# IV. Protein folding and Artificial Intelligence

## Definition and mechanisms

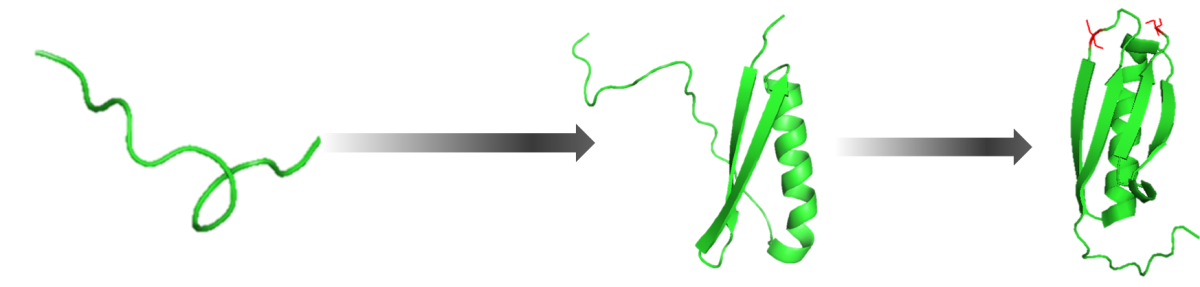
Protein folding refers to all the complex processes that take place after the amino acid sequence is linked in the cell after translation from the genetic information and by which the proteins assume their native three dimensional conformations. As it was mentioned in Chapter II, protein structure can be viewed at 4 different levels, but ultimately the spatial arrangement of the atoms that gives the molecule its shape is what defines its function in the organism.

Thermodynamically, a protein folds from a higher energy unfolded state to a lower energy folded state [3]. This process is usually a rapid one, often lasting from under one second to several seconds. The speed of folding suggests that this action takes a directed pathway rather than searching for random conformations until stumbling on the most stable structural arrangement.

Considering the large number of possible shapes for a macromolecule, it was argued that there should be pathways to simplify choices in the folding mechanism. Three mechanisms [1] were proposed, that simplified the search for the folded state:

* The **framework model** suggests that local elements of native secondary structure could form independently of tertiary structure, thus removing the stringent requirement of simultaneous formation of these two structures. The secondary structure elements would diffuse until they collided, successfully adhered and coalesced to give the tertiary structure.
* The classical **nucleation model** proposed that some neighboring amino acid residues would form native secondary structure that would act as a nucleus from which the structure would propagate in a stepwise manner. Thus, tertiary structure would form as a necessary consequence of the secondary structure.
* The **hydrophobic collapse model** hypothesized that a protein would collapse rapidly around its hydrophobic (nonpolar) side chains and rearrange from the restricted conformational space occupied by the intermediate.

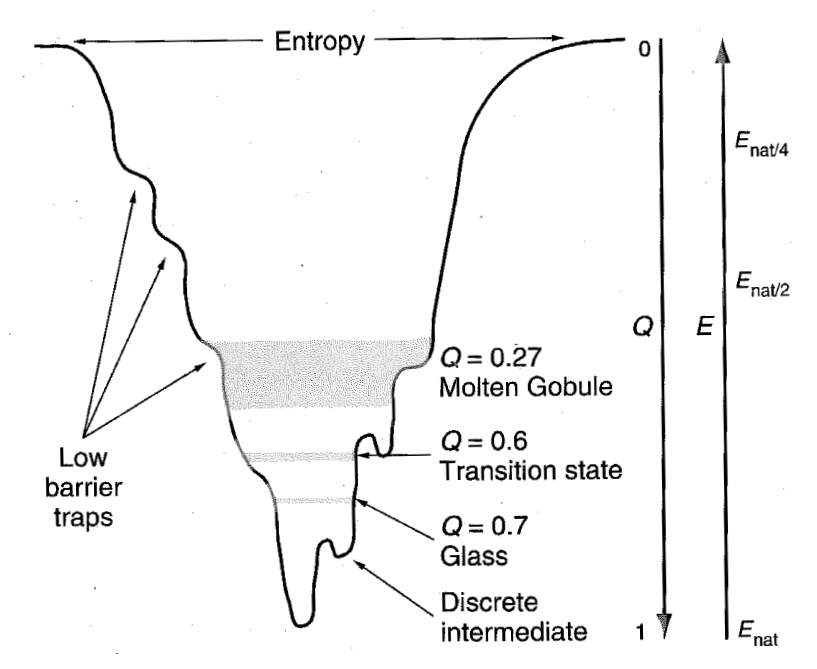
Because of the subsequent finding of so many apparent folding intermediates, it was assumed that the presence of intermediates on pathways is an essential requirement for folding. Therefore the nucleation mechanisms has fallen out of favor, as it is the only model that does not imply the existence of folding intermediates.



**Figure IV.1.** Steps a protein takes in order to assume its native three dimensional conformation [21]

Figure IV.1 illustrates the path from a linear sequence of amino acids to the native three dimensional structure of a protein, including a folding intermediary product with typical secondary structure elements (α-helix and β-sheets).

Theoreticians have compared the process of a protein falling into its native configuration to a progression down a funnel [1]. A cross section through an energetic funnel is given in Figure IV.2, where we can see that it represents a conceptual mechanism for understanding the self-organization of a protein to reach a lower free energy state. At the top of the funnel, the protein exists in a large number of random states that have high entropy. Progress down the funnel is accompanied by an increase in native-like structure as folding proceeds, such that the funnel is a progressive collection of geometrically similar collapsed structures, one of which is more thermodynamically favorable than the rest.



**Figure IV.2.** Cross section through a folding funnel, where E corresponds to the free energy of the conformation [1]

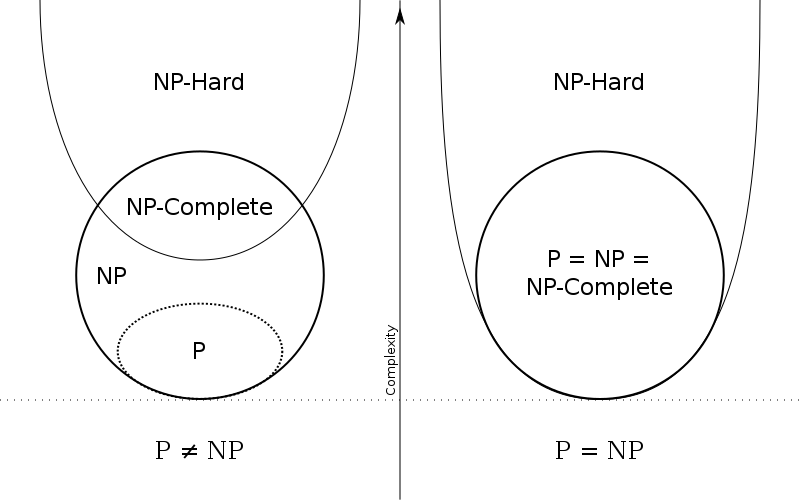
But proteins vary so much in structure, size and properties that there are bound to be many variations to these mechanisms and it is unlikely that there is a single mechanism for protein folding. Furthermore, evolution towards a specific function may be at the expense of stability or optimization of folding rate [1]. Nonetheless, understanding the mechanisms in which protein folding takes place helps us in choosing appropriate techniques for predicting protein structure, that correlate with the underlying forces that drive this process.

## NP-completeness and NP-hardness

There are three classes of problems, according to their computational complexity [22]:

* **Class P** consists of problems that are solvable in polynomial time: O(nk), for some constant k, where n is the size of the input to the problem.
* **Class NP** (nondeterministic polynomial time) includes problems that are verifiable in polynomial time, meaning that given a solution to a problem, we could verify that it is indeed correct in polynomial time with respect to the size of the input. Since we can solve any problem in P in polynomial time, without being provided a solution, any problem in P is also in NP.
* A problem belongs to the **NPC class** if it is in NP and is as “hard” as any problem in NP, referring to it as being NP-complete. If any NP-complete problem can be solved in polynomial time, then every problem in NP has a polynomial-time algorithm.

Furthermore, **NP-hard** is a class of decision problems which are at least as hard as the hardest problems in NP, but they do not necessarily have to be elements of NP.



**Figure IV.3.** Diagrams for the P, NP, NP-Complete and NP-Hard problems. The left side applies when P and NP are different and the right side is valid otherwise

It is unknown whether P = NP, but most researchers believe that P and NP are not the same class. Intuitively, the class P consists of problems that can be solved rather quickly. On the other hand, the class NP consists of problems for which a solution can be verified quickly. Considering that it is often more difficult to solve a problem from scratch than to verify if a solution is correct under some time constraints, theoretical computer scientists generally believe that NP includes problems that are not in P (left side of Figure IV.3).

Computational intractability [23] refers to the inability to construct efficient (polynomial time) algorithms that can solve a given problem, both in terms of the present state-of-the-art algorithmic research, as well as possible mathematical statement that no such algorithms exist. Usual statements about the intractability of a problem are made by showing that the problem is NP-complete, since the best known algorithm for any NP-complete problem takes an exponential number of computational steps with respect to the number of inputs, which makes these problems “practically intractable”.

Formally, NP-complete problems are decision problems, for which the answer is either yes or no. Optimizations problems like protein structure prediction are not directly considered within the framework of NP-completeness [23], but it can be transformed into a decision problem by defining a threshold with which the solution will be compared to. The corresponding optimization problem is at least as hard as the decision problem, since finding the optimal solution would answer this decision problem for every value of the threshold. Therefore, an optimization problem is NP-hard if its corresponding decision problem is shown to be NP-complete.

Following the thermodynamic hypothesis of proteins folding, computational models of protein structure prediction are typically formulated to find the global minimum of a potential energy function. Many protein folding models use lattices to describe the space of conformations that a protein can assume. Two or three dimensional lattices provide a natural discretization of the space of protein conformations, which are often viewed as a self-avoiding path in the lattice in which the vertices are labeled by amino acids. An energy value is associated with every configuration taking into account relationships between the amino acids on the lattice. But the specifics of these algorithms differed in many aspects, from the domain representation to the geometry of the lattice. The NP-completeness problem has been studied in the past considering some of these models, but results that transcend specific problem formulations are of significant interest because they may say something about the general biological problem with a higher degree of confidence.

Hart and Istrail [23] have managed to present a robust complexity analysis of a generalized lattice model, as well as general energy functions to predict protein folding. Their results suggest that the protein structure prediction problem is NP-hard for any reasonable lattice and for a class of energy formulas for which the energy monotonically increases to zero with the distance between amino acids.

But nature seems to be able to solve NP-hard problems in polynomial time, given the short duration of the entire folding process for a given protein. The exact principles and mechanisms by which it succeeds are still eluding researchers, but prediction algorithms are trying to bridge the gap between theory and nature by using the available data about protein structure to extract new information and knowledge.

## The role of CASP

CASP (Critical Assessment of methods of protein Structure Prediction) [24-26] has been monitoring the state of the art in modeling protein structure from amino acid sequence since its first round in 1994. CASP is a large-scale community experiment conducted every two years that aims to provide an independent validation benchmark for protein folding prediction.

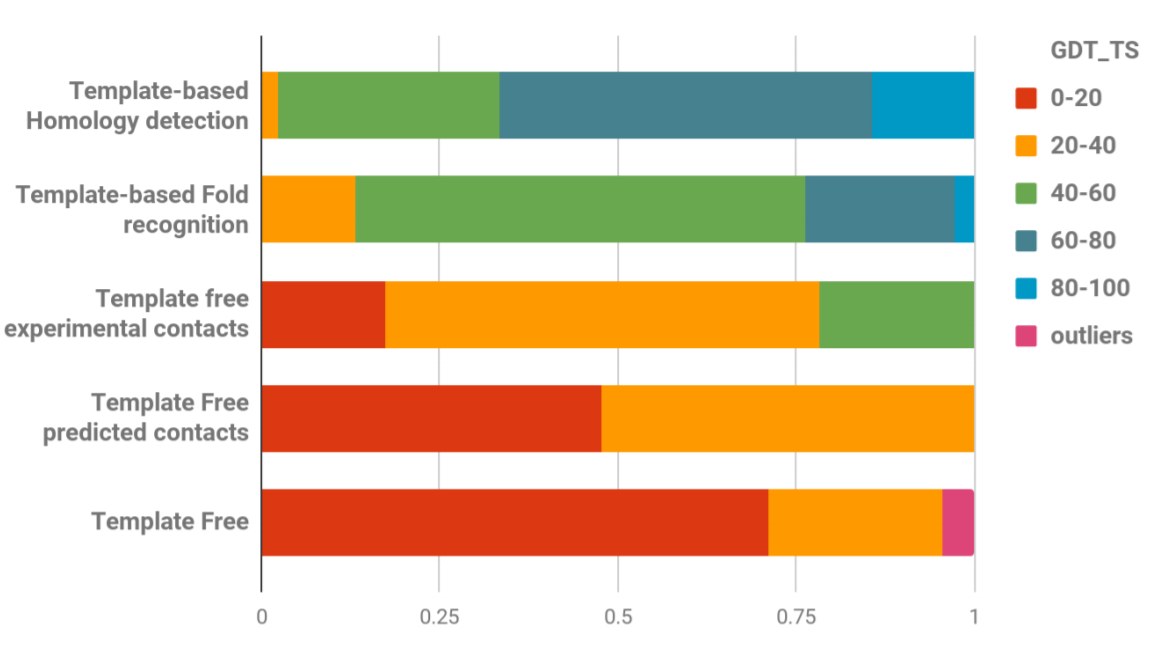
Since the first attempts until know, the problem of protein structure prediction has been claimed to be solved many times [25], only to be proven to be an ongoing struggle in the field of bioinformatics. The problem was that he algorithms used for prediction were trained using datasets that included the structures that were later evaluated. This is where the need for a standardized means of comparing different prediction tools and methods arose.

CASP is a double-blinded experiment [24] in which neither the predicting teams nor the organizers or the ones assessing the results know the structure of the target proteins at the time the predictions are made. Moreover, the independent assessors do not know the identity of the participants to ensure maximum objectivity.

Information about soon-to-be experimentally determined protein structures is collected and passed on to registered predictors from the modeling community. Research groups may participate via servers using fully automated methods or as experts, where a combination of computational methods and human expertise may be used. The structures gathered from the experimental community are called targets and the predicted conformations for a given target are called models. Expert groups are usually allowed up to three weeks to submit a model, while servers have three days.

The models are compared with the corresponding experimental structures using a range of numerical evaluation criteria and then independent assessors are asked to interpret the results and develop new measures of assessment if they see fit. The easiest way to compare the results given in terms of atom coordinates is to calculate the root-mean-square deviation (RMSD) after a structural superpositioning with the target [25]. But RMSD is overly sensitive in cases in which the model gets a loop very wrong, even though the remaining structure may be reasonably accurate. The global distance test total score (GDT\_TS) is a more robust structural similarity measure that is well defined given an alignment between two structures. The key idea is to count the number of residues that can maximally be fitted within a certain distance cutoff, expressed as a percentage.

For a typical difficult CASP target, no model comes close to the experimentally solved structure and results with a performance of DT\_TS < 20% are not an exception. Figure IV.4 shows the GDT\_TS scores for different model categories (that are discussed in the next subsection) in CASP11 where we can see that more than half of the results have a score of less than 40% (red and orange).



**Figure IV.4.** The GDT\_TS scores for different model categories in CASP11 [24]

The last round of CASP from 2016 (CASP12) gathered 34 experimental groups that provided 71 targets for assessment using methods from 8 modeling categories and almost 55 thousand models where submitted. This edition saw substantial progress in four areas, particularly in the protein contact prediction category and follows the long-term trend in CASP of increased cumulative modeling accuracy. Also, two new categories were included in response to the evolution of the field and also to encourage new directions: modeling of protein assemblies and evaluating the suitability of models for interpreting aspects of function [24].

Since 1994, CASP has continued to encourage researchers to work on better and improved methods to determine the conformation of proteins and has provided a benchmark for this dynamic bioinformatics domain. Many web-based prediction tools have been developed to participate in this competition, such as: ROSETTA, i-Tasser or Phyre2 [27-29], and are now reference points for future methods in this area.

## Types of protein structure prediction

## Methods

[21] \*\*\*, Protein folding image, Single-molecule protein dynamics, Department of Chemical Physics, Weizmann Institute of Science.

[22] Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, 3rd Edition, The MIT Press, Cambridge, Massachusetts, 2009.

[23] Hart, W.E., Istrail, S.: Robust proofs of NP-hardness for protein folding: general lattices and energy potentials, Journal of Computational Biology, 1997(4), 1-22.

[24] Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., Tramontano, A.: Critical assessment of methods of protein structure prediction (CASP) – Round XII, Proteins, 2018(86), 7-15.

[25] Feenstra, K.A., Abeln, S.: Structural Bioinformatics, Centre for Integrative Bioinformatics, Vrije Universiteit, Netherlands, 2017.

[26] Moult, J.: A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction, Current Opinion in Structural Biology, 2005(15), 285-289.

[27] Kaufmann, K.W., Lemmon, G.H., DeLuca, S.L., Sheehan, J.H., Meiler, J.: Practically useful: What the ROSETTA modeling suite can do for you, Biochemistry, 2010(49), 2987-2998.

[28] Zhang, Y.: i-Tasser server for protein 3D structure prediction, BMC Bioinformatics, 2008(9).

[29] Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., Sternberg, M.J.E.: The Phyre2 web portal for protein modeling, prediction and analysis, Nature Protocols, 2015(10), 845-858.