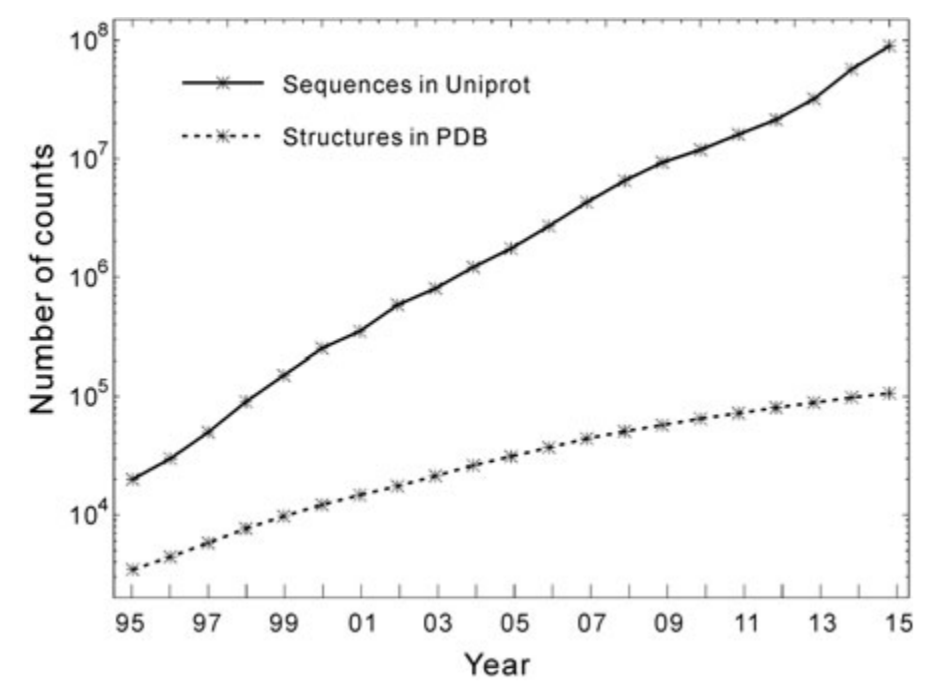
# IV. Protein folding and Artificial Intelligence

The number of known protein sequences has been increasing exponentially in the last years, mainly because of the success of an array of genome sequencing projects. But, as we have seen in Chapter II, the sequences on their own cannot distinguish the function of each protein in the cell. The speed of protein structure determination lags far behind the increase of sequences, due to the technical difficulties and laborious nature of structural biology experiments.

By the end of 2015, there were approximately 90 million protein sequences in the UniProtKB database, while the number of corresponding protein structures in the PDB was only about 100 000. The gap is rapidly widening, as it is illustrated in Figure IV.1, with a ratio of sequences over structures increasing to around 3 orders of magnitude.



**Figure IV.1.** The number of available protein sequences (in UniProt) and solved protein structures (in the PDB) from 1995 until 2015 [21]

Thus, developing efficient computer-based algorithms that can generate high-resolution three dimensional predictions becomes one of the major ways to fill this gap. In order to do that, it is also important to look at how proteins reach their native conformation and the forces that drive this process. Also, in this chapter we will analyze the computational complexity of predicting protein structure and inspect the importance of CASP (Critical Assessment of methods of protein Structure Prediction) in the development of this area of bioinformatics. Lastly, we will review the past and current methods of protein secondary and tertiary structure prediction.

## Mechanisms of protein folding

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.

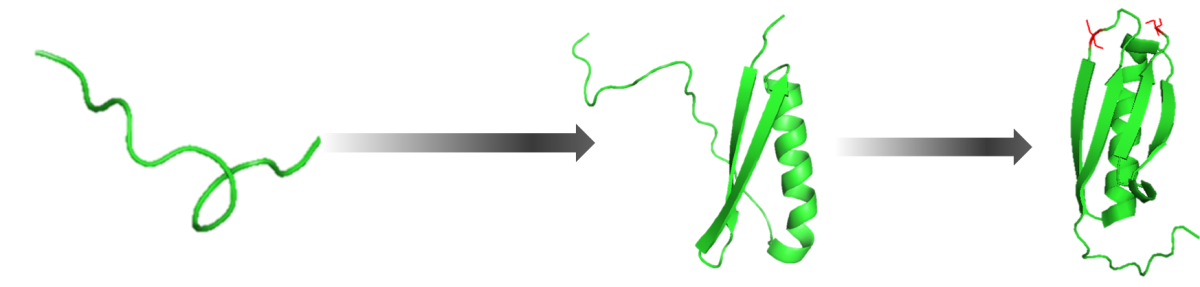
Thermodynamically, a protein folds from a higher energy unfolded state to a lower energy folded state [3]. This process is usually a rapid one, often lasting from under one second to several seconds. The speed of folding suggests that this action takes a directed pathway rather than searching for random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

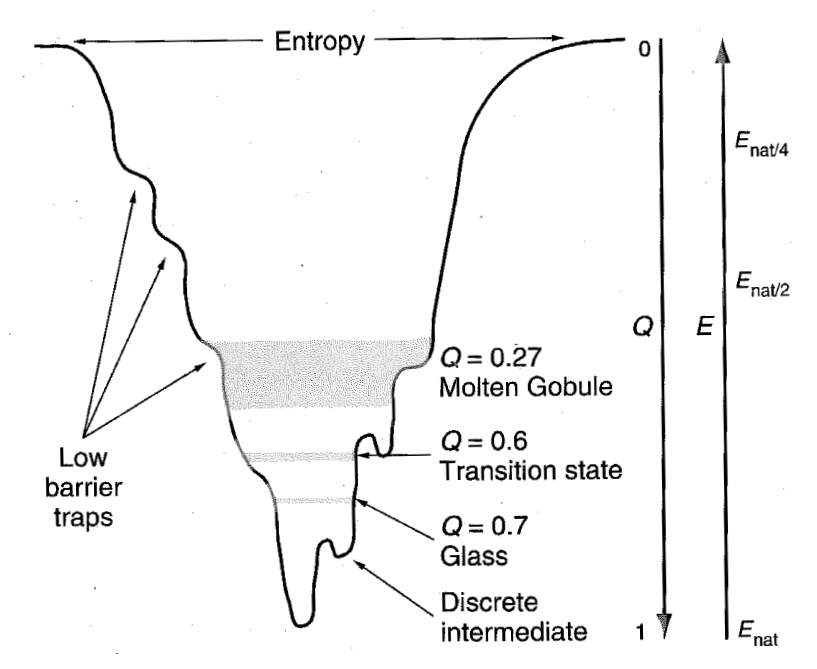
Because of the subsequent finding of so many apparent folding intermediates, it was assumed that the presence of intermediates on pathways is an essential requirement for folding. Therefore the nucleation mechanisms has fallen out of favor, as it is the only model that does not imply the existence of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

Figure IV.2 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).



**Figure IV.2.** Steps a protein takes in order to assume its native three dimensional conformation [22]

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.3, 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.3.** 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.

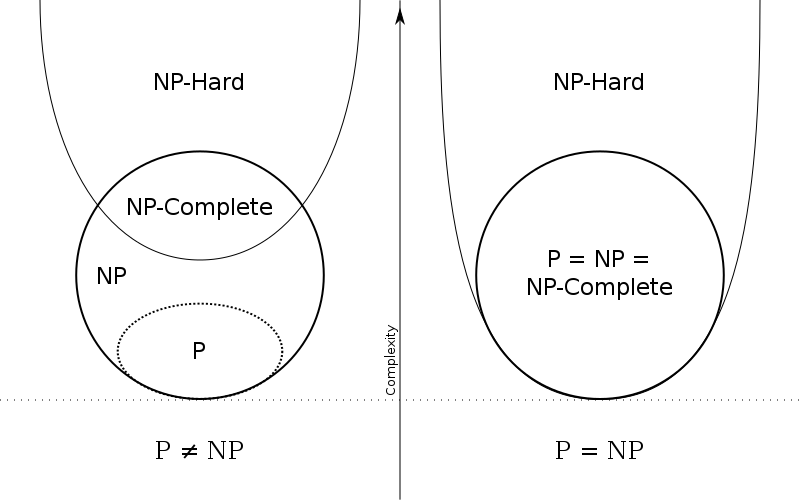
## NP-completeness and NP-hardness

Predicting the native three dimensional conformation of proteins has been a key concern in bioinformatics for many years and is still far from being solved. Therefore, it is useful to analyze the computational complexity of different prediction methods in order to explain the difficulties that limit the results in this domain.

There are three classes of problems, according to their computational complexity [23]:

* **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.4.** 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.4).

Computational intractability [24] 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 [24], 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 [24] 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. This is due to the vast conformational search space, considering that each atom in the molecule has 3 degrees of freedom and an entire macromolecule can have hundreds or thousands of degrees of freedom.

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) [25-27] 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.

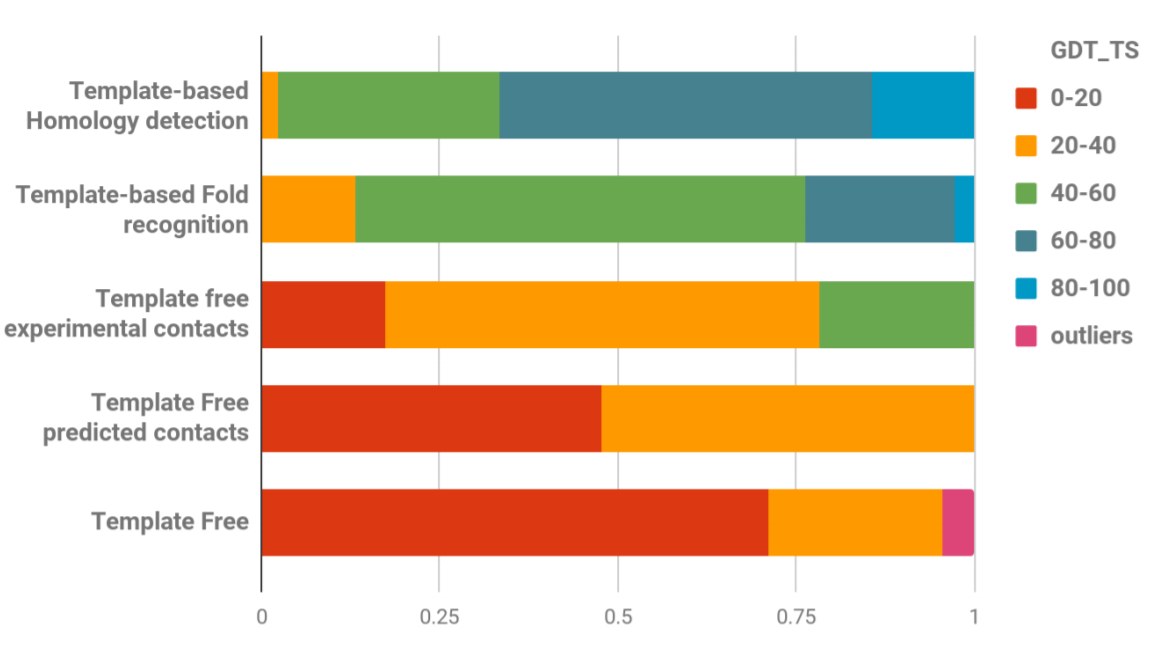
Since the first attempts until know, the problem of protein structure prediction has been claimed to be solved many times [26], only to be proven to be an ongoing struggle in the field of bioinformatics. The problem was that he algorithms used for prediction were trained using datasets that included 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 [25] 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 [26]. 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.5 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.5.** The GDT\_TS scores for different model categories in CASP11 [25]

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 [25].

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 [28-30], and are now reference points for future methods in this area.

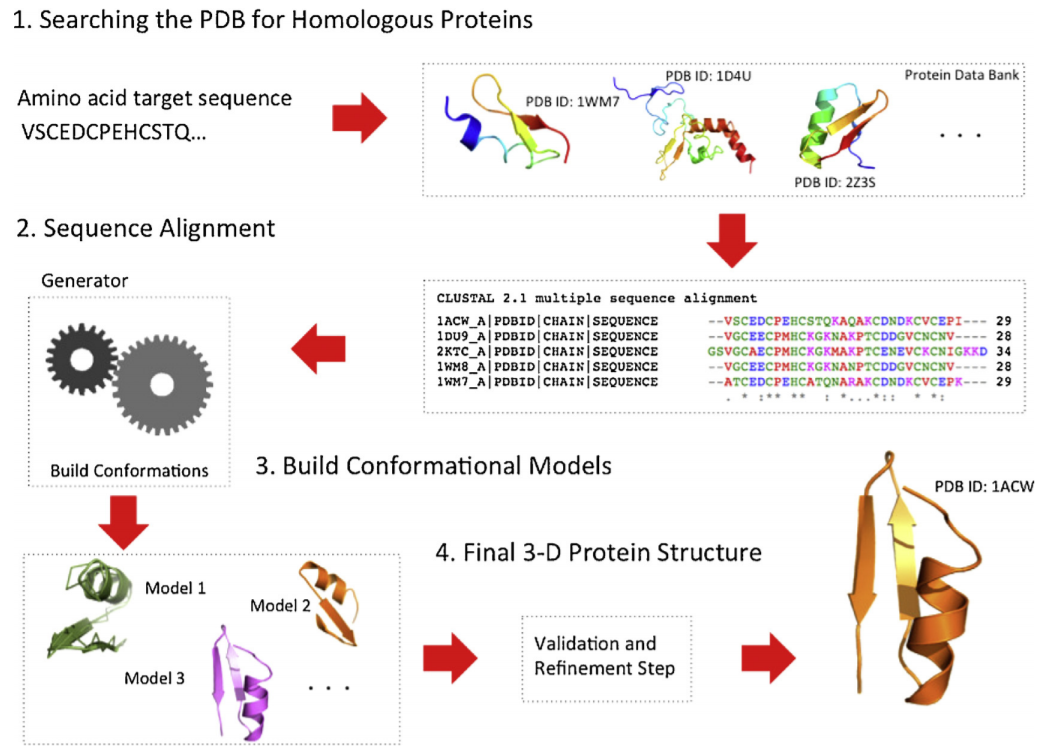
## Types of protein tertiary structure prediction

A large part of folding prediction relies on the fact that, for two homologous proteins (with similar functions), structure is more conserved than sequence [26]. When we think of it the other way around, we can deduce that if two protein sequences are similar, they are also likely to have a very similar structure. This means that if our sequence of interest is similar to a protein sequence with a known structure, we can use it as a starting point for our model. However, if a template structure is not available, models have to be constructed from scratch. Once a few models were created, we can assess which one performs better by scoring them using different quality assessment tools and in some cases apply model refinement methods [21].

### Template-based modeling

Template-based modeling (TBM) for tertiary structure prediction has been included in the CASP as a stand-alone prediction category since round VII in 2006, based on the fact that methods that use comparative modeling (either using homologous structures or fold recognition) have a higher accuracy than free modeling and could be grouped under a single name [31].

In **homology modeling**, a target sequence of amino acids is aligned against the sequence of another protein with known structure, acting as a template. The main idea is to create an atomic-resolution model of the target protein from its amino acid sequence and an experimentally determined structure of one or more related homologous proteins. Therefore, homology modeling can be applied whenever an evolutionary relationship between the target and template(s) can be detected [4]. The structures of these proteins are usually similar in the sense that amino acid residues with identical physico-chemical properties occupy the same position in homologous proteins, but also accounting for the possible additions and deletions of amino acids.



**Figure IV.6.** Schematic representation of the process of template-based modeling using homologous proteins for three-dimensional protein structure prediction [4]

The steps in constructing a three-dimensional model for a target protein is outlined in Figure IV.6, from finding homologous proteins, aligning the sequences, building the structure model and refining to best fit the target sequence [4].

The quality of homology modeling methods depends on the quality of the sequence alignment methods that compare the sequence of the target with the proteins of already known structure. There are two strategies to do this:

* Pairwise comparison, in which the target sequence is compared independently with each candidate sequence in the database. For example: FASTA, BLAST, PSI-BLAST.
* Multiple sequence comparison that performs multiple alignments to improve the sensitivity of the search. CLUSTALW, PSI-BLAST and T-COFFEE are examples of tools for multiple sequence alignment.

When building the model, usually the backbone from the homologous regions is constructed, continuing with the rest of the regions and the side chains. A variety of methods can be used to construct the structure of the target protein: segment matching, assembly of rigid bodies and modeling by satisfaction of spatial restraints. The main computational methods that use homology modeling are: SWISS-MODEL, MODELLER, ReformAlign, PyMod, and MULTALIN [4].

**Fold recognition** methods [4, 26], are motivated by the fact that proteins with no apparent sequence similarity could have similar folds. In contrast to homology-based template search, it is not strictly necessary for a target sequence and a template sequence to be homologous (evolutionary relationship or function similarity), they may have gained similar structures through convergent evolution. The library of potential folds is constructed from known native structures and the structural core elements are defined by the secondary structure elements: α-helix, β-sheet and coil, leaving a template of the backbone of the fold. A scoring schema to evaluate a particular placement of a sequence into a fold usually employs statistical references of each amino acid residue placement into a fold environment and describes how favorable a replacement of a query sequence and a template structure are. An algorithm to identify the optimal sequence –structure pairing is used next to search over the vast space of possible replacements. Some tools that employ fold recognition: GENTHREADER, 123D, ORFEUS, PROSPECT and Phyre [4].

Accuracy of protein models has increased dramatically from the early CASP experiments to the present day. It is routinely expected that a good structural model can be built for a target sharing more than 20% of sequence with at least one known protein structure, while cases where good models are built at lower sequence similarity are not unusual. Template-based modeling is currently the most reliable type of protein structure prediction [32]. Since the number of different protein folds is estimated to be limited ad fold coverage increases with the growth of protein structure database, the applicability of TBM is ever growing.

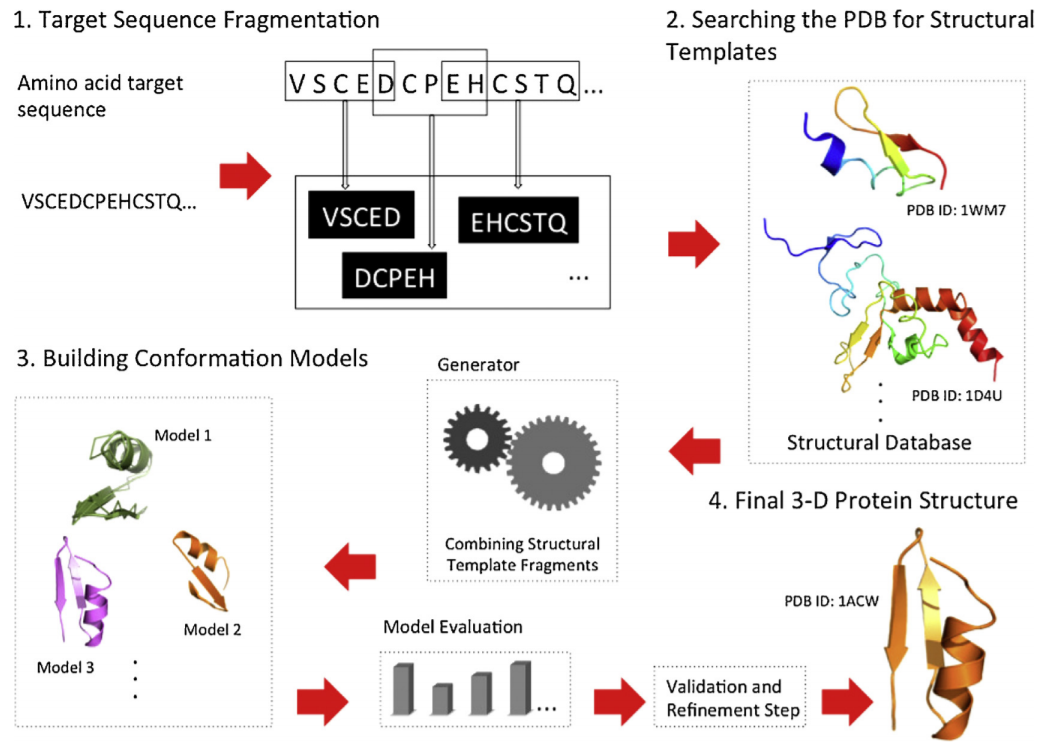
### Template-free modeling

Template-free modeling (or free modeling – FM) aims to predict tertiary structure without the use of a protein template, particularly when no suitable template is available [4]. In this case, there are two possible procedures: to try and predict the three dimensional structure without the use of any database information or to incorporate knowledge about the structure of small protein fragments (similar to template-based modeling but at a smaller scale).

FM without database knowledge also called *de novo* or ***ab initio* modeling** is based on the thermodynamical principles and the fact that the native structures of proteins correspond to the global minimum of its free energy, as it was discussed at the beginning of this chapter. *Ab initio* protein folding is considered a global optimization problem where the goal is to identify the positions of all atoms or a specific set of atoms in the protein structure that describe the minimum energy of the polypeptide conformation. They simulate the protein conformational space using an energy function, which describes the internal energy of the protein and its interactions with the environment. Ab initio methods, as opposed to TBM, can predict new folds because they are not limited to templates from the PDB. In general, this strategy requires the use of a geometric representation of the protein chain, a potential function and an energy surface searching technique. The most common tools used in FM without database information are: AMBER, CHARMM, UNRES and TINKER [4].

**The fragment-based FM methods** [4, 26] do not compare a target sequence to a known structure, but they compare fragments, short amino acid sub-sequences of a target against fragments of known protein structures. The general steps involved in the obtaining of a three dimensional structure using this strategy are review in Figure IV.7.

This procedure has its roots in the observation that when a new fold is discovered, it is composed of common structural motifs or fragments from secondary structures of proteins with known structures. When homologue fragments are identified, they are assembled into a structure through scoring functions and optimization algorithms. The fragments are assembled through a fragment assembly procedure with the purpose of finding the structure with the lowest energy potential, similar with *ab initio* methods. But in comparison with the previous FM methods, fragment-based methods take advantage of the reduction in the conformational search space given by the use of 3-9 amino acid long fragments. It is also important to note that because they do not rely only on physical principles and physico-chemical properties, such models are likely to share any of the biases that are present in the PDB. The most common tools that use the fragment-based strategy in predicting protein conformations are: i-Tasser, ROSETTA, FRAGFOLD and CABS-Fold [4]. Also, fragment-based methods produced very positive results in the CASP experiments.



**Figure IV.7.** Schematic representation of a fragment-based FM method for the prediction of protein tertiary structure [4]

### Refinement methods

Refinement methods [26, 33] refer to the improvement of the predicted model using different techniques and is included as a major step in both of the previous structure modeling techniques, as shown in Figures IV.6 and IV.7. But for many years, model refinement was not taken into consideration in the CASP experiments (until the eighth round). For the last two rounds in 2014 and 2016, the category of mode refinement has seen considerable improvement, with an enhancement of 3-5% over 70% of models. A broad variety of methods have been used in proteins structure refinement, ranging from knowledge-based and fragment based approaches to molecular dynamics with physics-based force fields. While some methods are relatively more conservative, providing a reliable but small refinement, other approaches are more adventurous providing significant improvement of the global and local structure for some targets while making a few others worse. Nonetheless, coupling refinement methods with TBM of FM has been shown to improve prediction accuracy.

## Methods

[21] Rigden, D.J. (editor): From protein structure to function with bioinformatics, Springer, London, 2017.

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

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

[24] 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.

[25] 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.

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

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

[28] 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.

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

[30] 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.

[31] Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., Hubbard, T., Tramontano, A.: Critical assessment of methods of protein structure prediction-Round VII, Proteins, 2007(69), 3-9.

[32] Kryshtafovych, A., Monastyrskyy, G., Fidelis, K., Moult, J., Schwede, T., Tramontano, A.: Evaluation of the template-based modeling in CASP12, Proteins, 2018(86), 321-334.

[33] Hovan, L., Oleinikovas, V., Yalinca, H., Kryshtafovych, A., Saladino, G., Gervasio, F.L.: Assessment of the model refinement category in CASP12, Proteins, 2018(86), 152-167.