Contents

I. Introduction

II. Proteins

1. Biological importance

Proteins are the major functional macromolecules of life [1] whose properties recommend them as therapeutic agents, catalysts, vaccines and materials. Among some of their important functions within organisms are: catalyzing metabolic reactions, intracellular molecular transporting, cell signaling and DNA replication. An alteration in any of these functions can lead to major negative consequences to the overall health of the organism.

Mutations in proteins can cause them to lose their function and are the source of many diseases. In some cases, metabolic pathways can be affected by the impaired catalytic activity of a particular protein. In other cases, when structural properties are altered, the loss of a physical function can be experienced. Some misfolded proteins, called infectious prions, can cause normal folded proteins to also become misfolded and can damage neurons, giving the affected brain a spongiform appearance. In a similar way, diseases can stem from proteins that gradually precipitate to form fibrils, long chains of polymerized sheets, in a process called amyloidosis. Approximately 50% of human cancers are caused by mutations that lower the stability of a protein that usually has the role to suppress the formation of tumors. In order to restore function or to destroy pathogens or cancers, current therapeutic agents target enzymes and receptors, two different types of proteins with respect to their function [1].

The properties and functions of cells and organisms are determined to a great extent by the proteins that they are able to make. Although the functions of proteins inside the cell are vast and diverse, their common mechanism of action is to bind to a substrate and act upon this interaction. This binding always shows great specificity, meaning that a protein can usually recognize just one or a few molecules out of many thousands that it encounters. This happens because the binding site of the protein has a three-dimensional structure that only matches a specific substrate, like a lock and key. If only a minor change occurs in the amino acid sequence of the protein, this binding site can have a totally different shape and the binding would not be possible [2].

This is one of the most important reasons why the study of protein structure is such an intensely researched domain in the field of bioinformatics and artificial intelligence. If we would know the three-dimensional structure of a protein that we want to target, we could design a molecule that perfectly fits inside its active site and purposefully interacts with the protein, either by enhancing its function or by blocking it, thus restoring the health of the organism. On the other hand, if we would want to act on a specific molecule in the organism that contributes to a disease, we could trace out a protein structure that would attach specifically to that molecule and then synthesize it using currently available techniques.

2. Protein Biosynthesis

The human genome has first been completely sequenced in 2001 and has been shown to include approximately 21 000 protein-encoding genes which give rise to a much greater number of distinct proteins, but this accounts for only about 1.5% of the total amount of DNA in a human cell [2]. The remaining is considered to be non-coding, regulatory DNA or sequences with functions not yet determined.

The order in which the amino acids are linked to form a specific protein is determined by the sequence of a corresponding gene. The mechanism [2] by which this process takes place has been shown to be universal in all species and it occurs in all living cells. It involves two stages that take place in two different regions of the cell.

The first step in producing proteins occurs in the nucleus of the cell, where the information from the DNA is transferred to another type of molecule capable of holding genetic data, the RNA. This process is called transcription and results in an intermediary product that is able to exit the nucleus and carry the information to the ribosomes, where the second stage takes place. The ribosomes are small structures in the cytoplasm of the cell that read the strand of RNA and produce the proteins by linking specific amino acids together, according to some particular rules. This process is also called translation, because it basically decodes the information from the 4-nucleotides alphabet of the RNA into the 20-amino acids alphabet of the proteins [2]. The entire operation is summarized in the Figure II.1.

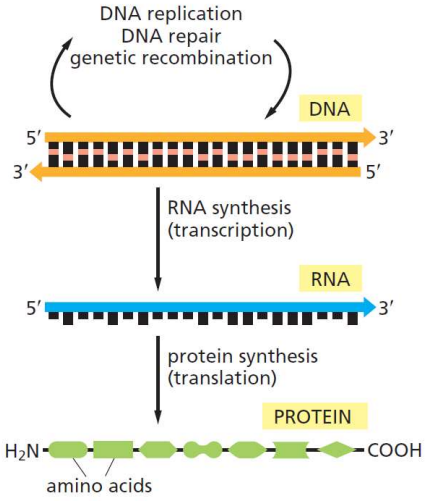

Figure II.1. The flow of genetic information from DNA to RNA and proteins [2]

Since it is obvious that the translation from nucleotides to amino acids cannot be accounted for by a direct one-to-one correspondence, the scientists tried to group together the nucleotides in order to try to solve this genetic code. It was shown in the early 1960s that a sequence of three consecutive nucleotides was able to represent one amino-acid, each group

being called a codon. Since there were four different nucleotides in the RNA, there were 4^3=64 possible combinations. With only 20 amino acids found in the structure of proteins, it was determined that some combinations are redundant and code the same amino acid. The fantastic feature of this genetic code is its universality, as it is applicable in every cell of every living organism [2].

3. Protein Structure

As stated earlier, all proteins contain a linear sequence of amino acids, molecules that contain two types of functional groups: carboxyl group (-COOH) and amino group ($-NH_2$). Each amino acid is linked with the next one by a peptide bond (-CO-NH-) between its carboxyl group and the amino group of the next molecule, giving the main protein two distinct ends: N-terminal end, with the free amino residue, and C-terminal end, with the last carboxyl residue. This is important because the counting of amino acids always starts from the N terminus [3].

Proteins contain an array of 20 different amino acids, listed in Table II.1, along with their abbreviations and the polarity of the side chains. There are an equal number of both polar (hydrophilic) and nonpolar (hydrophobic) molecules, a property that greatly affects the way in which the protein's three-dimensional shape will look like [2].

Table II.1. The 20 amino acids commonly found in proteins [2]

Polarity	Amino acid	Abbreviation (three letters)	Abbreviation (one letter)	Type of side chain
Polar	Aspartic acid	Asp	D	Negatively charged
	Glutamic acid	Glu	E	Negatively charged
	Arginine	Arg	R	Positively charged
	Lysine	Lys	K	Positively charged
	Histidine	His	H	Positively charged
	Asparagine	Asn	N	Uncharged polar
	Glutamine	Gln	Q	Uncharged polar
	Serine	Ser	S	Uncharged polar
	Threonine	Thr	T	Uncharged polar
	Tyrosine	Tyr	Y	Uncharged polar
Nonpolar	Alanine	Ala	A	Nonpolar
	Glycine	Gly	G	Nonpolar
	Valine	Val	V	Nonpolar
	Leucine	Leu	L	Nonpolar
	Isoleucine	Ile	I	Nonpolar
	Proline	Pro	P	Nonpolar
	Phenylalanine	Phe	F	Nonpolar
	Methionine	Met	M	Nonpolar
	Tryptophan	Trp	W	Nonpolar
	Cysteine	Cys	C	Nonpolar

The folding of a protein chain is also determined by many other interactions between residues from different regions. In Figure II.2 we have the amino acid sequence of the enzyme chymotrypsin, using the one-letter abbreviations from Table II.1. The enzyme is originally synthesized as a long polypeptide chain, but after the formation of the disulfide bridges between different cysteine residues, the initial chain is cleaved in three different pieces. We can see here the importance of these weaker interactions to the overall structure of the molecule [1].

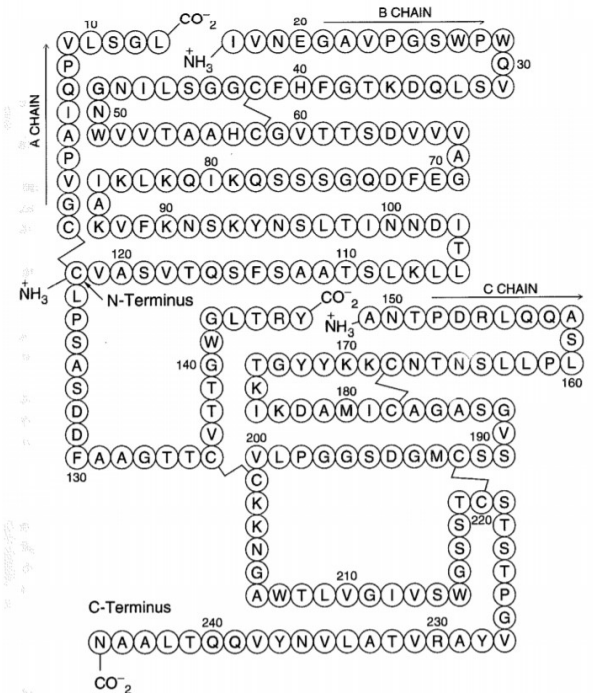

Figure II.2. Amino acid sequence of the enzyme chymotrypsin, consisting of three chains linked by weaker bonds [1]

Biologists have studied protein folding in a test tube using highly purified molecules and have found that adding certain solvents that disrupt the interactions between amino acids makes the protein unfold and converts it to a flexible polypeptide chain, losing its conformation. But when removing the solvent, the protein often refolds spontaneously into its original conformation, meaning that the amino acid sequence holds all of the information needed for specifying the three-dimensional shape of a protein [2]. The final folded conformation of any protein chain is generally one that minimizes its free energy.

3.1 Structure levels

Proteins can be analyzed at four levels:

- primary structure
- secondary structure
- tertiary structure
- quaternary structure

This hierarchy [4] facilitates the description and the understanding of proteins and it does not aim to precisely describe the laws that produce protein structures. It is an abstraction that intends to make the study of protein structures more manageable.

The **primary structure** describes the sequence of amino acids in a linear order, starting with the N-terminal region of the protein chain. **Secondary structure** can be described as the local spatial conformation of a polypeptide backbone, excluding the constituent amino acids' side chains. The major elements of the secondary structure are the α-helix and the β-sheet, with some regions of disorganized amino acids. The **tertiary structure** refers to the distribution of secondary structures in a three-dimensional space and is greatly influenced by weaker forces and interactions between side chains or with the surrounding medium. The **quaternary structure** refers to the overall spatial arrangement of polypeptide subunits within a protein composed of two or more polypeptide chains [3, 4].

Although the overall conformation of each protein is unique, when we compare the three-dimensional structures of many protein molecules, two regular folding patterns are often found within them. Both patterns were discovered more than 60 years ago from studies of hair and silk and are particularly common because they involve hydrogen bonds only between the atoms in the polypeptide backbone, and not those in the amino acid side chains. In each case, the protein chain adopts a regular, repeating conformation [2, 5]. These two secondary structure elements are commonly formed because they maximize formation of stabilizing intramolecular bonds and minimize repulsion between adjacent side chain groups, while also being compatible with the rigid nature of the peptide bonds [3].

3.2 The α-helix

The first folding pattern was called the **α-helix** and was identified in the protein α-keratin, which can be found in large quantity in the skin, hair and nails. This was able to explain the strength and elasticity of this protein and account for the fiber appearance at the X-ray diffraction. An α-helix is generated when a single polypeptide chain twists around itself to form a rigid cylinder, with a hydrogen bond between every fourth peptide bond and the amino acid side chains protruding outward from the helical backbone [2]. This gives rise to a regular helix with a complete turn every 3.6 amino acids, as can be seen in Figure II.3.

Stretches of α-helix can vary in length from one single helical turn to more than 10 consecutive turns, with the average length being of about three turns, in globular proteins [3]. The proteins located in the cell membrane, having transport and receptor functions, contain extensive regions of α-helix. Those portions of proteins that cross the membrane usually do so as α-helices composed of amino acids with nonpolar side chains. The hydrophilic polypeptide backbone is therefore shielded from the hydrophobic environment of the membrane by its protruding nonpolar side chains [2].

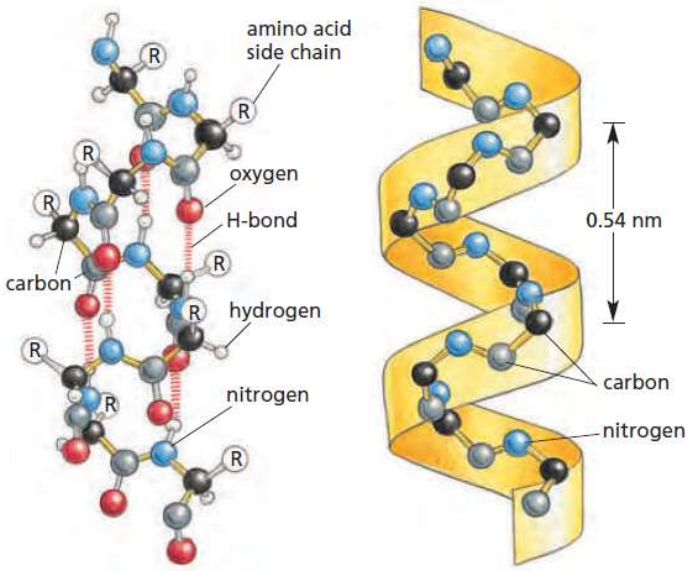

Figure II.3 The regular conformation of the polypeptide backbone in the α-helix [2]

3.3 The β-sheet

The other major structural element found in globular proteins is the **β-sheet** and it was first observed in the β form of keratin fibers. Although it was discovered a year after the first element, an approximate understanding of its molecular structure was achieved earlier than for the α structure [5]. The cores of many proteins contain extensive regions of β-sheet, which can be formed from neighboring sections of the polypeptide backbone that run in the same direction (parallel chains) or from a polypeptide backbone that folds back on itself, with each section running in the opposite direction to the one next to it (antiparallel chains), as in the Figure II.4 below [2]. Both types build a very rigid structure held together by bonds between neighboring chains.

Although these are the two major secondary structures that can be identified when looking at protein conformations, most proteins consist of several segments of α-helix and/or β-sheets separated from each other by various loop regions, or coils. These regions can vary in shape and length and allow the overall molecule to fold into a compact tertiary structure [3]. Beside their role in connecting regular secondary elements, loop regions often contribute directly to the biological function of the protein and are exposed to solvent.

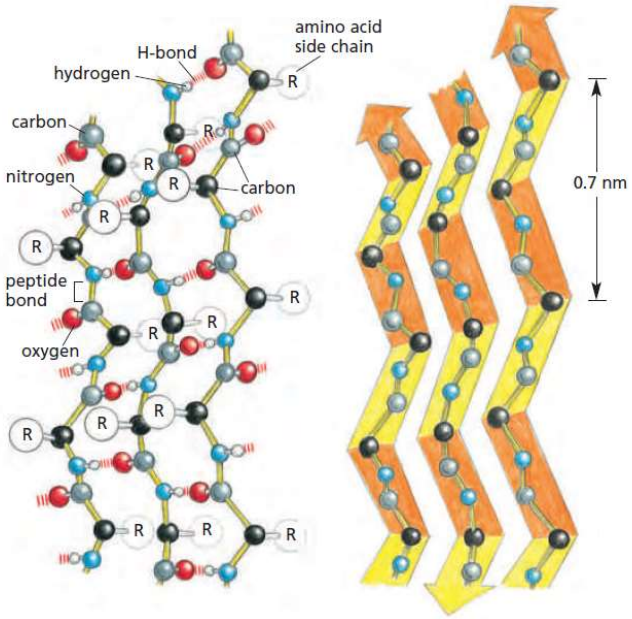

Figure II.4 The regular conformation of the polypeptide backbone in a β-sheet [2]

The major driving force for the folding of proteins seems to be hiding and clustering of hydrophobic side chains to minimize their contact with the water around the macromolecule. The basic requirements [1] for folding are that (1) the resulting structures are compact and so they have minimal hydrophobic areas in contact with solvent and that (2) the hidden groups that are bound by hydrogen bonds are all paired. The formation of the two secondary structures helps with the second point of the previous statement, as it maximizes the pairing of the hydrogen bonding groups. The helices and sheets pack by stacking their amino acid side chains.

4. Protein Structure Online Databases

To facilitate the understanding of, and access to the information available for protein structures, researchers have been gathering and structuring it in online databases, making the data easier to be queried and organized. In order to determine the unique primary structure through quaternary structure of a protein, different physico-chemical methods are employed, such as: X-ray crystallography, NMR spectroscopy or 3D electron microscopy [6].

Fifteen years after the determination of the first protein crystal structure corresponding to myoglobin, the **Protein Data Bank** (PDB) was created in 1971 [6] and initially contained only seven protein structures. The PDB currently archives approximately 130 000 entries and is managed by the Worldwide Protein Data Bank, which contributed to the evolution as the single global archive of macromolecular structure data. But in the first 30 years of its existence, the addition of new molecules was sparse and only by the mid-1990s a boost in the number of entries has been seen, as pictured in Figure II.5. This can be attributed to the advances in computer and information technology, which provided the much required computer power for

experiment automation, to the introduction of genetic engineering for easy production of basically any protein using bacterial cells and also to the development of powerful X-ray sources.

The RCSB PDB is the US regional center of the PDB and manages the website (rcsb.org) which offers multiple tools for structure query, browsing, analysis and molecule visualization. It enables users to perform simple searches based on PDB ID, name of the macromolecule, sequence or ligand, but also allows them to build complex search combinations of parameters and criteria. The PDB data is organized in hierarchical trees using external classification and annotation systems and visualization options enable the exploration of three-dimensional structure, structure/sequence information and correspondence between the two [8].

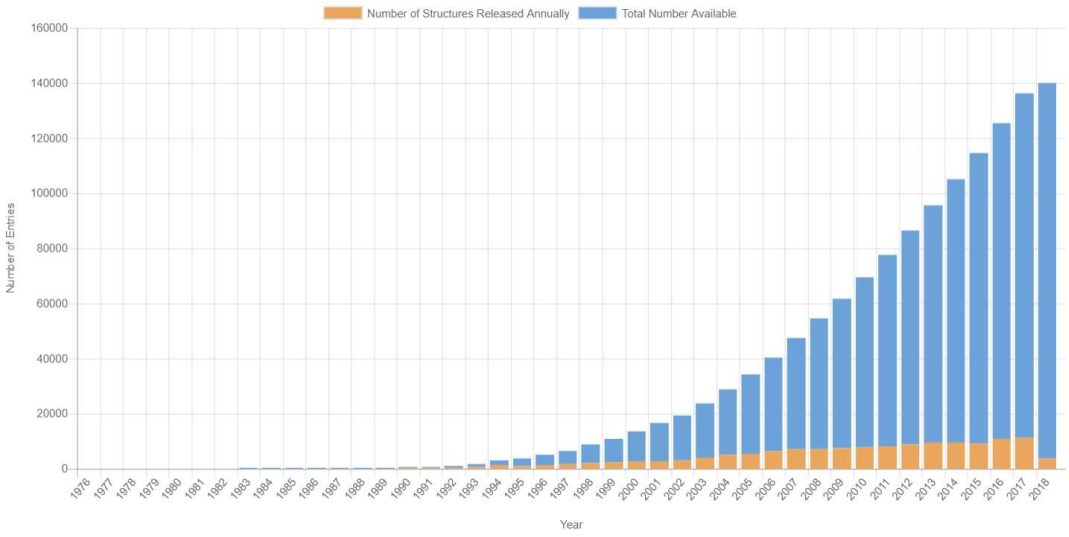

Figure II.5. Overall growth of released structures per year for the PDB [7]

Besides the PDB, there are many repositories and databases [6] used in structural biology, chemistry, life sciences and pharmaceutical industry, where they are crucial in the drug discovery process. The growing number of macromolecular structures in the PDB provides a solid foundation and increases the scientific potential of derivative data resources. Fold classification databases such as **CATH** and **SCOP** (Structural Classification of Proteins) aim to classify protein folds in terms of evolutionary relationships as well as sequence similarity, and are references for nonredundant folds and domains used by many structural bioinformatic tools. There are also other specialized data resources [6] that catalog and classify different structural aspects: the **Protein Data Bank of Transmembrane Proteins** (PDBTM), the **KnotProt** database (contains three-dimensional structures of proteins that form knots), **MPStruc** (the database of Membrane Proteins of Known 3D structure).

III. Artificial Intelligence

Artificial Intelligence (AI) is one of the newest fields in science and engineering and currently covers a huge variety of subfields, from the more general, as learning and perception, to the specific, such as playing chess, proving mathematical theorems, driving a car and diagnosing disease. AI is truly a universal field that aims not just to understand but also to build intelligent entities [9].

1. The foundations of AI

The beginnings of AI can be traced to philosophy and fiction, while early inventions in electronics, engineering and many other disciplines have greatly influenced the path of AI. Some early milestones include work in problem solving, including basic work in learning, knowledge representation and inference as well as programs in language understanding, translation, theorem proving, associative memory and knowledge-based systems [10].

AI sits at the intersection of a number of important disciplines, listed in Table III.1 below, each of them contributing in some way to the development of this field. In its formative years, AI was influenced by ideas from many fields of study. These came from people working in engineering(such as Wiener's work in cybernetics), biology(Ashby, McCulloch and Pitt's work on neural networks in simple organisms), experimental psychology, communication theory, game theory(notably by von Neumann and Morgenstern), mathematics and statistics, logic and philosophy(for example, Church and Hempel) and linguistics(such as Chomsky's work in grammar) [10].

Table III.1. The disciplines and the personalities that lead to the development of AI by finding answers to important questions [9]

Discipline	Questions	Personalities
Philosophy	• Can formal rules be used to draw valid conclusions? • How does the mind arise from a physical brain? • Where does knowledge come from?	Wilhelm Leibniz René Descartes Rudolf Carnap
Mathematics	• What are the formal rules to draw valid conclusions? • What can be computed? • How do we reason with uncertain information?	George Boole Kurt Gödel Alan Turing Steven Cook Thomas Bayes
Economics	• How should we make decisions so as to maximize payoff? • How should we do this when the payoff may be far in the future?	John von Neumann Richard Bellman Herbert Simon
Neuroscience	• How do brains process information?	Hans Berger, Camillo Golgi, Santiago Ramon y Cajal

Psychology	• How do humans and animals think and act?	H. Helmholtz, F. Bartlett, K. Craik, N. Chomsky
Computer Engineering	• How can we build an efficient computer?	J. Eckert, C. Babbage, J.M. Jacquard
Control theory and cybernetics	• How can artifacts operate under their own control?	N. Wiener, W.R. Ashby
Linguistics	• How does language relate the thought?	B.F. Skinner, N. Chomsky

These areas made their mark and continue to influence this field of study, but after having assimilated much, AI has grown beyond them and has, in turn, occasionally influenced them back [10]. Only in the last half century computational devices and programming languages have become sufficiently powerful to build experimental tests of ideas about what intelligence is.

2. Brief history

The first work that is now seen as belonging to AI was done by McCulloch and Pitt in 1943 and proposed a model of artificial neurons, drawing knowledge from three different sources: the basic function and physiology of neurons in the brain, a formal analysis of propositional logic and Turing‘s theory of computation. Their network of connected neurons was able to compute any computable function and could also implement all the logical connectives [9].

But the birth of AI is considered to have taken place in **1956** at the Dartmouth College in Hanover, where a two-month workshop gathered 10 scientists interested in the automata theory, neural nets and the study of intelligence from all over the US, in an attempt “to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans and improve themselves” [9].

Although the workshop itself did not lead to any new breakthroughs, it succeeded in introducing all the major figures involved in the discipline to each other. For the next 20 years, the field would be dominated by these people and their students and colleagues at major universities and study groups in the US [9].

The early years (**1952-1969**) of AI were full of successes, even though in a limited way. Taking into account the primitive computers and programming tools of the time, whenever a computer did something even remotely clever it was considered astonishing. Some accomplishments from this period are:

- the General Problem Solver (GPS) of Newell and Simon, probably the first program to incorporate the “thinking humanly” approach and could handle a limited class of puzzles
- the Geometry Theorem Prover of Gelernter, which was able to prove theorems that were considered tricky by many mathematics students

- the definition of the high-level language Lisp by McCarthy, which would become the dominant AI programming language for the next 30 years
- perceptrons and flourishing work on neural networks

Although these years where full of successes and enthusiasm was high, the period between **1966 and 1973** [9] was marked by a dose of reality. The predictions stated by many scientists did come true, but it took 40 years for this to happen, rather than 10. This overconfidence came from the fact that the early AI systems showed promising performance, but failed to take into account three major difficulties:

- The programs succeeded only by means of simple syntactic manipulations and knew nothing of their subject matter. An example of a failed project because of this aspect would be the efforts of early machine translation, when it was thought that simple syntactic transformations and word replacements would suffice to preserve the meaning of a sentence.
- The combinatorial explosion. It was thought at the time, before the theory of computational complexity was developed, that scaling up to more difficult tasks would be a matter of faster hardware and larger memories, but this assumption was soon proven wrong, when researchers failed to prove theorems involving more than a few dozen facts.
- The basic structures used to generate intelligent behavior had some fundamental limitations. For example, the perceptrons, although they were shown to be capable to learn anything that they could represent, they in fact could represent very little.

Until 1969, the problem solving techniques employed were using a general-purpose search mechanism attempting to put together elementary reasoning steps to find complete solutions, and they weren't able to scale up to larger or more difficult problems. The alternative was to build more powerful, domain-specific knowledge that would allow larger reasoning steps and could easily handle typically occurring cases in narrow areas of expertise. The **decade after 1969** [9] was marked by the emergence of projects that did just that, such as:

- DENDRAL – it was the first successful knowledge-intensive system and was used to solve the problem of inferring molecular structure from the information provided by a mass spectrometer. The first naïve version generated all possible structures for the given formula, predicted the spectrum that would be observed for each one and then compared these results with the actual spectrum of the molecule, but couldn't manage even moderate-sized molecules. So the researchers consulted analytical chemists and all the relevant theoretical knowledge gathered from them was mapped into rules that helped in restricting the search space.

- HPP – the Heuristic Programming Project was developed to investigate the extent to which the new methodology of expert systems could be applied to other areas of human expertise.
- MYCIN – was developed to aid in the diagnosis of blood infections. It had 450 rules acquired from extensive interviewing of medical experts, took into account the uncertainty associated with medical knowledge and was able to perform as well as some specialists.
- SHRDLU – a system for understanding natural language which was able to overcome ambiguity and understand pronoun references.
- Prolog – logic based reasoning language widely used in Europe at the time.

Since **1980** [9], AI has become an industry, with the first successful commercial expert system, R1, being employed at the Digital Equipment Corporation to help configure orders for new computer systems and saved the company an estimated $40 million a year. Also, in the mid 1980s, the back-propagation learning algorithm gained the spotlight and was applied to many learning problems in computer science and psychology. The content and methodology of work in AI has seen a revolution in recent years and is more common to build on existing theories than to propose new ones, to base claims on rigorous theorems or experimental evidence rather than on intuition and to show relevance to real-world applications.

Up until the years **2000s** [9], the emphasis in computer science has been on the algorithm, but recent work in AI suggests that for many problems, it is better to focus in the data and be less meticulous about what algorithm to apply, also taking into consideration the increasing availability of very large data sources. This suggests that the problem of how to express all the knowledge that a system needs may be solved by learning methods, rather than hard coded rules, provided that the learning algorithms have sufficient data to work with.

3. Domains of application

The multidisciplinary trait of AI can also be observed in the number of fields to which AI has contributed to, not only in the ones from which it originated. Although initially the research was much narrower, considering the multitude of areas in which AI has been proven useful until now, AI has been able to gain popularity thanks to its very efficient and general techniques. They allowed the methods to be easily adapted to different data and representations, from the financial field, to healthcare and robotics.

Some of the most notable examples of projects that incorporate AI method currently in use today are listed in Table III.2, along with their corresponding domain. Some projects include not just only one technique, but make use of AI for a multitude of tasks, such as the humanoid robot Sophia, created by Hanson Robotics. Sophia uses facial and speech recognition, imitates human gestures and facial expressions and is able to maintain a conversation [11].

Table III.2. Some of the more prominent domains in which AI is currently being applied in and a few corresponding examples of AI projects [9, 12-14]

Domains	Examples
Automotive	STANLEY, a driverless robotic car equipped with cameras, radar, sensors and an onboard software to command the steering, braking and acceleration won the DARPA Grand Challenge in 2005. Today there are more than 30 companies using AI to develop driverless cars.
Games	Deep Blue became the first computer program to defeat the world champion in a chess match in 1997. Also, AI is used in video games to produce bots that play the game alongside humans.
Military	Although many AI researchers seek to distance themselves from military applications, AI is currently used to develop military drones capable of autonomous actions and unmanned combat aerial vehicles.
Healthcare	AI has been successfully used to extract information on treatment patterns and diagnoses from large digital databases. Furthermore, robotic surgeries are being developed and performed, with the first unassisted surgery taking place in 2006 on a patient having heart arrhythmia.
Finance and economics	Systems to detect unauthorized use of debit cards have been in use since 1987 and AI also has an impact in online trading, stock investment decisions and preventing financial fraud.
Robotics	The iRobot Corporation has sold over two million Roomba robotic vacuum cleaners for home use. In addition to this, robotic manipulators are often used in industrial workflows, where repetitive actions are needed or precision is required.
Speech and image recognition	Image recognition methods are used in the analysis of medical imaging results and the subsequent diagnosis of disease, but also in day to day objects, such as cameras with face recognition or surveillance systems. Speech recognition has been proven very useful in the development of online assistants, such as Siri.
Aviation	Airlines use expert systems in planes to monitor the atmospheric conditions and system status, enabling planes to be put in autopilot. Also, the use of artificial intelligence in building simulators and analyze the data gathered by using them is proving to be very beneficial to the industry.
Education	Intelligent tutoring systems have been used to teach Air Force technicians to diagnose electrical problems in aircrafts and to train Navy recruits in technical skills in a shorter amount of time.
Marketing	AI techniques are used to back up marketing decisions by analyzing trends, providing forecasts, reducing information overload and allowing for up-to-date information.

4. Applications of AI in medicine and bioinformatics

Medicine seems particularly amenable to AI solutions [20] and has been the focus of much interest in thriving technological economies. The impact of AI can be grouped in two main topics: extracting meaning from large amounts of medical data in the search domain and aiding clinicians in delivering care to patients. Data-driven predictions of drug effects and interactions, identification of type 2 diabetes subgroups and the discovery of comorbidity clusters in autism spectrum disorders are just some of the successful results of using AI to extract information from large databases of Electronic Health Records. In the United States, machine learning approaches have been used to create a decision support system for physicians treating cancer patients, with the intention of improving diagnostic accuracy and reducing costs using large volumes of patient cases and scholarly articles.

As both the number of imaging studies and the number of images per study grows, the incorporation of computer-aided detection systems into the diagnostic process could improve the performance of image interpretation by providing quantitative support for clinical decision making, particularly the differentiation of malignant and benign tumors.

Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data. The technical advances in the last years have increased the amount of data that biologists can record about different aspects of an organism at the genomic, transcriptomic and proteomic levels, and the discipline of bioinformatics has allowed scientists to exploit the advances in computer science and computational statistics in analyzing this data. But as the volume of data grows, the techniques used must cater for large-scale data [15].

Such an approach is ideal because of the ease with which computers can handle large quantities of data and probe the complex dynamics observed in nature [16]. But this merger of disciplines is not as surprising considering that life itself is information. An organism's physiology is largely determined by its genes, which at its most basic can be considered as digital information.

Bioinformatics tackles three main topics of handling biological data [16], particularly data regarding macromolecules such as DNA, RNA and proteins, and those are: organizing data in a way that makes it easily accessible for researchers, building specially developed tools and using those tools to analyze the data and interpret the results in a biologically meaningful manner. AI has a role in developing and applying particular methods that use biological data, such as DNA sequences or amino acid chains, to help understand different physiological functions or pathological processes within an organism.

DENDRAL was the first rule-based system applied to a "real-world" problem. Its development began at Stanford University in 1965 under the guidance of E. Feigenbaum, B. Buchanan, J. Lederberg and C. Djerassi and it spanned approximately half the history of AI research. It was used by chemists to determine the molecular structure of different organic

compounds by analysis of certain physical spectra of the molecules. It was one of the first large-scale programs to incorporate the strategy of using detailed, task-specific knowledge about the problem domain as a source of heuristics and to seek generality through automating the acquisition of such knowledge. It used a substantial amount of knowledge of chemistry and thus managed to reach a high level of performance. DENDRAL was a knowledge driven program and one of the first to conceptually separate the knowledge base that could be edited or redefined for new problems, from the code that would remain the same for interpreting and using that knowledge [17, 18].

DNA sequence analysis is another [19] topic that has attracted computer scientists to use AI techniques because of the availability of digital information. But there are also some challenges related to this area, such as:

- Parsing a genome in order to find the segments of DNA with various biological roles (sequences that encode proteins or that control when and where proteins are expressed).
- Aligning the sequences of DNA in order to check for similarities between them.

A summary of some of the most important applications of AI in bioinformatics is given in Table III.3, grouped by the techniques employed to analyze the data.

Table III.3. Some important applications of AI in the field of bioinformatics [15]

Technique	Bioinformatics applications
Nearest neighbor and clustering approaches	Both algorithms could provide good solutions where implementation and computation time are a priority. They can be used to determine useful information from high-dimensional data, but also as a method for pre-processing data for use by other algorithms.
Decision trees	• HIV and Hepatitis C protease cleavage prediction: See5 was developed to determine whether there was a pattern of amino acids in the substrate that could help determine whether the viral protease did or did not cleave, for the design of possible future protease inhibitors. • Classification of cancer by using diagnosis data: using a committee of decision trees to decide the outcome of the classification task (deciding if the data corresponds to ovarian cancer or not). Performs better than See5, but may require significant extra computation.
Neural Networks	• Gene expression analysis: using neural networks or perceptrons to attempt to distinguish between diseased and normal individuals, or to distinguish between two types of a disease by solely using the expression values of genes taken from those individuals. • Identifying protein subcellular location: using a Kohonen neural network to predict where a protein was located, based on its amino acid make-up, because it can provide important clues as to its function in the cell.

Genetic Algorithms	• Reverse engineering of regulatory networks: taking snapshots of a system at different times, consisting of the gene expression data and constructing a graph representing the regulatory network of the system. • Multiple sequence alignment: matching two or more DNA or amino acid sequences in order to find similarity between genes or proteins that may lead to similarity also in function.
Cellular Automata	CA allows the behavior of molecules to be investigated in highly complex environments where there might be many hundreds of molecules interacting at once. • Simulation of an apoptosis(cellular death) network • Cellular automata model for enzyme kinetics

Another topic of major interest at the crossroads of molecular biology, chemistry and artificial intelligence is the prediction of protein structure from the amino acid sequence. This is also the focus of this paper and will be discussed in depth in the next chapter, classifying the different techniques based on algorithm, exterior knowledge used or the structure type that is being predicted.

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.

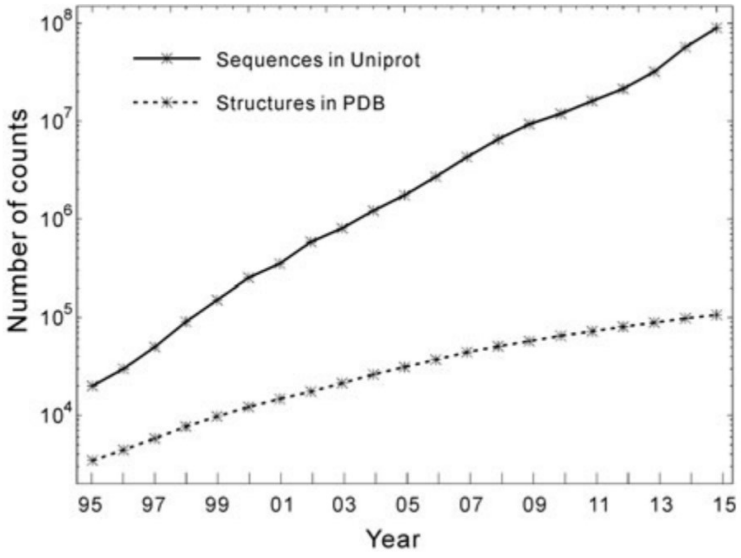

Figure IV.1. The number of available protein sequences (in UniProt) and solved protein structures (in the PDB) from 1995 until 2015 [21]

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.

1. 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).

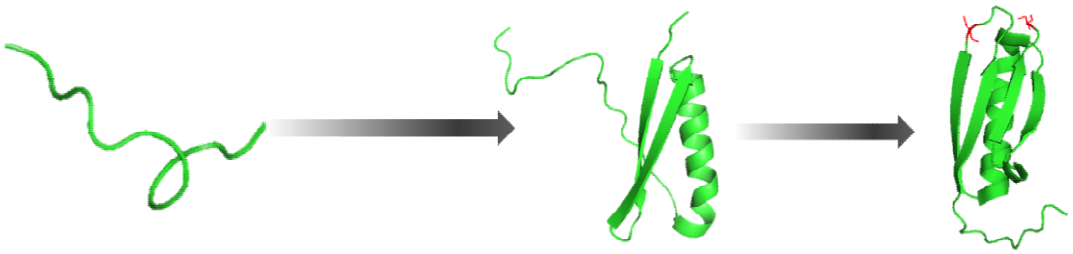

Figure IV.2. Steps a protein takes in order to assume its native three dimensional conformation [22]

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.

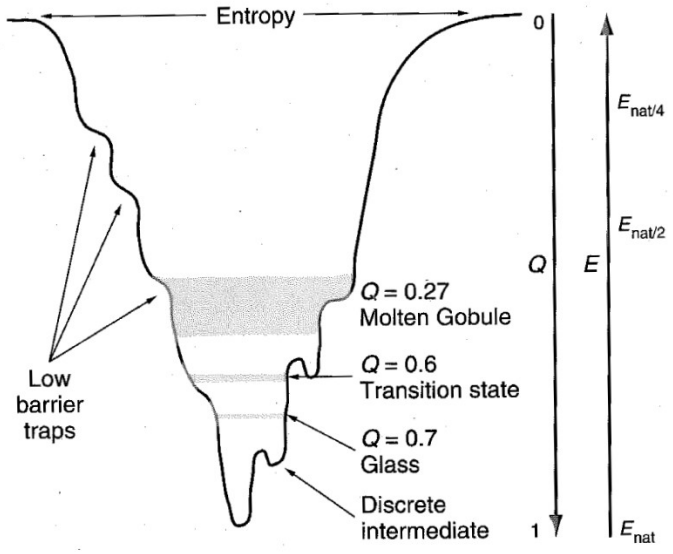

Figure IV.3. Cross section through a folding funnel, where E corresponds to the free energy of the conformation [1]

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.

2. 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: $O(n^k)$, 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.

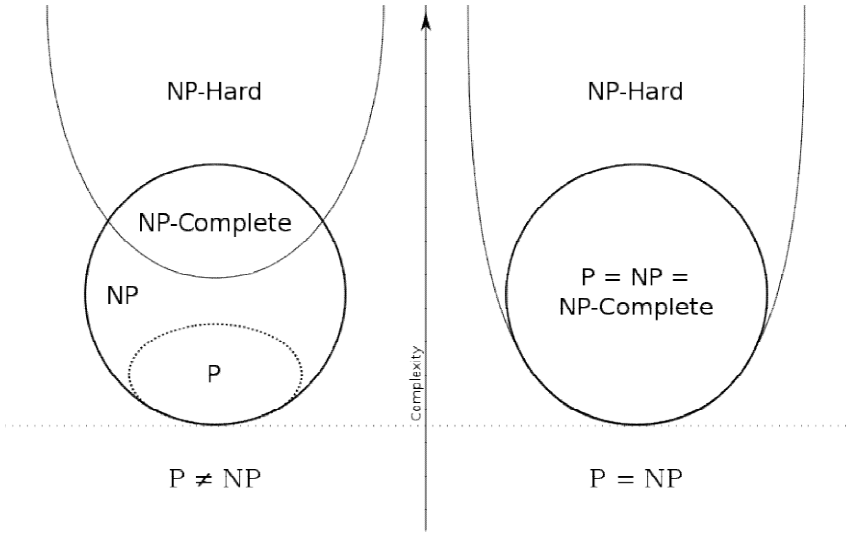

Figure IV.4. Diagrams for the P, NP, NP-Complete and NP-Hard problems. The left side applies when P and NP are different and the right side is valid otherwise

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.

3. 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).

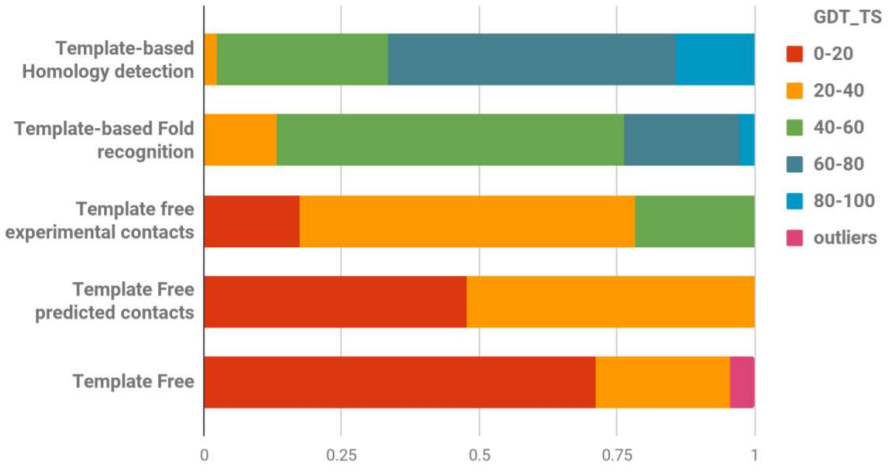

Figure IV.5. The GDT_TS scores for different model categories in CASP11 [25]

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.

4. 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].

4.1 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.

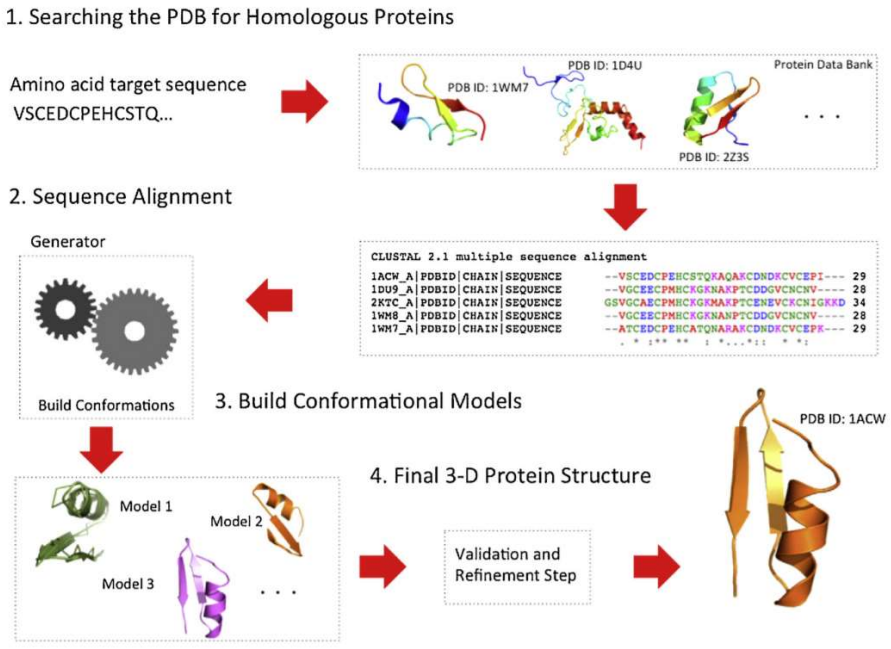

Figure IV.6. Schematic representation of the process of template-based modeling using homologous proteins for three-dimensional protein structure prediction [4]

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.

4.2 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.

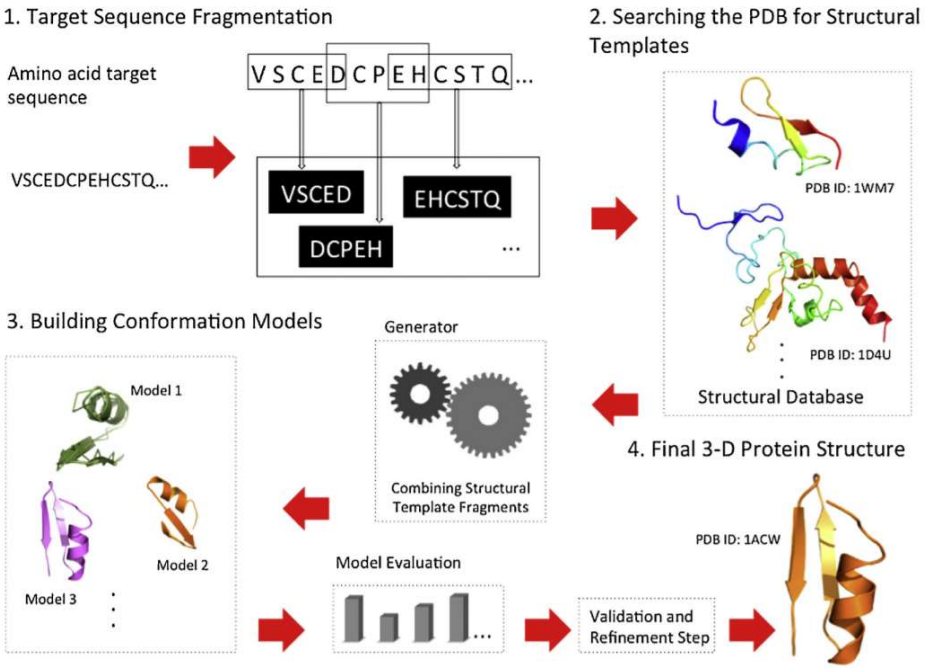

Figure IV.7. Schematic representation of a fragment-based FM method for the prediction of protein tertiary structure [4]

4.3 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.

5. AI techniques for protein structure prediction

VII. References

[1] Fersht, A.: Structure and Mechanism in Protein Science, A Guide to Enzyme Catalysis and Protein Folding, W. H. Freeman and Company, New York, 1999.

[2] Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., Walter, P.: Molecular Biology of the Cell, 6th Edition, Garland Science, Taylor & Francis Group, New York, 2015.

[3] Walsh, G.: Proteins, Biochemistry and Biotechnology, 2nd Edition, Wiley Blackwell, 2014.

[4] Dorn, M., Barbachan e Silva, M., Buriol, L.S., Lamb, L.C.: Three-dimensional protein structure prediction: Methods and computational strategies, Computational Biology and Chemistry, 53(2014), 251-276.

[5] Richardson, J.S: The Anatomy and Taxonomy of Protein Structure, Advances in Protein Chemistry, 34(1981), 167-339.

[6] Wlodawer, A., Dauter, Z., Jaskolski, M.(editors): Protein Crystallography, Methods and Protocols, Springer, New York, 2017.

[7] PDB Statistics, https://www.rcsb.org/stats/growth/overall

[8] Rose, P.W. et al.: The RCSB protein data bank: integrative view of protein, gene and 3D structural information, Nucleic Acids Research, Database issue, 45(2017), D271-D281.

[9] Russel, S.J., Norvig, P.: Artificial Intelligence, A Modern Approach, Third Edition, Prentice Hall, New Jersey, 2010.

[10] Buchanan, B.G.: A (very) brief history of Artificial Intelligence, AI Magazine, 4(2006), 53-60.

[11] Goertzel, B., Mossbridge, J., Monroe, E., Hanson, D., Yu, G: Loving AI: Humanoid Robots as Agents of Human Consciousness Expansion, 2017.

[12] Kim, S.S.Y., Dohler, M., Dasgupta, P.: The Internet of Skills: The use of 5th generation telecommunications, haptics and artificial intelligence in robotic surgery. BJU International, 2018.

[13] Hashimoto, D.A., Rosman, G., Rus, D., Meireles, O.R.: Artificial Intelligence in Surgery: Promises and Perils, Annals of Surgery, 2018.

[14] Binner, J.M., Kendall, G., Chen, S.H.: Applications of Artificial Intelligence in Finance and Economics, Advances in Econometrics, 19(2004).

[15] Keedwell, E., Narayanan, A.: Intelligent Bioinformatics, John Wiley & Sons, Ltd, 2005.

[16] Luscombe, N.M., Greenbaum, D., Gerstein, M.: What is Bioinformatics? A proposed definition and overview of the field, Methods of Information in Medicine, 40(2001), 346-358.

[17] Lindsay, R.K., Buchanan, B.G., Feigenbaum, E.A., Lederberg, J.: DENDRAL: a case study of the first expert system for scientific hypothesis formation, Artificial Intelligence, 61(1993), 209-261.

[18] Feigenbaum, E.A., Buchanan, B.G, Lederberg, J.: On generality and problem solving: a case study using the DENDRAL program, Stanford Artificial Intelligence Project Memo No.131, 1970.

[19] Ezziane, Z.: Applications of artificial intelligence in bioinformatics: a review, Expert Systems with Applications, 30(2006), 2-10.

[20] Jones, L.D., Golan, D., Hanna, S.A., Ramachandran, M.: Artificial Intelligence, machine learning and the evolution of healthcare, Bone Joint Research, 7(2018), 223-225.

[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.