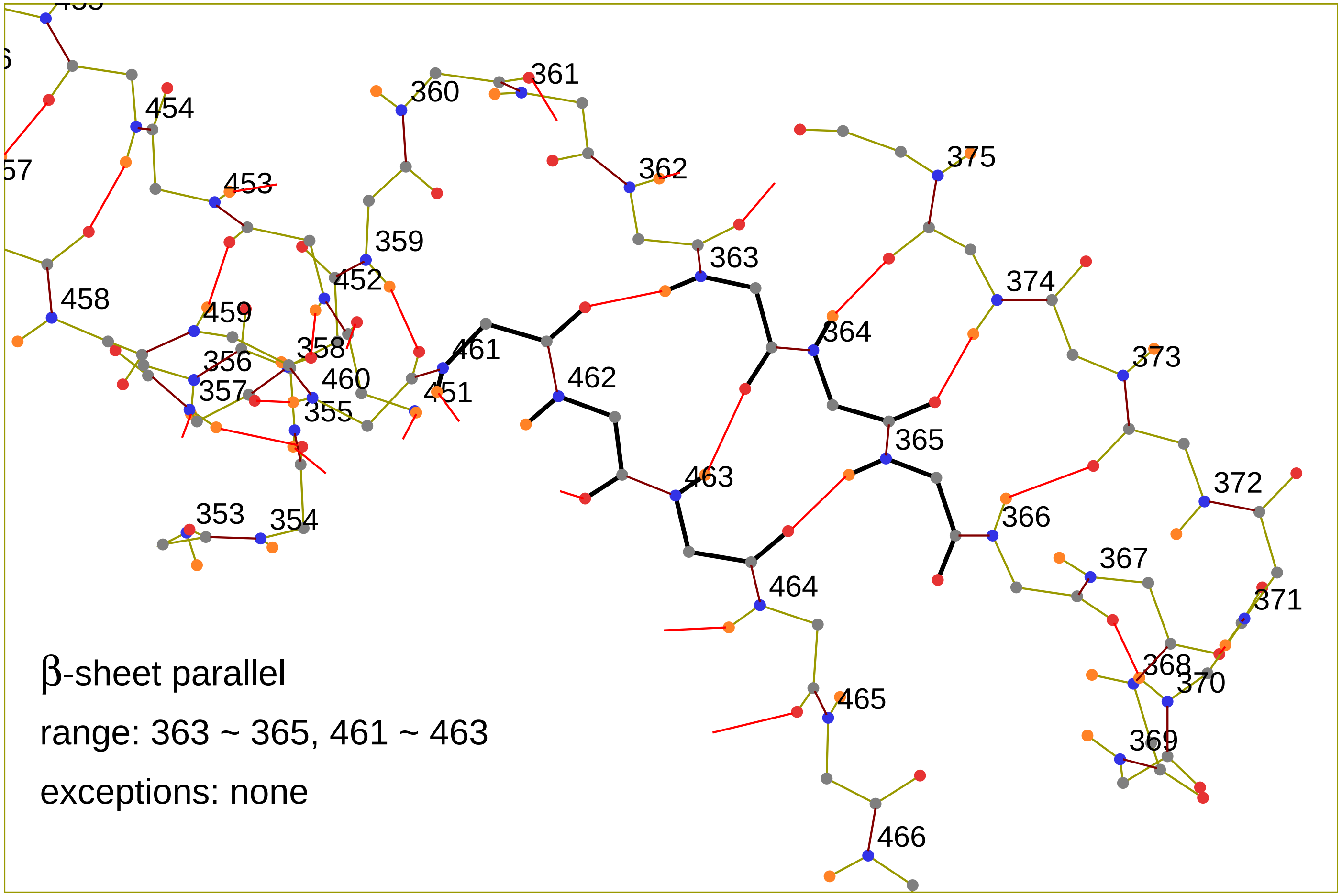


An example of anti-parallel β-sheets:



β-sheet anti-parallel
ranges: 252 ~ 269, 275 ~ 291
exceptions: (266 277) (268 274) (269 272) (269 271)

An example of a parallel β-sheet:



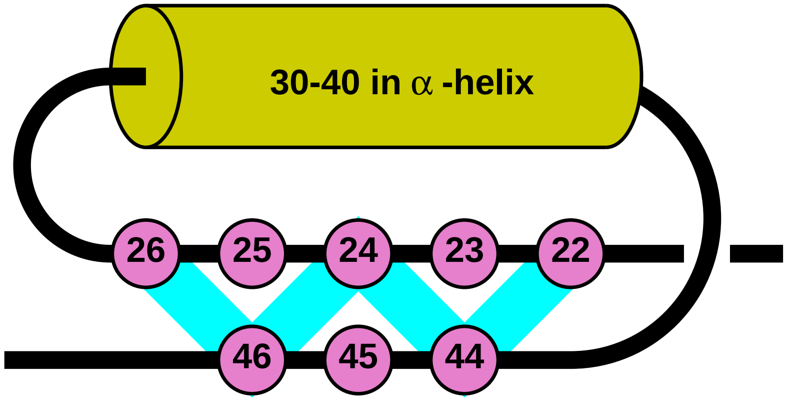
β-sheet parallel

range: 363 ~ 365, 461 ~ 463

exceptions: none

Besides the examples we have seen: α-helices and strands of β-sheets there are other structures like π-helices, β-turns, turns, β-hairpins. They are a bit less interesting because they cannot form periodic patterns and they provide much smaller proportion of the entire protein. Predicting them is important, in particular, they give strong clues about the α-helices and β-strands.

Tertiary structures are arrangements of secondary structures. The most ubiquitous is a 2-stranded β-sheet. Larger tertiary structures are called *motifs*. Many motifs can be defined in terms of α-helices and β-sheets. Hence discovering β-sheets is a major portion of identifying tertiary structures of various sizes.



$\beta - \alpha - \beta$ motif

Existing methods:

To predict if a residue is in an α-helix, β-strand etc. we look at the sequence of 15 residues, with 7 neighbors to the left and right. The information is fed into a *neural network* and out comes a prediction. This method was pioneered by Rost in 1995.

The success rate of prediction was improved by using *profiles*, multiple alignments of protein sequences. The input to the network that describes a residue may have a form "always Phenylalanin", "Phenylalanin or Proline" etc. Some benefits of profiles are analogous to the benefits of multiple alignments for gene identifications – structures are conserved better than loops.

Neural network can be replaced with *support vector machines*, which is basically the same thing, but with a different method of *training*.

Among further improvements, Meiler and Baker coupled neural network predictions with Rosetta program, which basically allows to check if predictions fit together in three dimensions. In turn, Rosetta may find possible structures that were not predicted initially and we get an improved set of predictions for the next run of Rosetta.

Meiler and Baker reported very impressive gains. It would be nice to reproduce their level of success with "white box" method. It is hard to get extra insight from thousands of coefficients produced by training of neural networks.

Possible global optimization method: maximum weight matching.

Around 1995, Hubbard tried to predict β-sheets based on a matrix: given two aminoacids, what is their propensity to be opposite each other in a β-sheet. The results were showing some predictive power, but not as good as the subsequent results of Rost.

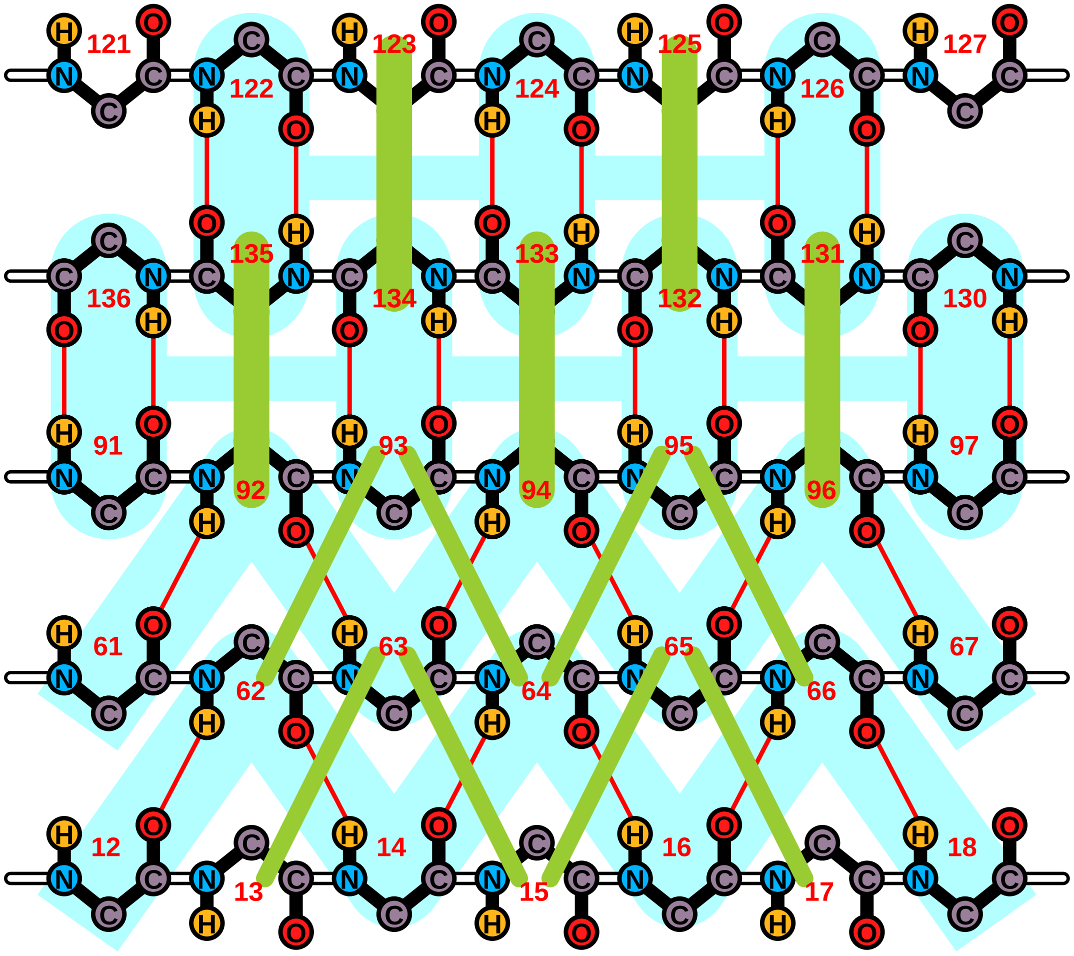
We propose to refine Hubbard's approach in two ways.

First, we want to base our “propensity” assesment based on triples that may face each other rather than single residues. Importantly, such two triples may contain 3-4 hydrogen bonds and they force a number of side-chains to be in contact with each other, so there should be more dependencies.

Second, given such two triples, we can introduce an edge connecting their central residues and with the weight equal to the propensity value. Given such a set of edges, we will search for a *maximum weight* matching. (See the next picture.)

The hope is that wrong prediction would be sufficiently inconsistent to fail to be present in the maximum weight matching.

Fragment of the matching that corresponds to secondary structures.



🟦 highlighting of hydrogen bonds of a β-sheet

▬ an edge of the 2-matching

▬ an edge of the matching

Challenges:

getting propensity values of pairs of triples, given that there are 64M possibilities; we can use protein-alignment distance to tuples observed in the structures recorded in the training set

refining propensity values, can we decrease the values that more often in wrong solutions than others etc., Given edges with a high score, they are meaningful only if used in groups corresponding to plausible structures. We can eliminate isolated edges in the matching problem. We can also use consistent groups as predicted structures. This way each predicted structure obtains a weight.

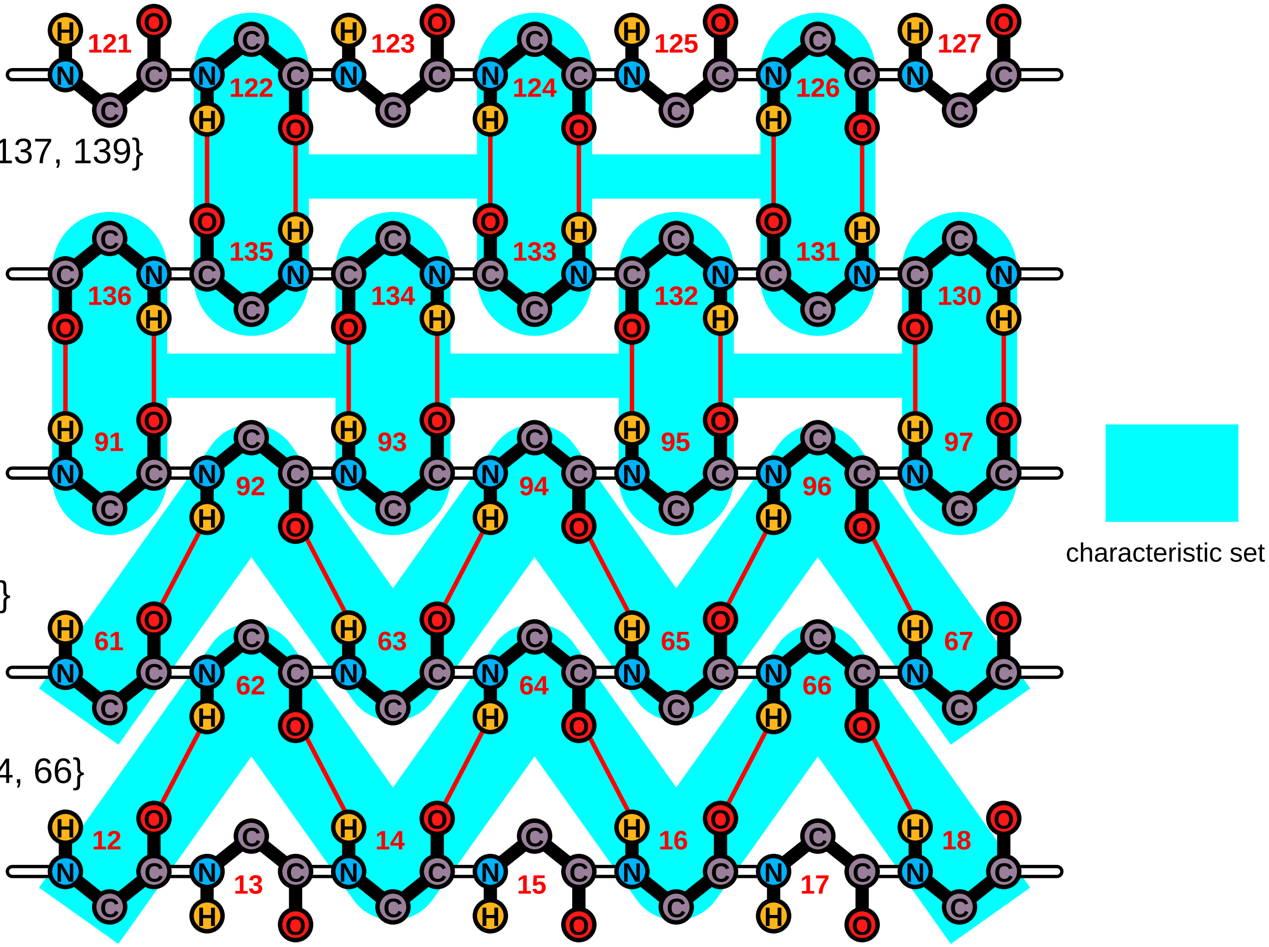
Now we have a combinatorial problem: given a set of plausible predictions, find a consistent subset of maximum weight.

By formalizing the notion *consistent* in several ways we obtain several possible problems.

Possible global optimization method: set packing.

For each predicted structure we can define a *characteristic set* of residue numbers. For an α-helix, this is the set of residues that it includes. For a β-sheet, this is the set of residues that contain hydrogen bonds that define the sheet.

Example of characteristic sets of β-sheets:



{122, 124, 126, 135, 137, 139}

{91, 93, 95, 97,
130,132,134,136}

{61,63,65,67, 92, 94, 96}

{12, 14, 16, 18, 62, 64, 66}

🟦 characteristic set

The definition of characteristics sets of β-sheets: “numbers of residues of the hydrogen bonds of the sheet” has two good consequences:

1. sets of different 2-stranded sheets are disjoint, so we have a set-packing problem;
2. after separating odd numbers from even numbers, characteristic sets have the form of a pair of contiguous intervals of integers, moreover, these intervals differ in size by at most one.

We can define consistency of the predicted structures as the disjointness of their characteristic sets. In that case, we have to solve a weighted set packing problem:

> given a family of sets, each with a weight, maximize the joint weight of a subfamily in which sets are pairwise disjoint.

Bad news: set packing is as hard to approximate as independent set problem, which means, very, very hard.

Good news: property (2) of our sets allows to find 4-approximation in time $O(n^2)$, where n is the number of sets.

Packing of k-tuples of intervals has a $2k$-aproximation based on Lagrangean relaxation (Haldorsson and others). Because intervals have almost equal lengths, one can use a much faster *local ratio* algorithm of Berman and DasGupta.

More bad news: this is an insufficient notion of consistency.

Full consistency: the predicted structures fit together in three-dimensional space.

Checking: running Rosetta, like Meiler and Baker?

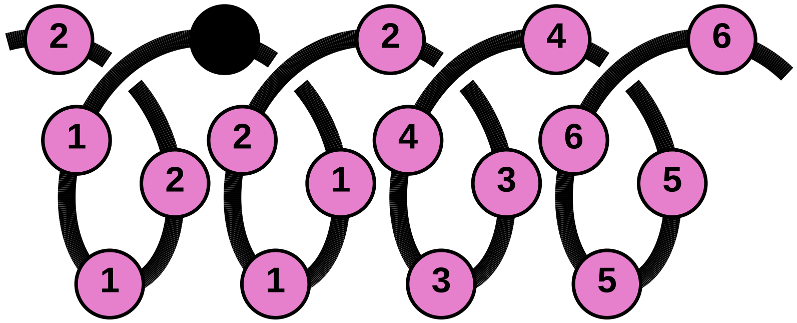
Alternative: intermediate notions of consistency.

Metric consistency: we can assume that the distance between consecutive residues on a sequence is exactly 1, plus we can make assumptions about the exact distances within α-helices and β-sheets. Such assumptions roughly corresponds to geometric facts about these structures.

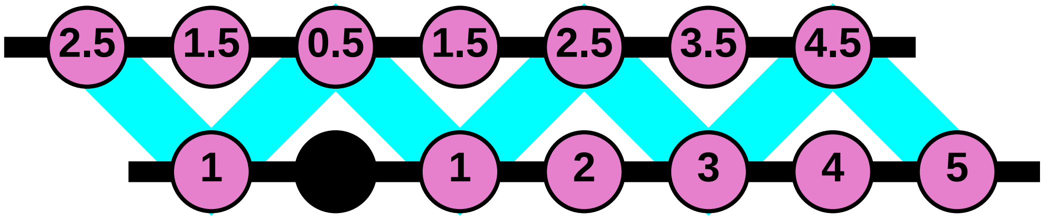
We may require that for a selected set of structures these assumptions — and the triangle inequality — do not yield a contradiction.

Why use distances that only roughly correspond to the geometric facts? We want to choose distances that impose as stringent conditions as possible, provided that this conditions are satisfied by all known protein structures.

Distances inside an α-helix (from the black residue):

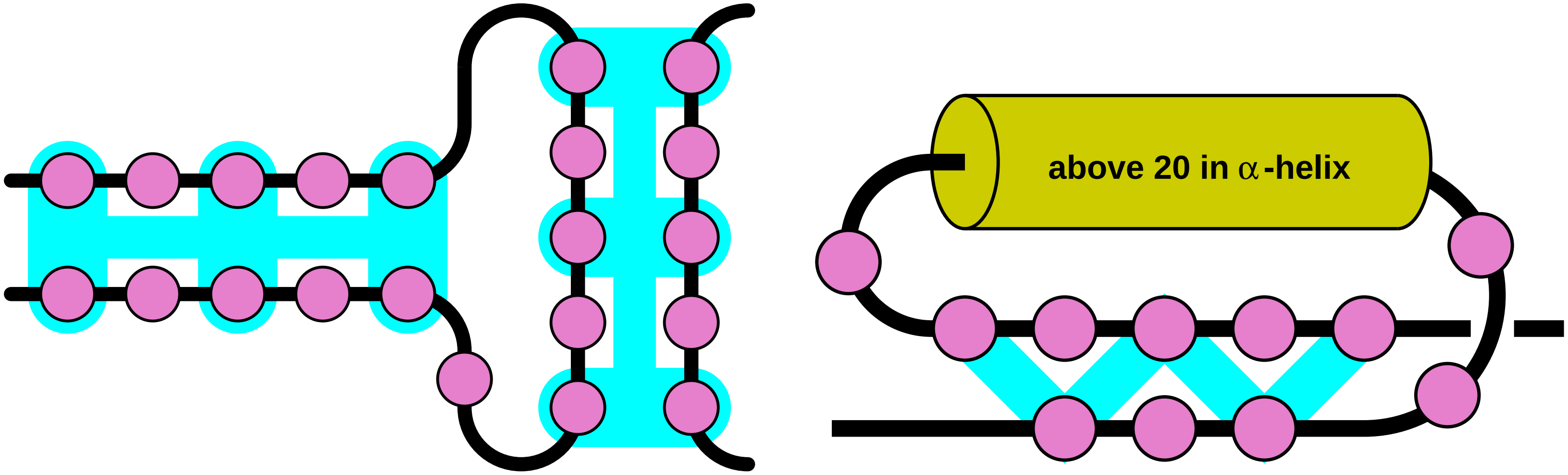


Distances inside a β-sheet (from the black residue):



Pairwise metric consistency: find set of plausible structures with maximum total weight such that their characteristic sets are disjoint and no two of them imply a metric contradiction.

Examples of metric contradictions:



Left example: in the vertical β-sheet, the distance between top and bottom residues is exactly 4 and at most 3.5.

Right example: in the α-helix, the distance between first and last residue is 10 (or more), and at most 8.5.

Good news: pairwise metric consistency defines a problem that can be approximately solved using local ratio method.

Metric consistency can be applied in other ways as well. If the number of plausible structures is not too large (50? 90?), one can apply an exact algorithm, of branch and bound type, for Maximum Weight Independent Set, and maintain the table of metric implications of currently selected structures. Increasing the number of detected conflicts improves the running time of branch and bound methods.